Veridical Data Science towards Trustworthy AI

Bin Yu

Statistics, EECS, Center for Computational Biology

Simons Institute for the Theory of Computing

IMS ICSDS, Seville, Spain

Dec. 15, 2025

Alternative Title:

Veridical Data Science is a Frontier of Statistics in the Age of AI

This talk is dedicated to

Berkeley Statistics Department
at its 70th Anniversary (1955-2025), and

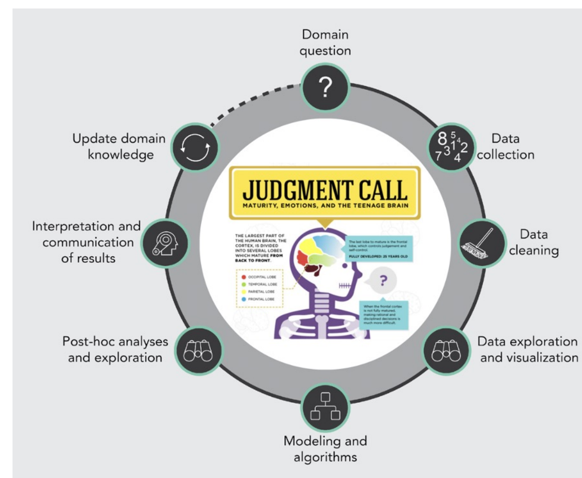
Bell Labs
at its 100th Anniversary (1925-2025)

What does “veridical” mean in VDS?

Veridical means “truthful” in two ways in VDS:

1. It seeks truth in data conclusions

1. It is truthful to the data science life cycle (DSLC)



Outline of talk

1. Statistics needs to adapt to the AI age
2. VDS with core principles of Predictability-Computability-Stability (PCS) is a frontier of statistics
3. VDS success stories …
4. Theory and processes of productive theoretical research
5. PCS current directions and resources

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**, … **carcinogenesis**."*

– 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",** …

Collaborative research culture and **"in-context"** research **norm**.

How does statistics thrive?

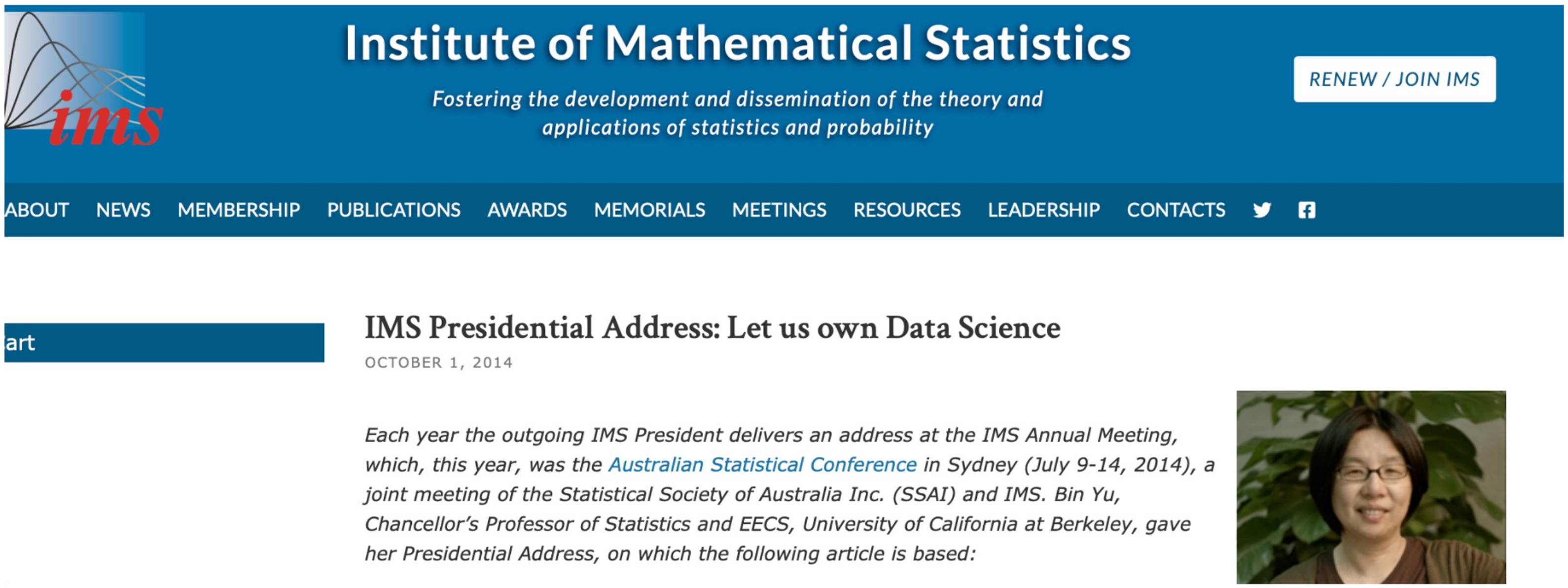
How does statistics thrive?

"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, ...

In 2025, We Are in the Age of AI

"Life is complicated, but not uninteresting." – J. Neyman

“**Statistics** is the science of learning from data, and of measuring, controlling and communicating uncertainty.”

– ASA

Statistics is at an **inflection point**...

Goal of Statistics in the Age of AI:

To provide **data evidence in context** for trustworthy conclusions and decisions, which rely on an entire DSLC.

Berkeley statistics in a new college CDSS (Computing, Data Science, and Society)

Two popular courses: Intro DS class (**Data8**, 2015) and DS Techniques (**Data100**, 2017) (of 1500 students each course each semester).

A new college CDSS in 2019, which now houses the DS Major (co-owned by stats and EECS), Stats Major (Stats), and CS Major (EECS)

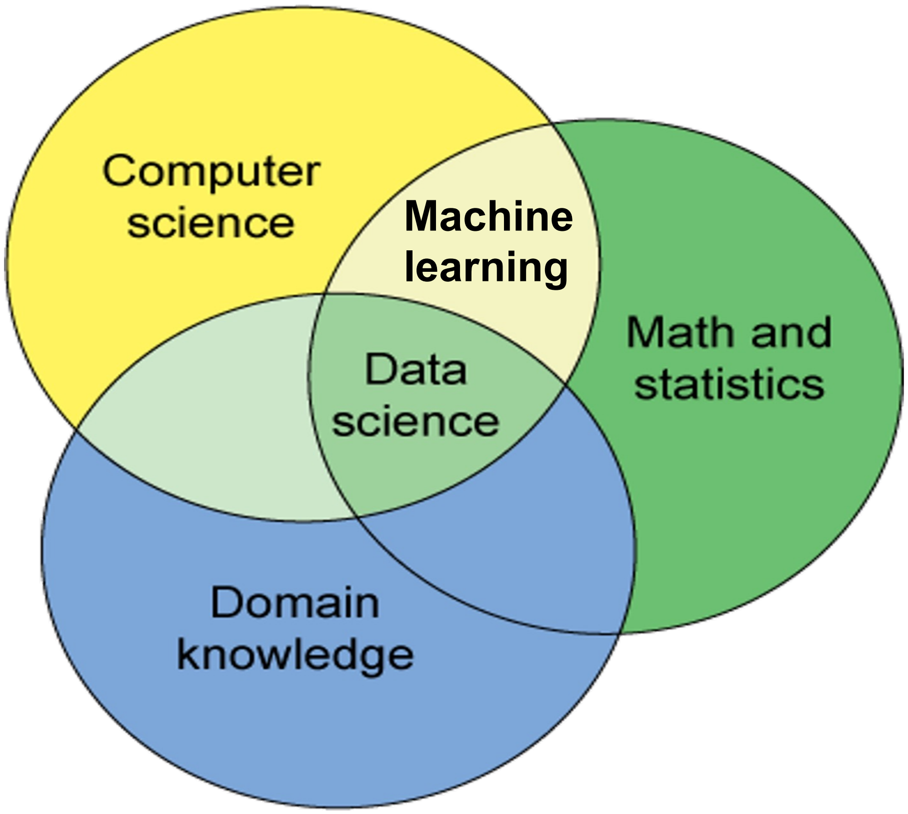
UC Berkeley College of Computing, Data Science, and Society

CDSS units

- Statistics Department
- EECS Department (joint with College of Engineering)
- Center for Computational Biology
- Computational Precision Health (joint with UCSF)
- BIDS (Berkeley Institute of Data Science)
- IDSI (Interdisciplinary Data Science Institute) (in the process)

We explore solutions to society's greatest challenges through computing and data science

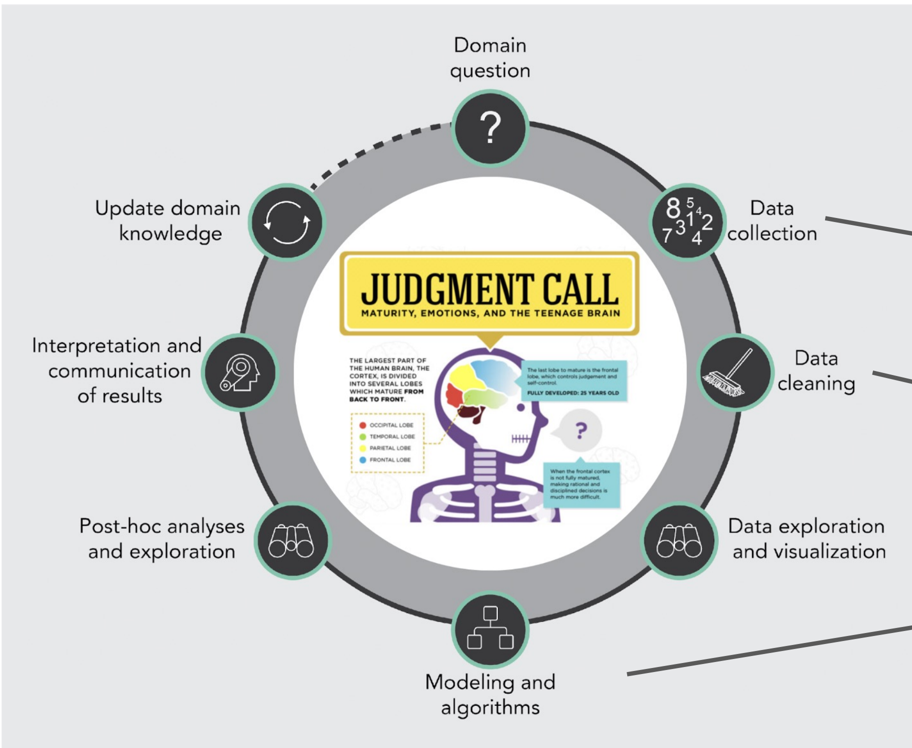
Data science (DS) is Foundation of AI



Conway's Venn Diagram

Thanks to chatGPT.

Data Science Life Cycle (DSLC)



Box (1979). Cox and Snell (1981), Nelder (1991)....

Image credits: R. Barter and toronto4kids.com

Human judgment calls in every step and they create missing **uncertainty**:

What choices were made while collecting data?

How was the data cleaned?

Modeling choices

A DSLC creates uncertainty in every step!

Uncertainty from analyst choices:

social science (there is a similar paper from biologists)

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

Nate Breznau, Eike Mark Rinke, Alexander Wuttke, +162, and Tomasz Żółtak Authors Info & Affiliations

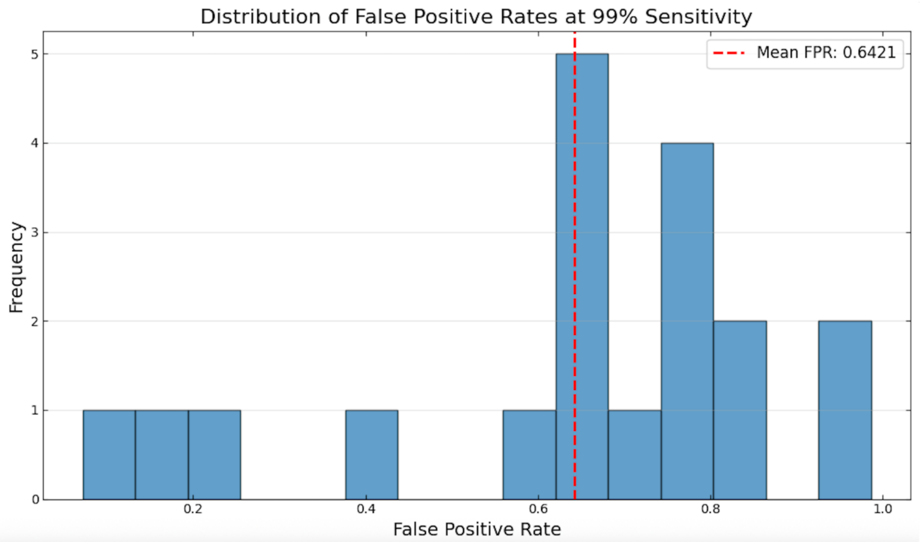
Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022

"… **Seventy-three independent research teams** used identical cross-country survey data to test a prominent social science hypothesis… **teams' results varied greatly, ranging from large negative to large positive effects of immigration on social policy support.**"

Another uncertainty-source NOT accounted for: data cleaning choices (MA stat class, UCB)

- Goal: clinical decision rule on CT-scan or not for pediatric patients (with traumatic brain injuries), using clinical variables and labels

- 18 indep. students: the same raw data , same data cleaning guidelines



At 99% sensitivity, estimated false positive rates range from 7.3% to 98.7%

– a **90% difference!**

Judgement calls (data cleaning) creates **uncertainty**!

Patterns (2023)

Leakage and the reproducibility crisis in machine-learning-based science

Highlights

- Data leakage is a flaw in machine learning that leads to overoptimistic results

Authors

Sayash Kapoor, Arvind Narayanan

Foundations of science are eroded when errors propagated from the initial 41 papers to 648 papers.

15 years earlier: reproducibility crisis

"Scientists from biotech companies Amgen and Bayer Healthcare reported alarmingly **low replication rates (11–20%)** of landmark findings in preclinical oncological research."

-Wikipedia on "replication crisis"

Begley CG, Ellis LM (March 2012). "Drug development: Raise standards for preclinical cancer research". *Nature*. **483** (7391): 531–533.
Prinz F, Schlange T, Asadullah K (August 2011). "Believe it or not: how much can we rely on published data on potential drug targets?". *Nature Reviews. Drug Discovery*. **10** (9): 712.

Image from https://www.nature.com/articles/533452a

We are in an AI Reproducibility Crisis in Science

Science

Current Issue First release papers Archive About

HOME > SCIENCE > VOL. 359, NO. 6377 > ARTIFICIAL INTELLIGENCE FACES REPRODUCIBILITY CRISIS

IN DEPTH | COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify.

MATTHEW HUTSON Authors Info & Affiliations

SCIENCE • 16 Feb 2018 • Vol 359, Issue 6377 • pp. 725-726 • DOI: 10.1126/science.359.6377.725

2018

2023

nature

Explore content About the journal Publish with us Subscribe

nature > news feature > article

NEWS FEATURE | 05 December 2023

Is AI leading to a reproducibility crisis in science?

Scientists worry that ill-informed use of artificial intelligence is driving a deluge of unreliable or useless research.

By Philip Ball

95% Failure Rate of AI Projects based on an MIT Report

Why 95% Of AI Projects Fail And How Better Data Can Change That

Forbes

By Gary Drenik, Contributor. Gary Drenik is a writer covering AI, analytics a...

Follow Author

Published Oct 15, 2025 at 10:00am EDT

Share Save Comment 0

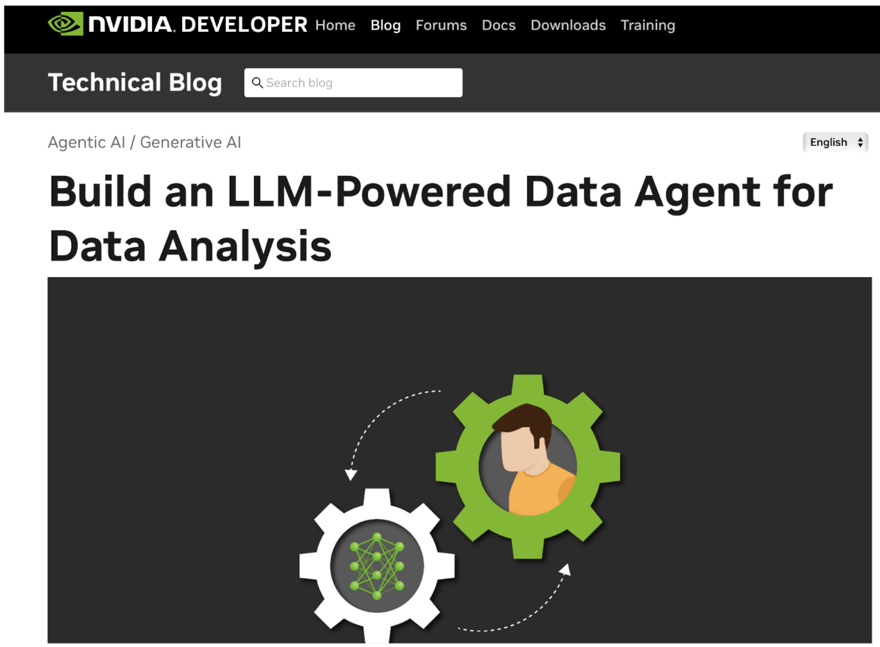
Add Us On Google

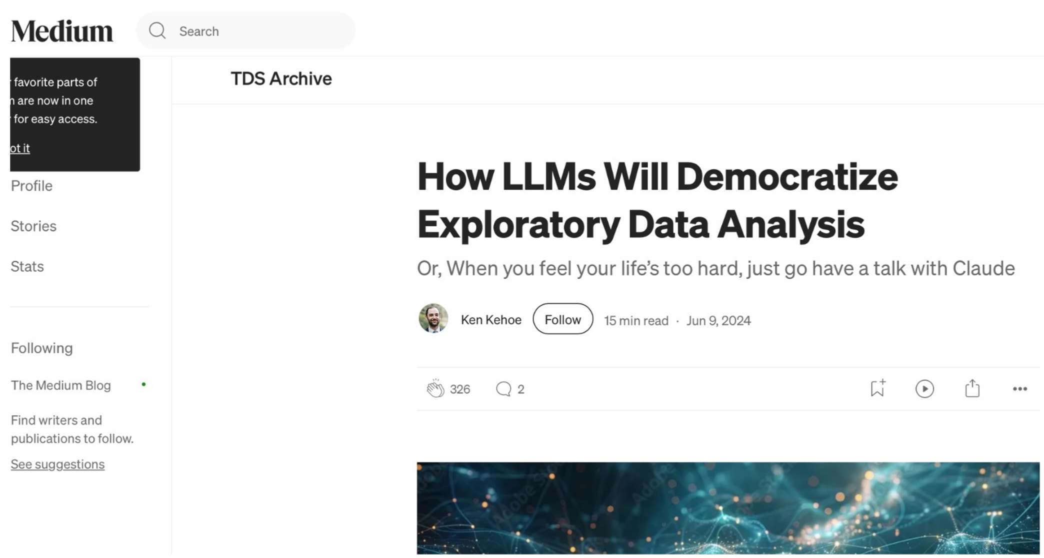
ADVERT

https://www.forbes.com/sites/garydrenik/2025/10/15/why-95-of-ai-projects-fail-and-how-better-data-can-change-that/

As we will see later in this talk, better data is not enough.

LLMs are doing data analysis, like it or not





AI frontiers: safety (and rigorous evaluation)

California Just Passed the First U.S. Frontier AI Law. Here's What It Does.

SB-53 offers a blueprint for evidence-generating transparency measures that could shape the next few years of frontier AI governance.

By Scott Singer and Alasdair Phillips-Robins
Published on October 16, 2025

https://carnegieendowment.org/emissary/2025/10/california-sb-53-frontier-ai-law-what-it-does?lang=en

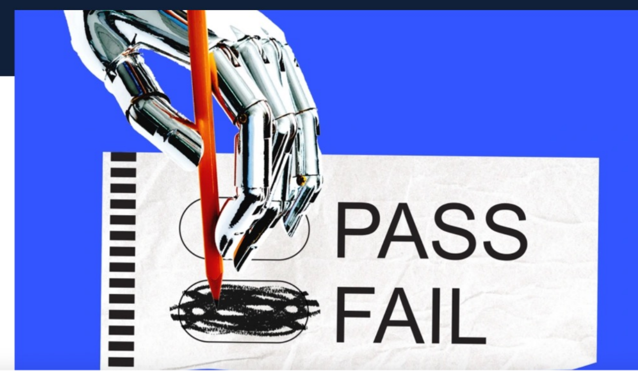
NBC NEWS

Nov. 5, 2025

ARTIFICIAL INTELLIGENCE

AI's capabilities may be exaggerated by flawed tests, according to new study

A study from the Oxford Internet Institute analyzed 445 tests used to evaluate AI models.



By Jared Perlo

Researchers behind a new study say that the methods used to evaluate AI systems' capabilities routinely oversell AI performance and lack scientific rigor.

The study, led by researchers at the Oxford Internet Institute in partnership with over three dozen researchers from other institutions, examined 445 leading AI tests, called benchmarks, often used to measure the performance of AI models across a variety of topic areas.

Questions for statistics community

- How well can genAI do basic data analysis in 2 year? How well can genAI do mid-level data analysis in 5 years?
- Do we leave genAI development for data analysis to CS/AI people?
- Entry level software engineer jobs are disappearing, will this happy to entry level statistics and data science jobs in the next few years?
- How do we prepare for this likely event?

A unique opportunity and responsibility for statisticians

To engage in

AI research,

AI reproducibility research, and

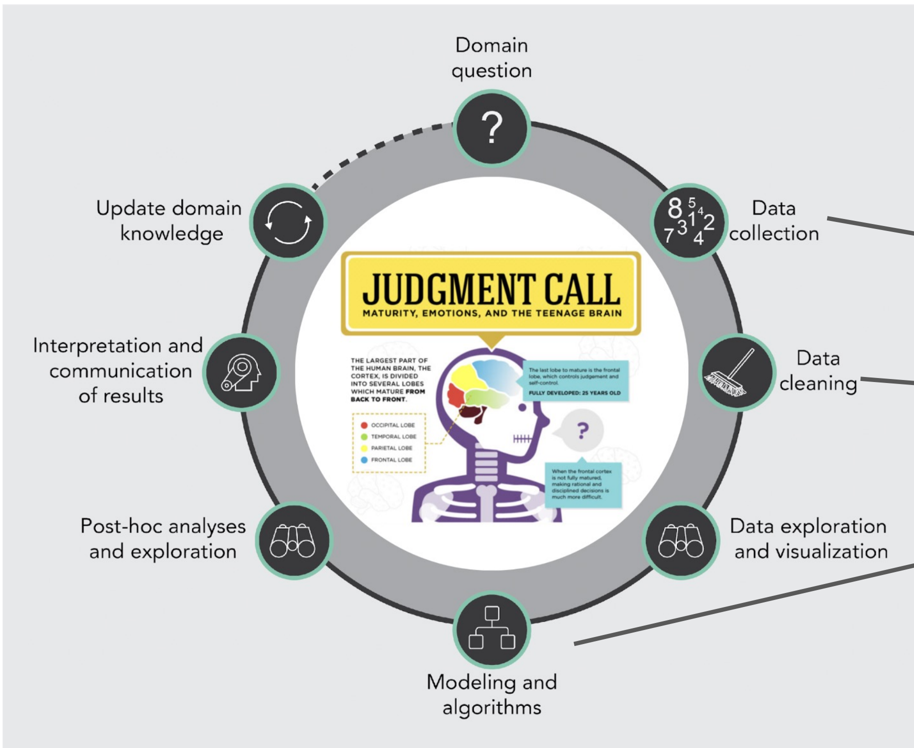
AI safety research,

for statistics to thrive in the age of AI.

In our house of “uncertainty”

https://www.housebeautiful.com/room-decorating/living-family-rooms/g715/designer-living-rooms/
https://www.forbes.com/sites/janicegassam/2019/11/27/the-pink-elephant-in-the-workplace-how-to-have-conversations-about-race-politics-and-religion-at-work/

Data Science Life Cycle (DSLC)



Human judgment calls in every step and they create missing uncertainty:

What choices were made while collecting data?

How was the data cleaned?

Modeling choices

A DSLC creates uncertainty in every step!

Box (1979). Cox and Snell (1981), Nelder (1991)....

Image credits: R. Barter and toronto4kids.com

Neyman recognized the human element or “act of will” in statistical modeling work

“the mental processes behind the new method of estimation consist of deductive reasoning and of **an act of will**.”

– J Neyman (1951, "Foundations of the General Theory of Estimation"

Neyman recognized the human element or “act of will” in statistical modeling work

“the mental processes behind the new method of estimation consist of deductive reasoning and of **an act of will**.”

– J Neyman (1951, "Foundations of the General Theory of Estimation"

Formal recognition of human judgment calls in DSLC not only demands **transparency** in reporting, but also provides a great opportunity for **aggregation** for better results – not unsimilar with seeking second opinions and combining opinions in medicine.

Statistics thrives by meeting AI challenges

Goal of Statistics Today:

To provide **data evidence in context** for trustworthy conclusions and decisions, which rely on an entire DSLC.

Thus statistics becomes a **systems science**. It is in need of fundamental principles that apply to the multiple steps of a DSLC or a DS or AI development workflow.

Outline of talk

1. Statistics needs to adapt to the AI age
2. VDS with core PCS principles is a frontier of statistics
3. VDS success stories …
4. Theory and processes of productive theoretical research
5. PCS current directions and resources

VDS is built on three core princip

Veridical Data Science (VDS) for trustworthy DS and AI

Y. and Kumbier (2020)

PNAS

VDS is built on three core principles of data science for **every step** of the **data science life cycle** (DSLC):

(**P**)redictability [ML and Stats] (**“reality-check”**)

(**C**)omputability [ML, “R”]

(**S**)tability [expanding uncertainty sources with user defined perturbations]

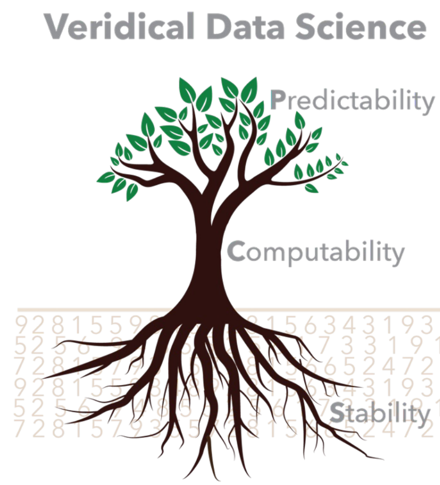


Image credit: R. Barter

PCS documentation for **transparency** and **trust**

[on GitHub (JupyterNotebook Quarto)]

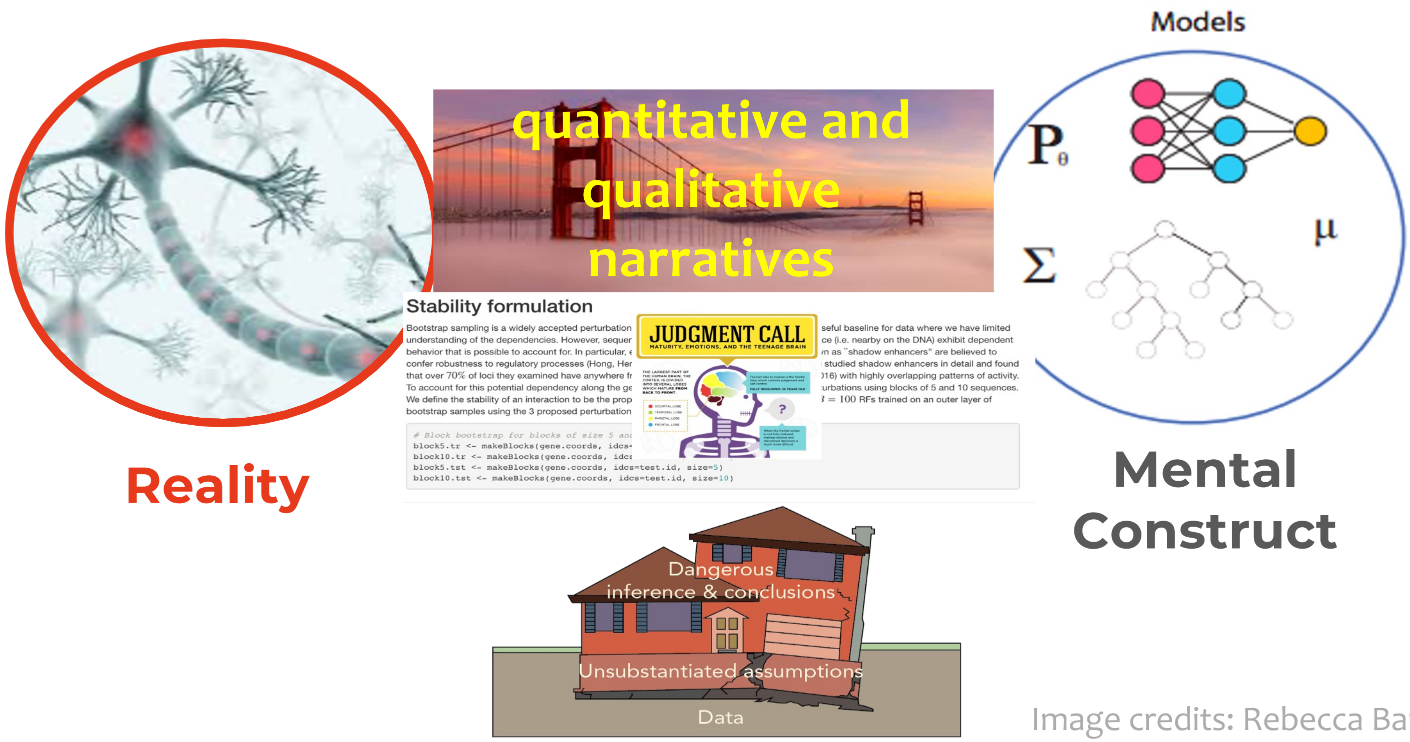


Image credits: Rebecca Barter

PCS documentation template: https://yu-group.github.io/vdocs/PCSDoc-Template.html

PCS for VDS – a new paradigm

In hindsight (after a decade of development), PCS principles are common-sense principles:

“Pred-check” is about general reality check including model checking…

“S” addresses new sources of uncertainty in a DSLC

“C” is indispensable and includes data-inspired simulations

PCS for VDS as a systems science

Guiding every step of a DSLC or an AI workflow (e.g. data cleaning)

- Moving away from “true model” framing or rejecting the assumption that a model equals reality.
- Enforcing “reality checking” and differentiating different types of probabilistic models (hence differentiating diff. strengths of evidence), through “P” and documentation
- Uncertainty quantification (as a special form of “S”) not based on limiting distributions, but in the spirit of bootstrap (more later)

PCS for VDS – a new paradigm

Unifying, synthesizing, and **expanding** on ideas and best practices in **Machine Learning** and **Statistics** to cover the entire **DSLC**.

Providing a unified language/concept to assess and improve both reality-check and stability/robustness towards reproducible results and decisions, and to **communicate**

among statisticians, data scientists, and AI researchers, to and among **scientists or domain experts, managers in industry, and users of DS and AI, …**

Trustworthy AI or AI safety need communication/interpretation

How to choose perturbations for “S”?

For **each step** of DSLC, there are **multiple reasonable choices determined by context**, possibly favored with different weights based on prior knowledge, and subject to resource constraints; there might also be multiple stability metrics for each perturbation.

Meta judgment calls are still needed; aggregations could improve final results.

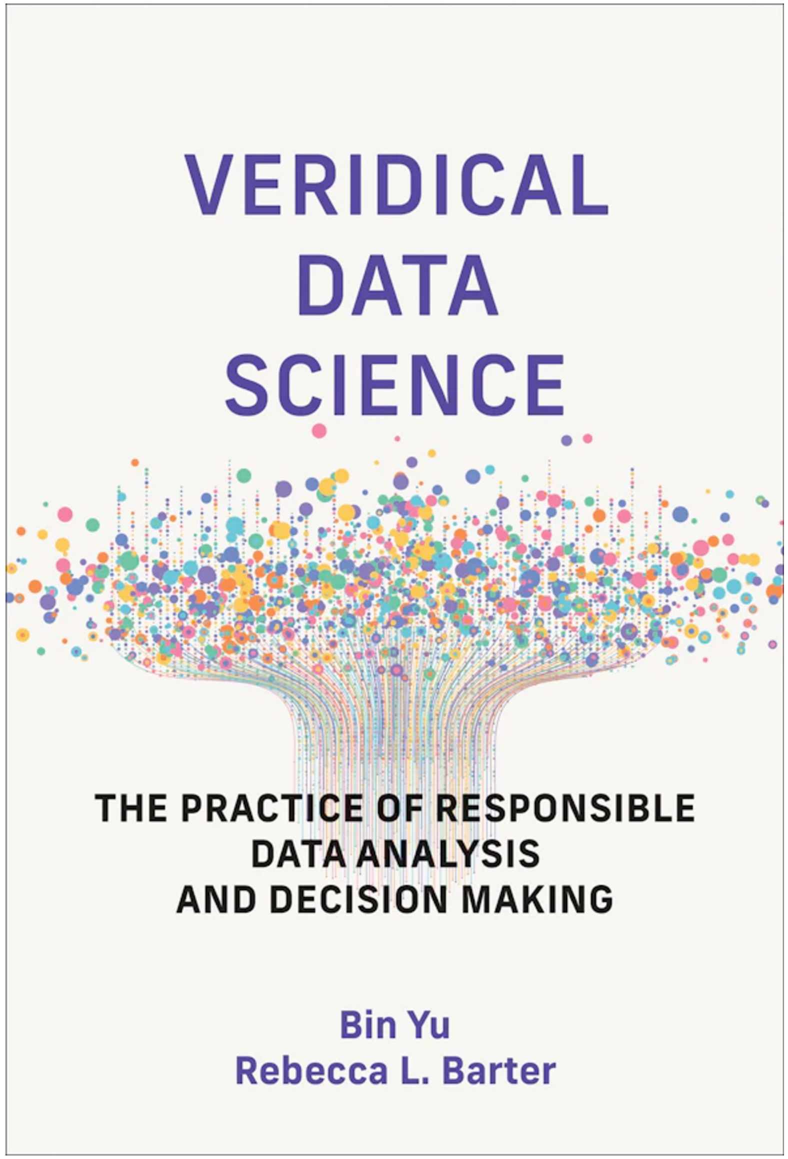
Record all human reasoning and judgment calls using **PCS documentation.**

Design issues in PCS implementation are research questions in PCS.

A related work is “Forking” by Gelman and Loken, 2014.

Bin Yu Rebecca Barter

Free version
vdsbook.com



MIT Press，'24
(ML Series)
Paper Book

Book review by Yuval and Yoav Benjamini

A Review of "Veridical Data Science" by Bin Yu and Rebecca L. Barter

Full article forthcoming.

by Yuval Benjamini and Yoav Benjamini

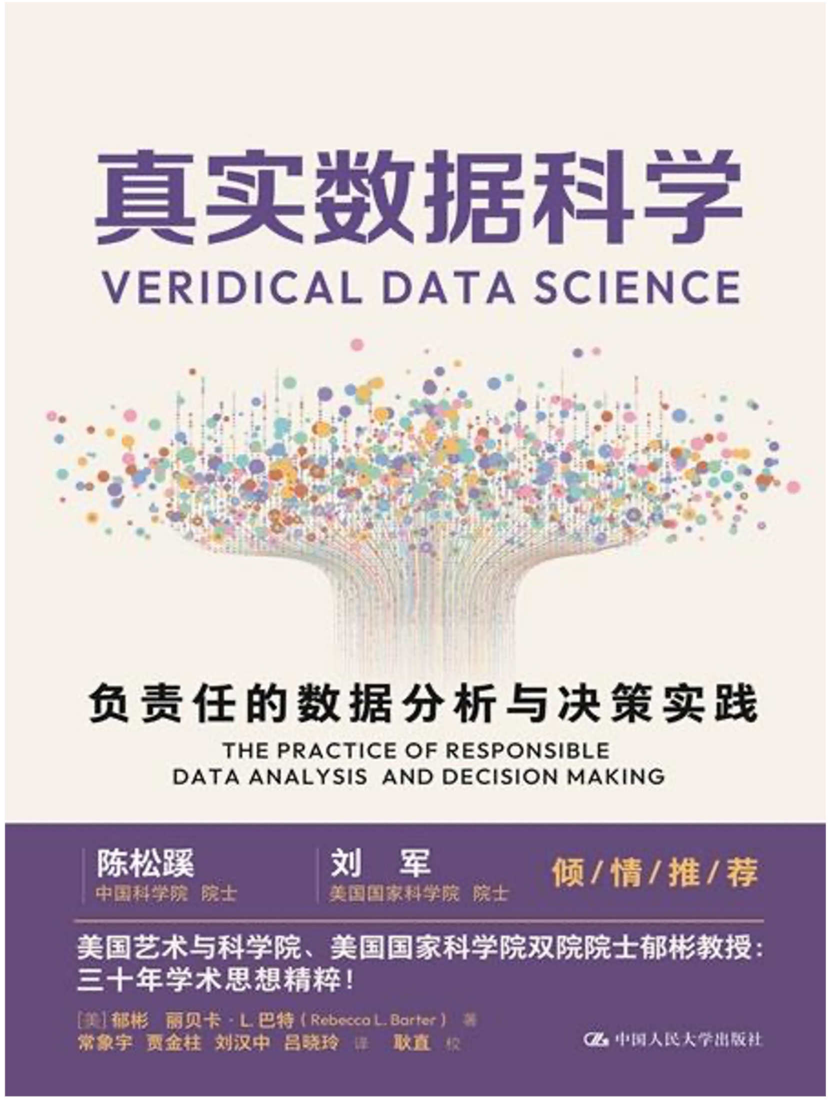
HARVARD DATA SCIENCE REVIEW

Editor-in-Chief (Xiao-Li Meng)'s Note: "*In this **inaugural book review** for* Harvard Data Science Review, *… The Benjamini duo discuss the potential uses and prospective readers of the book, concluding that its **pedagogical excellence, diverse examples, and projects** make* Veridical Data Science *a suitable textbook for students of all levels, in addition to being a valuable resource for data scientists in general.*"

免费线上英文版

vdsbook.com

郁彬　　丽贝卡 L 巴特



作者: 郁彬 / 丽贝卡 L 巴特

出版社: 中国人民大学出版社

译者: 常象宇 / 贾金柱 / 刘汉中 / 吕晓玲

校对：耿直

出版年: 2025-9

页数: 356

定价: 96

装帧: 平装

ISBN: 9787300337500

真实数据科学- 中国高校教材图书网

Accepted (2025) in Special Issue on Workflows of Applied Data Analysis
of *Philosophical Transactions of Royal Society A* (edited by Andrew Gelman)

Zach Rewolinski

PHILOSOPHICAL
TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

PCS Workflow for Veridical Data Science in the Age of AI

Zachary T. Rewolinski[1] and Bin Yu[1,2,3]

"This paper presents an updated and streamlined PCS workflow, tailored for practitioners and enhanced with guided use of generative AI..."

VDS draws on my research and teaching experience

I live in both the world of theory

https://lifeology.io/blog/2020/07/27/the-spherical-cow-an-insight-into-science/

and the world of practice

I work as a bridge between the two worlds.

https://photos.com/featured/cow-in-field-low-angle-robas.html?srsltid=AfmBOooIxDXoQM3yKMMjN9nJ4atS0rgVGQIm_lKCCZGHK55NXsvzQOsC

VDS is built on real-world data experience

"How can we differentiate between clouds in polar regions in satellite imagery?" (Shi et al. 2007) **(remote sensing)**

"How can we concisely summarize text documents using natural language processing" (Jia et al. 2014) **(NLP)**

"How do embryonic fruit flies form their organs?" (Wu et al. 2016) **(dev. biology)**

"How does the brain respond to visual stimuli (such as from movies and images)?" (Nishimoto et al. 2011) **(neuroscience)**

"How can we extract diagnostic information stored in pathology reports using NLP" (Odisho et al. 2020) **(digital health)**

"Which subgroups of patients are more likely to experience side effects when taking certain drugs?" (Dwivedi et al. 2020) **(clinical trial for drug dev.)**

PCS (in context) has had many successes

- **Finding genetic drivers of HCM** (experimentally validated) (**genomics**)
- **Cutting cost by ½ of a new prostate cancer detection algorithm** (**cancer res**)
- **Improving t-SNE and UMAP comp. biology**)
- Finding new meaningful subareas of the brain related to speech (**comp. neuro.**)
- Evaluating or stress-testing existing ER clinical decision rules (**medicine)**
- **New stat/ML algorithm developments in context** to add (appropriate) stability (e.g. **iterative random forests (iRF)**, **lo-siRF,** staNMF, staDISC, staDRIP, MDI+, **PCS ranking**, **NESS, PCS-UQ** ...)
- **Extensions by others** to veridical spatial data science, veridical network analysis, and **reinforcement learning** by others, and PCS-guided LLM development, ...

Outline of talk

1. Statistics needs to adapt to the AI age
2. VDS with core PCS principles is a frontier of statistics
3. VDS success stories …
4. Theory and processes of productive theoretical research
5. PCS current directions and resources

Three PCS success stories:
externally validated and refereed

Interpretability is critical for trustworthy AI and AI safety.

These stories are about interpretable and reproducible scientific results guided by **PCS**, which is a **prerequisite for interpretability.**

Causality Spectrum and PCS

Mechanistic Individual level	...	Average effect Group level
Stable, replicable		Effect depends on the group Stability implicit in causal inference: e.g. SUTVA

PCS works towards causality:

Predictability + stability (+ computability)

↓

interpretable hypothesis generation
recommendations for experiment

Success Story 1: new ML algorithm guided by PCS “in context”

Iterative random forests to discover predictive and stable high-order interactions

Sumanta Basu[a,b,c,1], Karl Kumbier[d,1], James B. Brown[c,d,e,f,2], and Bin Yu[c,d,g,2]

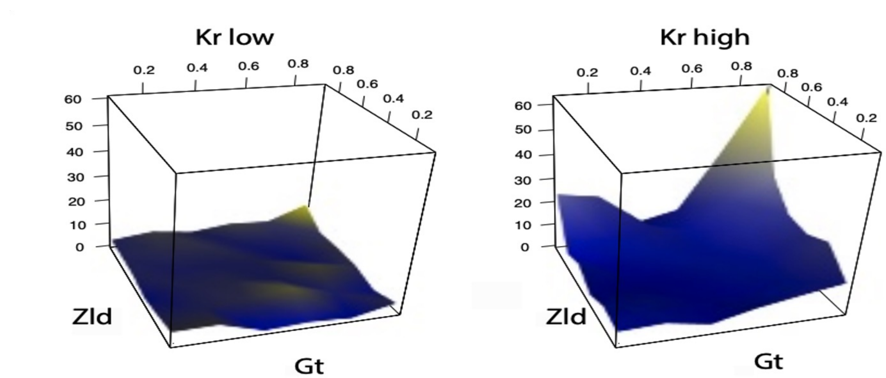
PNAS, 2018

Co-authors


S. Basu


K. Kumbier


B. Brown



Problem: predicting enhancer status from genomics data in fruitfly

80% success rate: 16 out of 20 iRF-found pairwise gene-gene interactions validated by past experiments.

iterative Random Forests (iRF)

Basu, Kumbier, Brown and Yu (2018)

Core idea: **add stability** to random forests (RF)

1. **Soft dim reduction** using importance index to sample features

1. Random interaction trees (RIT) to find intersections of paths

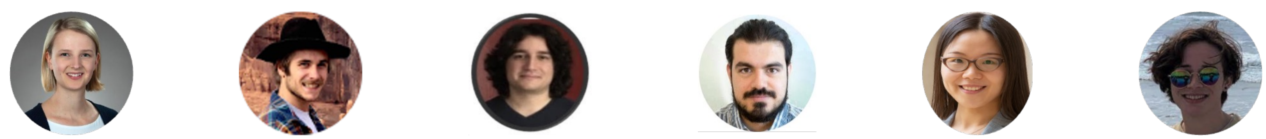
1. Outer-loop bagging assesses **stability**

Similar computational and memory costs as RF

Success story 2: finding **genetic drivers of** heart disease **HCM**

Berkeley UNIVERSITY OF CALIFORNIA | UCSF | Stanford | CHAN ZUCKERBERG BIOHUB

PIs: **Euan Ashley,** Rima Arnaout, Ben Brown, Atul Butte, James Priest, **Bin Yu**
Collaborators: Victoria Parikh, Chris Re, Deepak Srivastava



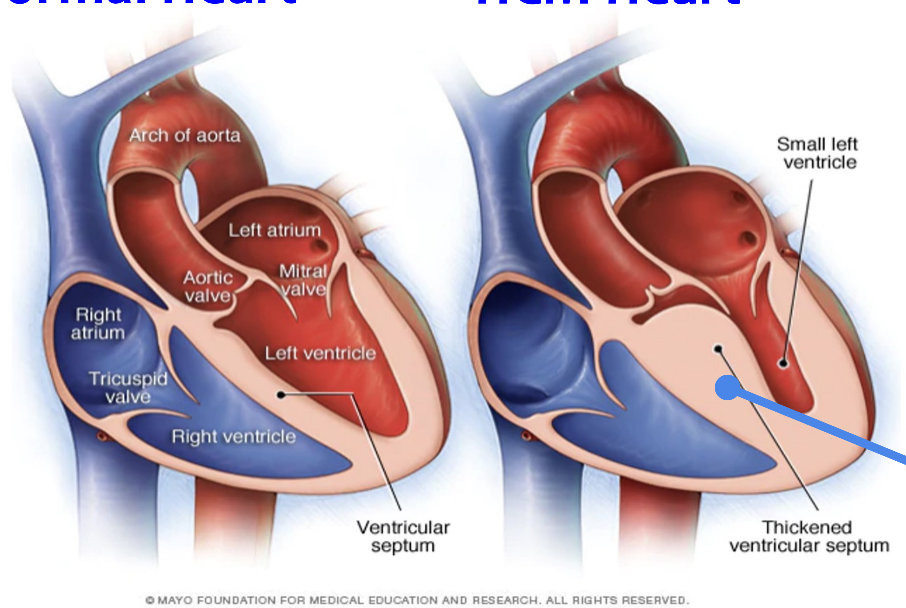
M. Behr *K. Kumbier* *M. Aguirre* *A. Cordova-Palomera* ***Q. Wang*** *N. Youlton*

C. Weldy *W. Hughes* *A. Agarwal* ***T. Tang*** *O. Ronen* *X. Li* *A. Kenney*

What is HCM?

Hypertrophic Cardiomyopathy (HCM) is a genetic heart disease, characterized by **thicker walls** of the heart chamber (left ventricle).

Normal Heart **HCM Heart**



Thickened heart wall

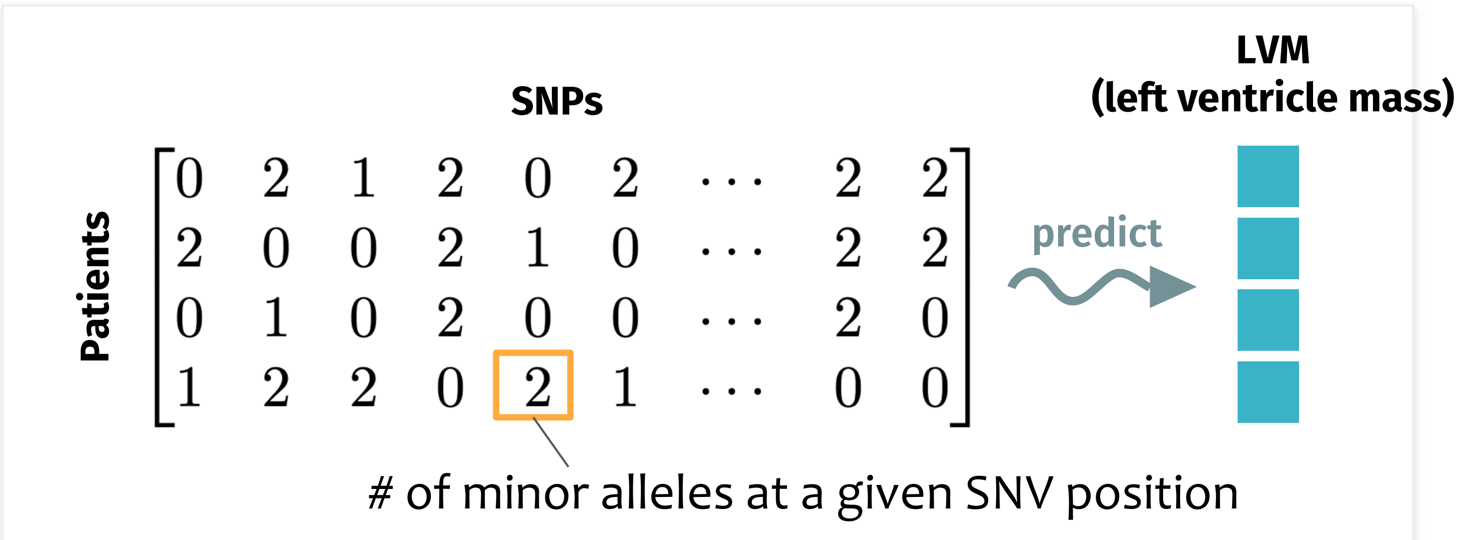
HCM rate is 1/500 in the US.

There is an important genetic component to it.

UK Biobank Data

n ~ 30K white British unrelated population with MRI data
p ~ 15 million imputed SNVs!!

HCM labels didn't work.



Cancer phenotypes?

Other disease phenotypes?

SNPs correspond to deviations from the "normal" genome states, hence could be predictive of diseases.

Finding genetic drivers of HCM: a low SNR problem (Wang et al (2025), *Nature Cardiovascular Res.*)

Problem reformulation (new phenotype, binarization for “P”)

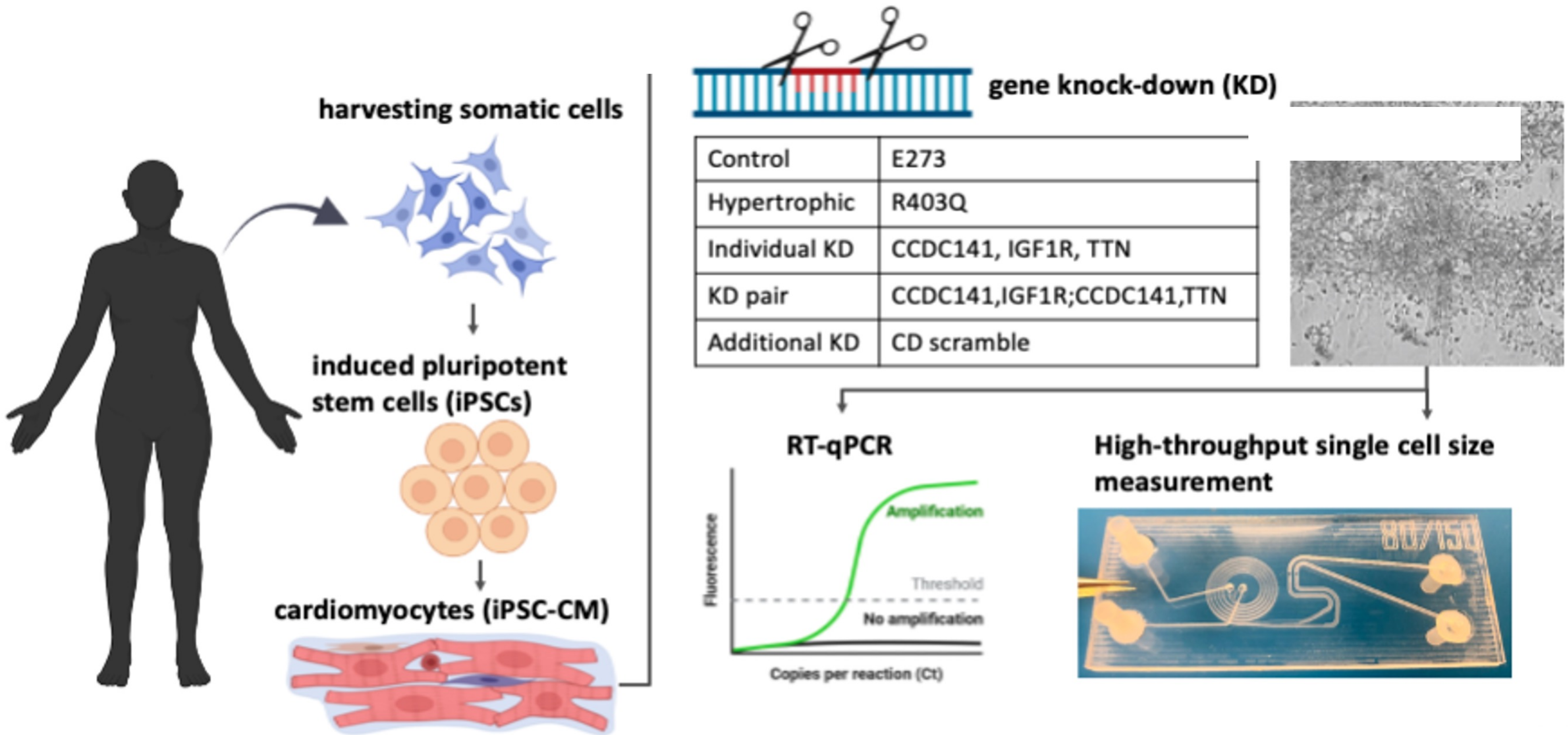
PCS-guided **lo-siRF prioritization** for gene and gene-gene interactions

Outperforming traditional interaction models that do not find credible gene-gene interactions (based on annotated databases)

Gene-silencing (intervention) causality validation experiments

Functional interpretation, Network analysis, Enrichment analysis

Gene Silencing experiments



Control	E273
Hypertrophic	R403Q
Individual KD	CCDC141, IGF1R, TTN
KD pair	CCDC141,IGF1R;CCDC141,TTN
Additional KD	CD scramble

Gene-silencing experimental validation results

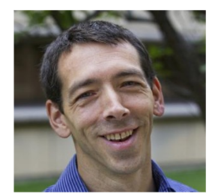
- **High yield rate 80%**: **4 out of 5 experiments successful (cost-effective)**
- Mechanistic interpretations for found epistatic interactions that drive HCM: **CCDC141-IGF1R** and **CCDC141-TTN – possible drug targets**

"Epistasis regulates genetic control of cardiac hypertrophy" (88 pp. + supp)
by Wang* and Tang*, ..., Y.* and Ashley* (2025)

Main co-authors:

Qianru Wang

Tiffany Tang



Euan Ashley

Stanford University

Berkeley UNIVERSITY OF CALIFORNIA

Success story 3: **PCS ranking** for

Cost-effective prostate cancer detection

Standard prostate cancer test
PSA has **very high false positive rate 90%** **(at 90% sensitivity)**

JAMA Oncology

Development and Validation of an 18-Gene Urine Test for High-Grade Prostate Cancer

Jeffrey J. Tosoian, MD, MPH[1,2]; Yuping Zhang, PhD[3]; Lanbo Xiao, PhD[3] ; et al

MyProstateScore2 (MPS2) (2024) lowers it to 60% (from Chinnaiyan group in UMich).

It uses 18 genes + clinical

LYNX DX

MICHIGAN

Success story 3: **PCS ranking** for

Cost-effective prostate cancer detection

Standard prostate cancer test
PSA has **very high false positive rate 90%** (at 90% sensitivity)

JAMA Oncology

Development and Validation of an 18-Gene Urine Test for High-Grade Prostate Cancer

Jeffrey J. Tosoian, MD, MPH[1,2]; Yuping Zhang, PhD[3]; Lanbo Xiao, PhD[3] ; et al

MyProstateScore2 (MPS2) (2024) lowers it to 60% (from Chinnaiyan group in UMich).

It uses 18 genes + clinical

LYNX DX

A simplified MyProstateScore2.0 for high-grade prostate cancer

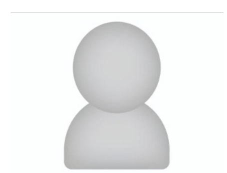
Cancer Biomarkers
Vol. 42(1): 1–17
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/18758592241308755
journals.sagepub.com/home/cbm
Sage | IOS Press

Data cleaning uncertainty 1-2% (AUC).
Our sMPS2 (2025) uses 8 genes instead of 18 and with similar false positive rate.

Joint US patent filed.


Tang


Kenney


Zhang


Chinnaiyan

MICHIGAN

Berkeley
UNIVERSITY OF CALIFORNIA

PCS for unsupervised learning: NESS

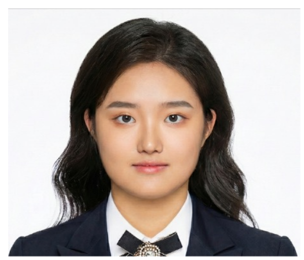
NESS: Neighbor Embedding for Smooth Structures using Stability Measure

PCS extended to unsupervised learning in Ch. 6-7 of Yu-Barter book.

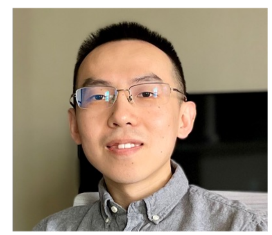
NESS is PCS-guided and on unsupervised learning.

Co-authors:

Rong Ma



Xi Li



Jingyuan Hu

"Uncovering smooth structures in single-cell data with PCS-guided neighbor

t-SNE & UMAP: non-linear 2D embedding methods

- **t-SNE** (van der Maaten & Hinton, 2008) & **UMAP** (McInnes et al., 2018) workflow:

  similarity graph + iteratively optimizing 2D embeddings **preserving local distances**

- Key hyperparameters: **graph connectivity** (perplexity for t-SNE, neighbor size for UMAP), **random initialization** (t-SNE), GD step size, ...

Main differences:

- t-SNE: random initialization, Gaussian-kernel-based similarity, early exaggeration stage
- UMAP: spectral initialization (KNN graph)

t-SNE & UMAP dominant in bio. research: visualization for biological insights and knowledge (e.g. cell types and cell progression)

Analysis | Published: 03 December 2018

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht, Leland M

Ng, Florent Ginhoux & E

Nature Biotechnology 3

116k Accesses | **4715**

Article | Open access | Published: 28 November 2019

The art of using t-SNE for single-cell transcriptomics

Dmitry Kobak & Philipp Berens

Nature Communications **10**, Article number: 5416 (2019) | Cite this article

Matters Arising | Published: 01 February 2021

Initialization is critical for preserving global data structure in both *t*-SNE and UMAP

Dmitry Kobak & George C. Linderman

Nature Biotechnology **39**, 156–157 (2021) |

24k Accesses | **275** Citations | **212** Altm

Technology Feature | Published: 24 May 2024

Seeing data as t-SNE and UMAP do

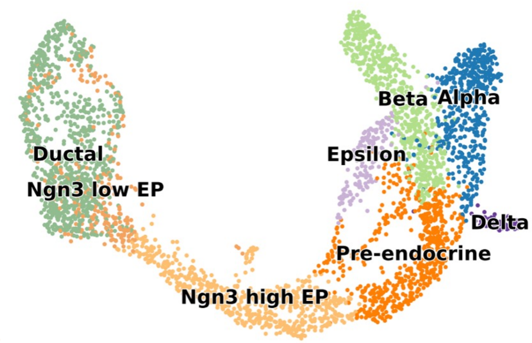
Vivien Marx

Nature Methods **21**, 930–933 (2024) | Cite this article

23k Accesses | **29** Citations | **44** Altmetric | Metrics

Used by 87% recent "single cell" papers

"Single cell" data: each cell has thousands or millions of measurements



UMAP visualization of pancreatic endocrine cell differentiation

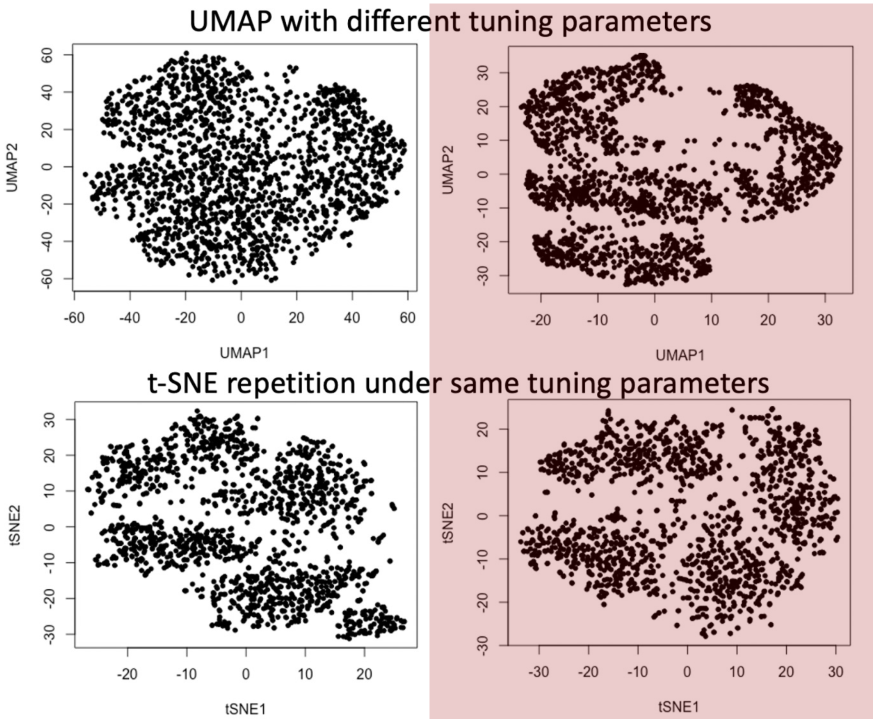
[Bastidas-Ponce *et al.*, 2019, *Development*]

Main Problems with t-SNE and UMAP: instabilities

Instability to
graph connectivity:
not enough separation

Instability to
initialization:
too much separation

“ground truths”



These **instabilites** cause **cell identification problems.**

Life is dynamic, as in normal development and disease progression

Smooth embedding structure captures dynamic life, but is a bigger problem for t-SNE and UMAP

“P” or reality-check in NESS is established through domain knowledge and/or simulation studies

NESS builds also on PCS-related theory

Cai, T. T., & Ma, R. (2022, *JMLR*). "Theoretical foundations of t-sNE for visualizing high-dimensional clustered data":

- t-SNE uses approximate power iterations to create clusters
- t-SNE achieves cluster consistency under mild conditions
- **Clusters are randomly located under random initialization**

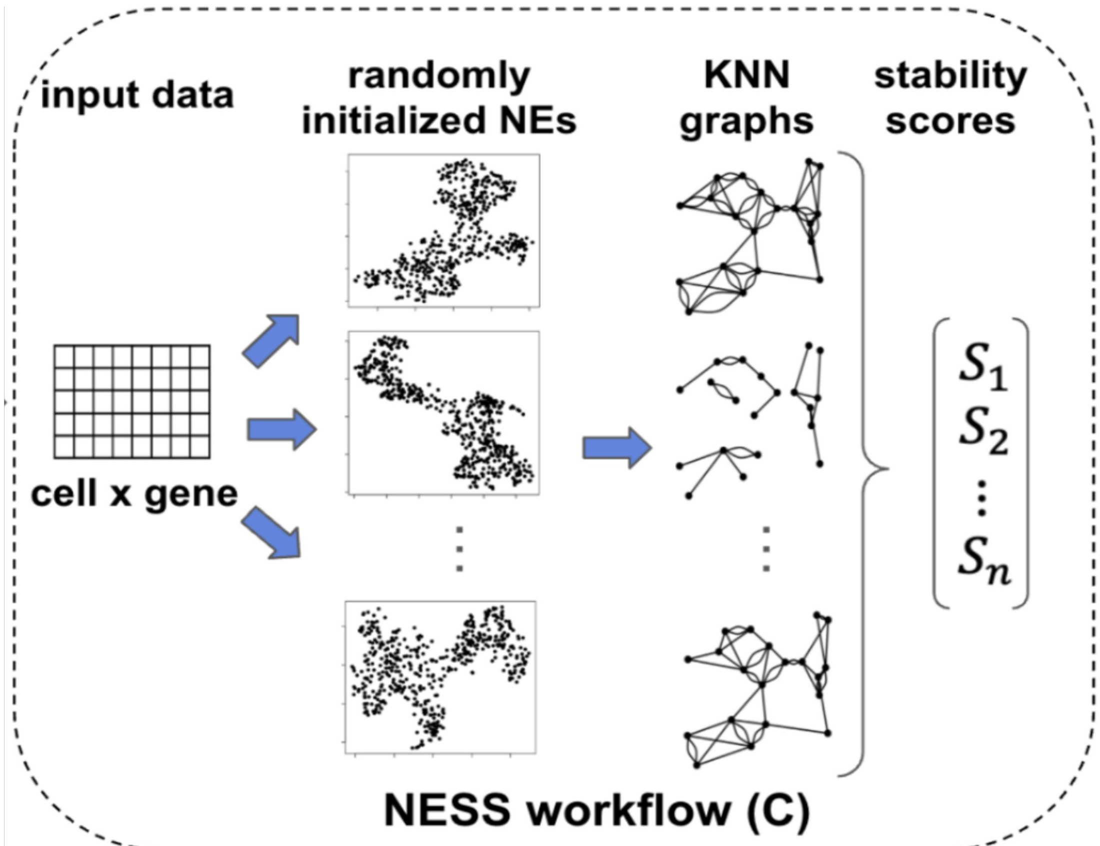
Liu, Z., Ma, R., & Zhong, Y. (2025, *Nat. Comm*). "Assessing and improving reliability of neighbor embedding methods: a map-continuity perspective"

t-SNE embedding map is intrinsically discontinuous for clustered data

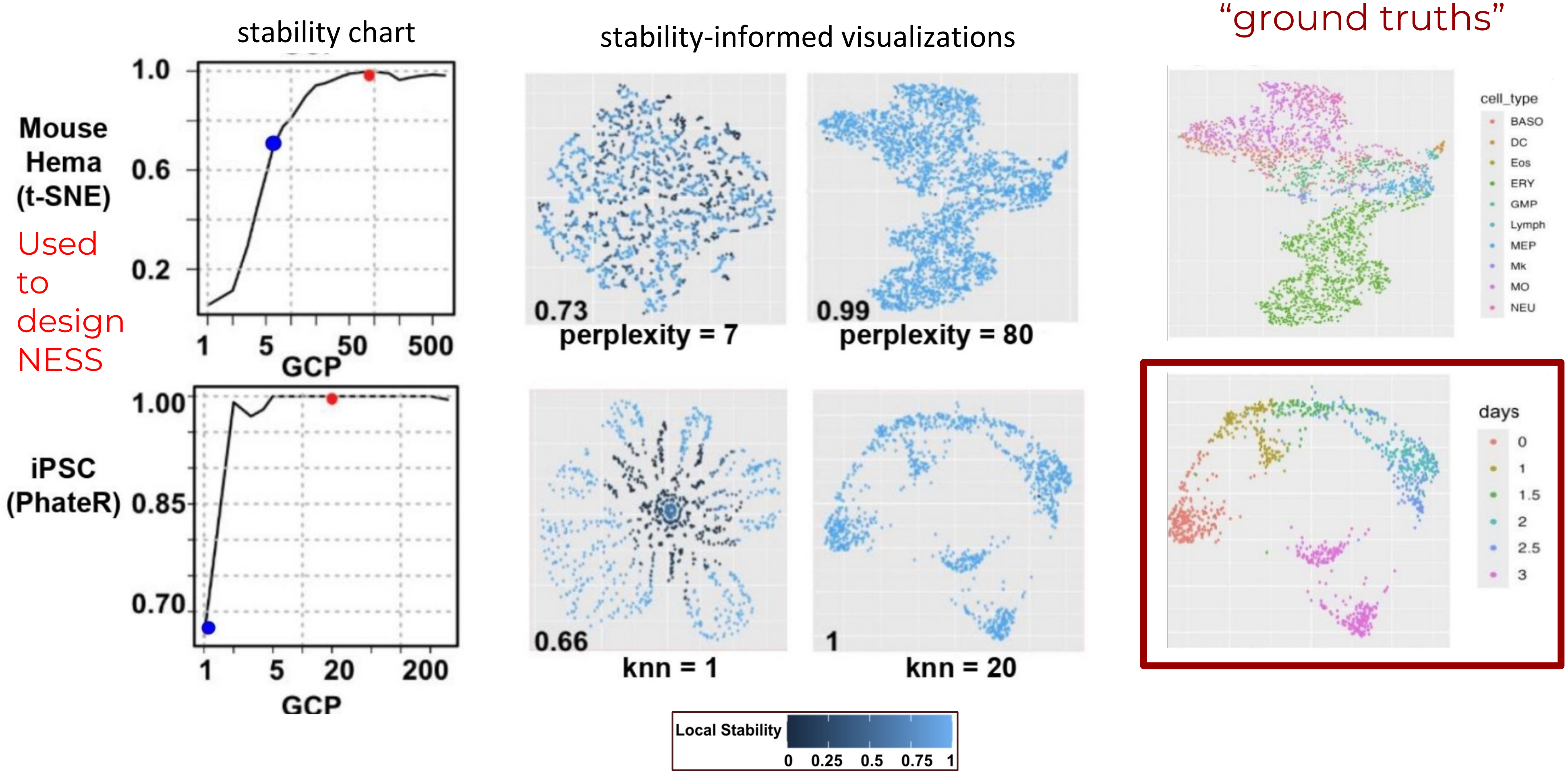
Our proposed NESS for smooth structures (S)

NESS develops a stability score for each data point that takes into account both “S” and “P” (to maintain important global structure)

building on both empirical evidence and PCS-related theory

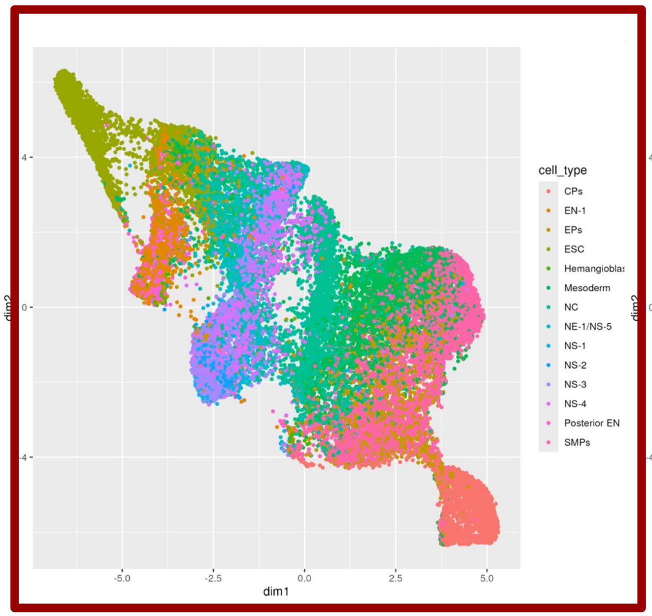


NESS selects best embeddings w. stability score

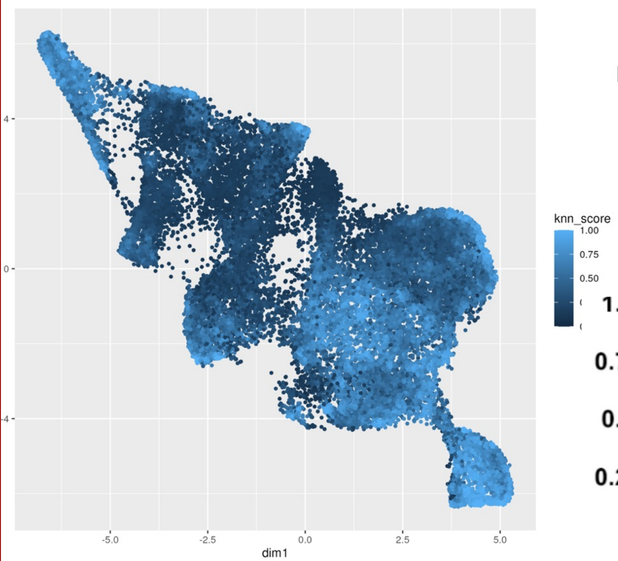


NESS stability score identifies transitional cell states in embryogenesis

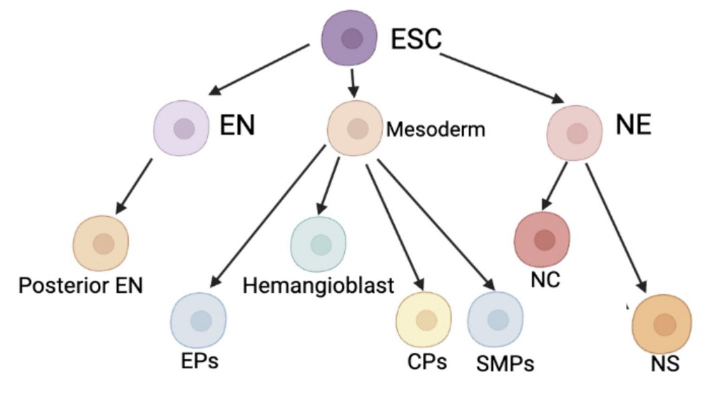
Mouse embryoid stem cell differentiation
scRNA-seq data (31029 cells x 19112 genes)



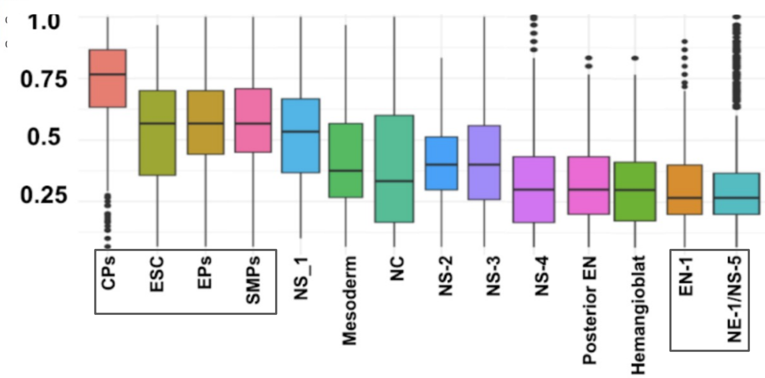
"Ground truth"



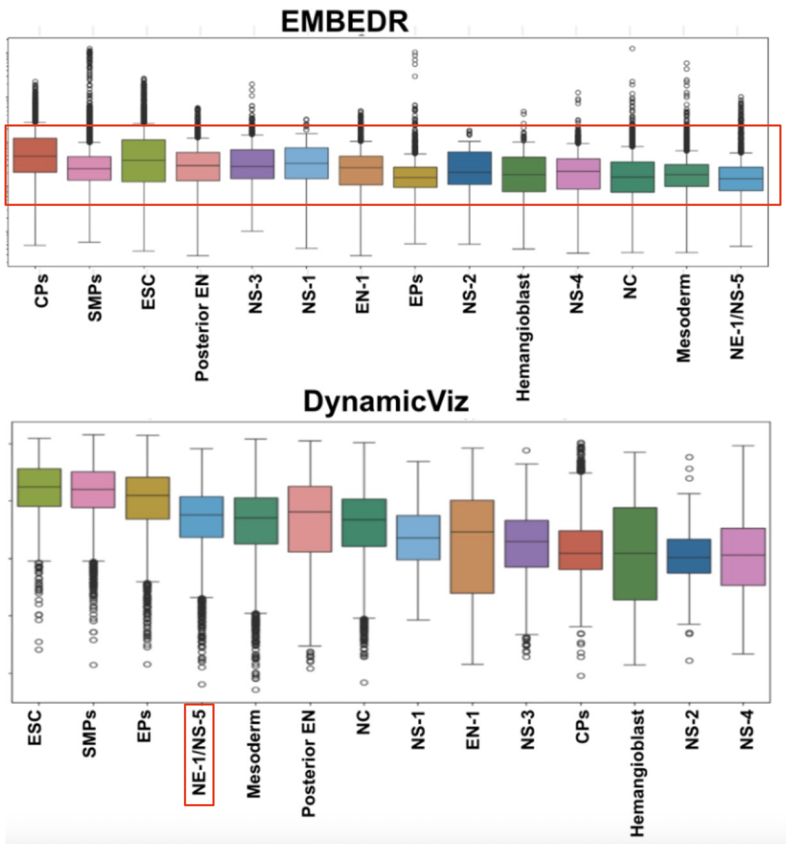
NESS UMAP with
Stability Scores



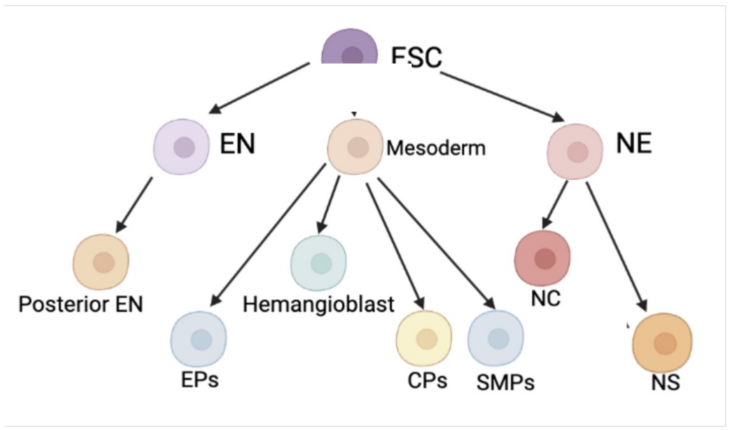
stability score grouped by cell states



Comparison methods: EMBEDR and DynamicViz



EMBEDR is not as differentiating.
DynamicViz gets one stage wrong.



4 other validated use cases of NESS in the paper

- NESS helps identify key genes associated with **human induced pluripotent stem cells (iPSC) differentiation**
- NESS identifies transitional cell states in **murine intestinal organoid development**
- NESS resolves distinct neuronal subpopulations during **embryoid formation**
- NESS reveals transcriptional dynamics during **mammalian spermatogenesis** and **neurogenesis**

PCS-Uncertainty Quantification (UQ)

PCS UQ for regression and classification

R. Barter*

PCS regression perturbation interval (Ch. 13 of Yu-Barter book) and classification and comp. efficient PCS UQ for deep-learning (new)

Abhineet Agarwal* **Michael Xiao*** Boyu(Boris) Fan Omer Ronen

* denotes equal contribution

"PCS-UQ: Uncertainty Quantification via the Predictability-Computability-Stability Framework"
https://arxiv.org/abs/2505.08784 (under review)

PCS Perturbation Interval

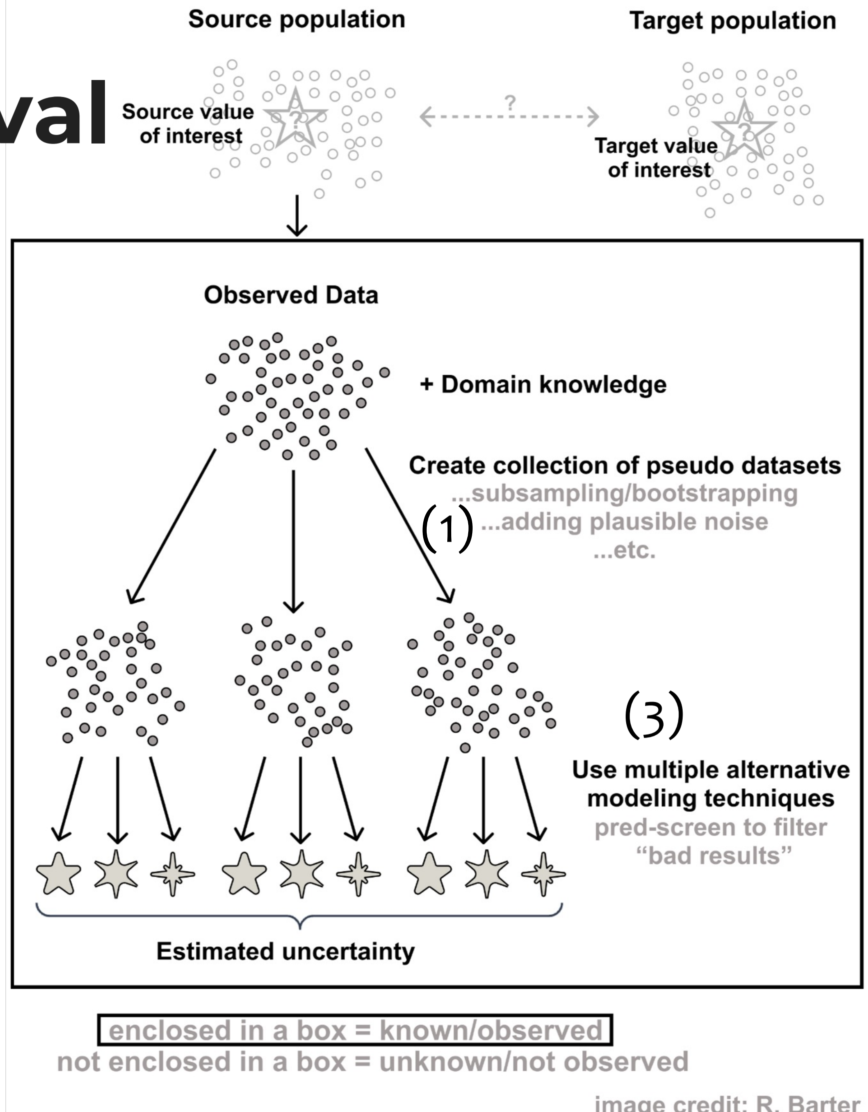
VDS book considers three sources of uncertainty in DSLC from

1. **Data collection process (existing)**

1. **Data cleaning choices (new)**

1. **Pred-checked modeling choices (new)**

PCS UQ relies on a finite collection of pseudo datasets or values of interest, as in bootstrap.



Comparison to classical conformal

PCS

- Data cleaning uncertainty allowed
- Uses **multiple** ML algorithms
- "**Pred-check**" to screen bad alg.
- "Local" calibration via stability (**bootstrap,** multiple pred-checked algorithms, data cleaning)
- **Multiplicative length calibration** to achieve empirical coverage on validation set under assumption that validation set is a good proxy to future data

Split Conformal

- No data cleaning uncertainty
- Uses **one** ML model
- No explicit model checking
- Global calibration using residuals

  (no bootstrap, no multiple alg.)
- **Constant length calibration**, to achieve coverage if exchangeability assumption holds between future data and current data

Classical Conformal Methods for Reg.

- **(Split) Conformal Inference:**

  Inductive confidence Machines for regression.
  Authors: Papadopoulos, H.,Proedrou, K.,Vovk ,V.,& Gammerman, A
  Journal: Machine Learning (2002)

  Distribution-free predictive inference for regression

  Authors: Lei J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., & Wasserman, L.

  Journal: *Journal of the American Statistical Association (2018)*

- **Studentized Conformal Inference:**

  Distribution-free predictive inference for regression

  Authors: Lei J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., & Wasserman, L.

  Journal: *Journal of the American Statistical Association (2018)*

PCS Hyper-parameters

- Candidate models:
  - Linear: OLS, Lasso, Ridge, ElasticNet,
  - Bagging: Random Forests (RFs), ExtraTrees
  - Boosting: XGBoost, AdaBoost,
  - DL: Multi-layer Perceptrons (1 hidden layer)

- **Top-3 best** performing models across **1000** bootstraps

How were hyper-parameters chosen?

- Candidate models: Popular choices across widely-used model classes
- Top-3 & 100 bootstraps chosen via **synthetic simulations & 5 pilot datasets**

No contamination

Conformal Hyper-parameters

- Candidate models: OLS, Lasso, Ridge, ElasticNet, Random Forests (RFs), XGBoost, ExtraTrees, Multi-layer Perceptrons (1 hidden layer)
- Try **all** candidate models and use **best** one for conformal
- For majority, try **all candidate models**

Critical importance of benchmarking for the success of ML and AI

Empirical **benchmark datasets, continuously enriched,** are a cornerstone for ML/AI algorithm development and evaluation…

Theory come later, and partially

Critical importance of benchmarking for the success of ML and AI

Empirical **benchmark datasets, continuously enriched,** are a cornerstone for ML/AI algorithm development and evaluation…

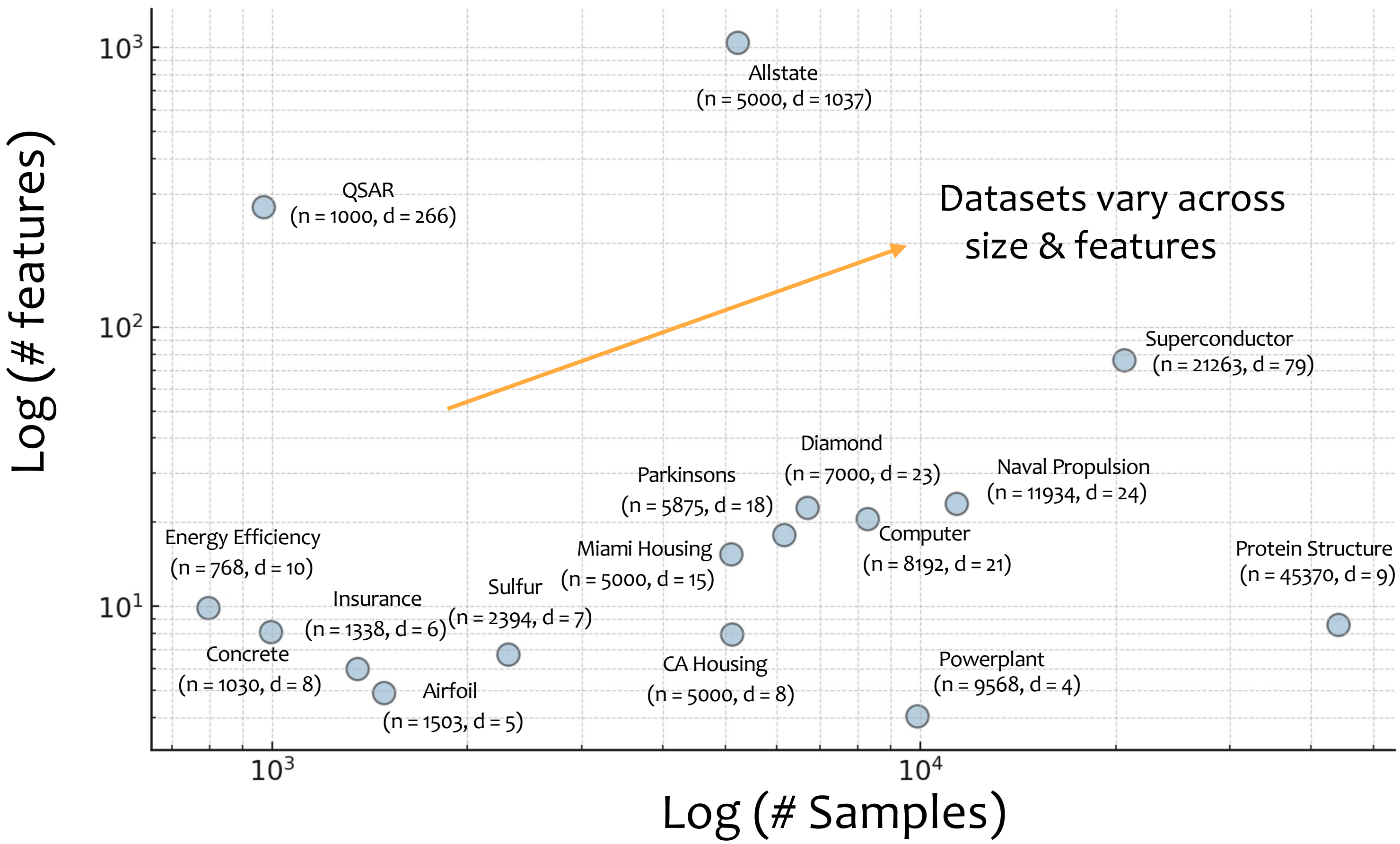
Theory come later, and partially

Benchmark datasets are imperfect, but are overall much better as **reality-check evidence (“P” in PCS)** than a small number of (non-standard, possibly cherry-picked) datasets and analytical simulations in most statistical papers.

Benchmarking is now an “emerging science” (Hardt, 2025)

17 Real-World Regression Datasets

(no data cleaning uncertainty)



PCS-UQ: more realistic with pred-checked multiple models

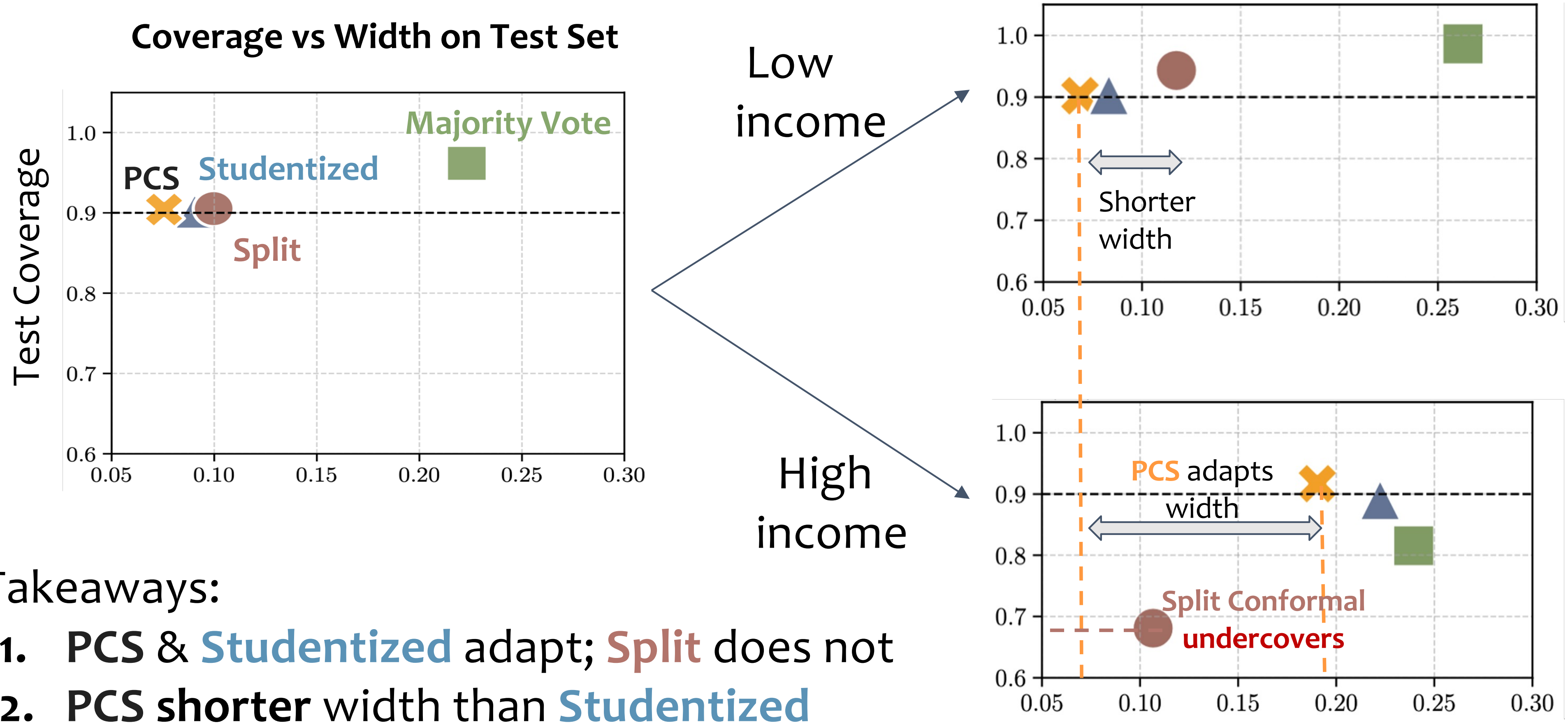
“Experiments across **17 regression datasets** show that **PCS-UQ achieves the desired coverage** and **reduces width over classical conformal approaches by ≈ 20%, marginally. Better performance over subgroups (coverage and width) than classical conformal methods.** Theoretically, we show a modified PCS-UQ algorithm is a form of split conformal inference and achieves the desired coverage with exchangeable data.” **Agarwal et al (2025)** https://arxiv.org/pdf/2505.08784

PCS-UQ: more realistic with pred-checked multiple models

“Experiments across **17 regression datasets** show that **PCS-UQ achieves the desired coverage** and **reduces width over classical conformal approaches by ≈ 20%, marginally. Better performance over subgroups (coverage and width) than classical conformal methods.** Theoretically, we show a modified PCS-UQ algorithm is a form of split conformal inference and achieves the desired coverage with exchangeable data.” **Agarwal et al (2025)** https://arxiv.org/pdf/2505.08784

Experiments use clean data sets so don’t have the data cleaning step. Ch. 13 of the VDS book deals with the data cleaning step.

Dataset: Miami Housing (n=13932, d=28)



Takeaways:

1. **PCS** & **Studentized** adapt; **Split** does not
2. **PCS shorter** width than **Studentized**

PCS-UQ: Multi-Class & DL Results

Takeaways across datasets

- **PCS** and **conformal** achieve **desired coverage** across 6 tabular datasets
- **PCS** reduces width over best conformal by **~20%**

Takeaways across datasets

- **PCS** and **conformal** achieve **desired coverage** across 6 tabular datasets

- **PCS** reduces width over best conformal by **~20%**

Deep-learning

- Provide approximation schemes to overcome computationally expensive bootstrap training

- **PCS approximation schemes out-perform conformal variants**

Deep-Learning Experiments

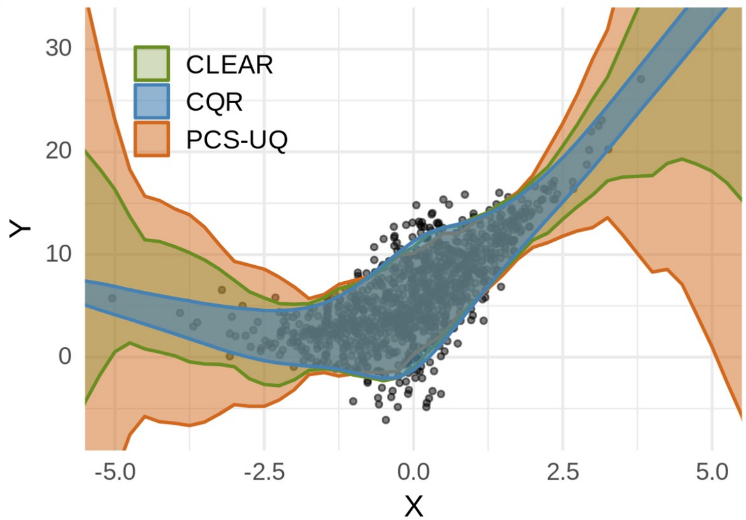
Method/Dataset		CIFAR 100		ImageNet Small		Places365 Small	
		Av. Size	Time (min)	Av.Size	Time (min)	Av. Size	Time (min)
APS		6.8	2	14.4	3	16.8	3
RAPS		6.5	2	10.6	3	11.2	3
TopK		8.5	2	12	3	13	3
PCS	**Original**	**3.7**	350	**8.3**	2000	**8.8**	2500
	Dropout	4.4	4	9.8	5	9.8	4
	Noise	4.2	3	9.4	5	9.6	3
	Embedding	**4.1**	10	**9.1**	25	**9.3**	30

Takeaways:

1. **Original PCS smallest size**
2. **PCS approximation schemes produce small sets & are efficient**

CLEAR: Calibrated Learning for Epistemic and Aleatoric Risk (Azizi et al., 2025) https://arxiv.org/abs/2507.08150

PCS-UQ reduces width by 19% over CQR (over 17 datasets)
(CQR: conformalized quantile regression)



CLEAR combines PCS-UQ and CQR;
scales them by two calibration constants.

➔ Width reduction:

- 7% compared to PCS-UQ
- 25% compared to CQR

Co-authors

I. Azizi*

J. Bodick*

J. Heiss*

* denotes equal contribution

Consistent ranking of CQR, PCS-UQ and CLEAR across 3 candidate algorithm sets – “S” analysis

- Variant (a):
  - PCS-UQ reduces width compared to CQR by **5.86%**
  - CLEAR reduces width compared to PCS-UQ by **13.34%**
  - CLEAR reduces width compared to CQR by **20.80%**
- Variant (b):
  - PCS-UQ reduces width compared to CQR by **22.73%**
  - CLEAR reduces width compared to PCS-UQ by **13.10%**
  - CLEAR reduces width compared to CQR by **32.03%**
- Variant (c):
  - PCS-UQ reduces width compared to CQR by **19.25%**
  - CLEAR reduces width compared to PCS-UQ by **7.46%**
  - CLEAR reduces width compared to CQR by **25.53%**

Conformal Theory

We made connection to conformal inference

- Multiplicative calibration step in PCS-UQ can be viewed as new form of conformal inference

- Implies modified PCS-UQ has theoretically valid coverage under exchangeability.

- PCS-UQ has two other steps (Pred-check and bootstrap) that underlie the better performance.

Outline of talk

1. Statistics needs to adapt to the AI age
2. VDS with core PCS principles is a frontier of statistics
3. VDS success stories …
4. Theory and processes of productive theoretical research
5. PCS current directions and resources

Theoretical Foundations of PCS

What is called PCS-related theory?

Explicit considerations of stability or sensitivity and/or computation budget in an algorithm or procedure or under multiple generative models. “P” is covered in the generative model which is chosen for capturing reality and analytical tractability.

Theoretical foundations for PCS

Stability is a key concept in stats/ML theory

Bernoulli **19**(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

- **Central Limit Theorem**
- **Uniform stability in ML (that bounds generalization error)**
- **Random matrix results**
- **Sensitivity bounds in causal inference**

Stability is more central than Gaussian distribution to stat/ML.

Recent theoretical PCS-related works

- Behr et al (2022) shows that a theoretical version of iRF is model selection consistent under additive Boolean regression models.

- Cai and Ma (2022) on understanding effect of initialization in t-SNE cluster locations, and consistency results for clustered data

- Trillos et al (2025) studies a general “insensitivity” quantity of an estimator, and establishes a theory of sensitivity based on Weissenstein geometry, similar in spirit to the Fisher-Rao geometry.

PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCESS

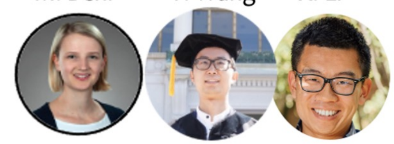
Provable Boolean interaction recovery from tree ensemble obtained via random forests

2022

Merle Behr[a,1], Yu Wang[a,1], Xiao Li[a], and Bin Yu[a,b,c,2]

Co-authors

M. Behr Y. Wang X. Li



- **New Local Spiky Sparse (LSS) model:** linear combination of Boolean interactions as regression function

$$E(Y|X) = \beta_0 + \sum_{j=1}^{J} \beta_j \prod_{k \in S_j} \mathbf{1}(X_k \gtrless \gamma_k),$$

- Theoretical tractable version of siRF: **LSSFind** based on Depth-Weighted Prevalence (DWP) computed from an RF tree ensemble
- Interaction discovery consistency of LSSFind under regularity conditions
- Simulation studies

JMLR (2022)
https://jmlr.org/papers/volume23/21-0524/21-0524.pdf

Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data

T. Tony Cai TCAI@WHARTON.UPENN.EDU
Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Rong Ma RONGM@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Editor: Ji Zhu

https://arxiv.org/abs/2511.07414

Wasserstein-Cramér-Rao Theory of Unbiased Estimation

Nicolás García Trillos[1], **Adam Quinn Jaffe**[2] **and Bodhisattva Sen**[2]

[1] *Department of Statistics, University of Wisconsin Madison, WI, e-mail:* garciatrillo@wisc.edu

[2] *Department of Statistics, Columbia University, New York, NY, e-mail:* a.q.jaffe@columbia.edu; b.sen@columbia.edu

Open Theoretical problems

Open problems motivated by PCS

Mathematical results at the modeling stage of a DSLC

- Conjecture: consistent model selection with **positive probability** of “LSSFind” when features are dependent and/or interaction sets in LSS model overlap.

- Syntheses of different notions of stability, and their relationships and connections with generalization, causality, transfer learning.

Open problems motivated by PCS

- Are combinations of good “algorithms” are desirable than a single algorithm under suitable probabilistic data generation models?

- We have some preliminary results on PCS-UQ using multiple classes of methods (joint work with M. Xie, T. Luo, A. Ozerov).

Open AI problems motivated by PCS

In a DSLC,

- What are reasonable models and specifications for the data cleaning step?
- For the EDA (exploratory data analysis) step?
- How about verified AI systems, which go beyond a DSLC?

Productive ways for impactful theory

- Sequential: substantial **data evidence first**, **math late**r:
  PCA, Bootstrap, Lasso, RF, Boosting, SVM, Spectral clustering, DL, iRF, t-SNE, …

These algorithms have been or could be analyzed under multiple reasonable data generation models ("S" for theory)

- Combined: **math results** under **appropriately simplified** models, and **data results** from multiple (realistic) simulation models and serious real-world benchmark datasets (e.g. muP and LoRA+ in DL).

Outline of talk

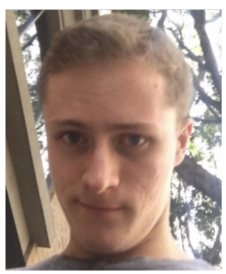
1. Statistics needs to adapt to the AI age
2. VDS with core PCS principles is a frontier of statistics
3. VDS success stories …
4. Theory and processes of productive theoretical research
5. PCS current directions and resources

PCS guides synthetic stimulus design (hypothesis generation) for fMRI experiments using LLMs

A generative framework to bridge data-driven models and scientific theories in language neuroscience

Richard Antonello[1†], Chandan Singh[2†], Shailee Jain[3,a], Aliyah Hsu[2,4], Jianfeng Gao[2], Bin Yu[4,5,6*‡], Alexander Huth[1,7*‡]

Main co-authors:



R. Antonello

C. Singh

A. Huth

https://arxiv.org/html/2410.00812v1 (revising for Nature Neuroscience)

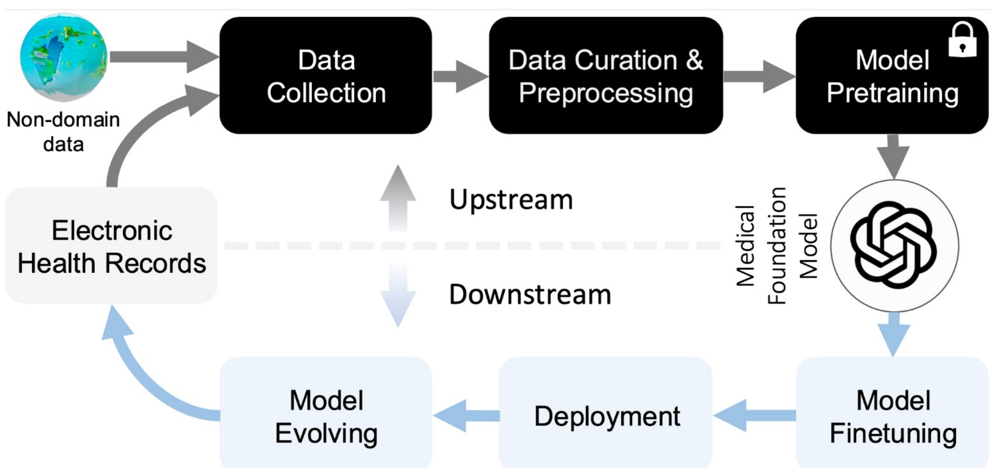
“Veridical data science for medical foundation models” (Alaa and Yu, 2024, https://arxiv.org/abs/2409.10580)

How is the foundation model life cycle (FMLC) different?

Grounding (“P”) and consistency (“S”) are two key issues in LLMs.

“Black-box” upstream process

Upstream:



Downstream:

Downstream process constrained by the black-box upstream process

Ahmed Alaa

PCS is applicable for synthetic data generation via genAI

Reality and stability checks are central problems in synthetic data generation via genAI, corresponding to grounding and consistency considerations there.

Detailed PCS examinations of synthetic data generation by genAI and PCS recommendations are PCS research problems.

For example, for text-to-image data generation,

what reality checks are feasible and reliable? And for what purpose?

what stability checks are necessary and feasible? And for what purpose?

Other PCS projects: all collaborative

Broad impact projects:

- **Implementing PCS on DS platforms (co-PI on proposal)**
- Three PCS video modules for HS DS, Math and Science classes
- Cryo-EM competition led by Flatiron Institute
- Veridical AI theme in new Norwegian AI Center (TRUST)

Research projects: PCS is applied

Interpretable DL

genomics using genAI

medical AI using LLMs

causal inference

interactive PCS data analysis platform using LLMs ("StatGenie")

AI safety (UK AISI grant)

PCS/VDS Resources

Software to address "C" in PCS

Veridical Flow: (v-flow) PCS-style data analysis made easy!
(Duncan et al, 2022, JOSS)

simChef: PCS-style simulations made easy!
(Duncan et al, 2024, JOSS)

Merits: simulation guidelines (Elliott et al 2025, JCGS (to appear))

More at my website: https://binyu.stat.berkeley.edu/ – click on code on top

MERITS of a high-quality simulation study

(Elliott et al, 2025): PCS-inspired simulation guidelines to address “C”

(Computability) **Modular:** Written in self-contained and logically partitioned segments of code.

(Computability) (Predictability) **Efficient:** Streamlined computationally and conceptually.

Realistic: Faithful to the physical world.

Intuitive: Sensible to the intended audience and, in a general sense, to a reasonably comprehensive readership.

(Stability) **Transparent:** Documented thoroughly and candidly.

Stable: Reproducible/replicable, and externally valid.

C.F. Elliott

T. Tang

J. Duncan

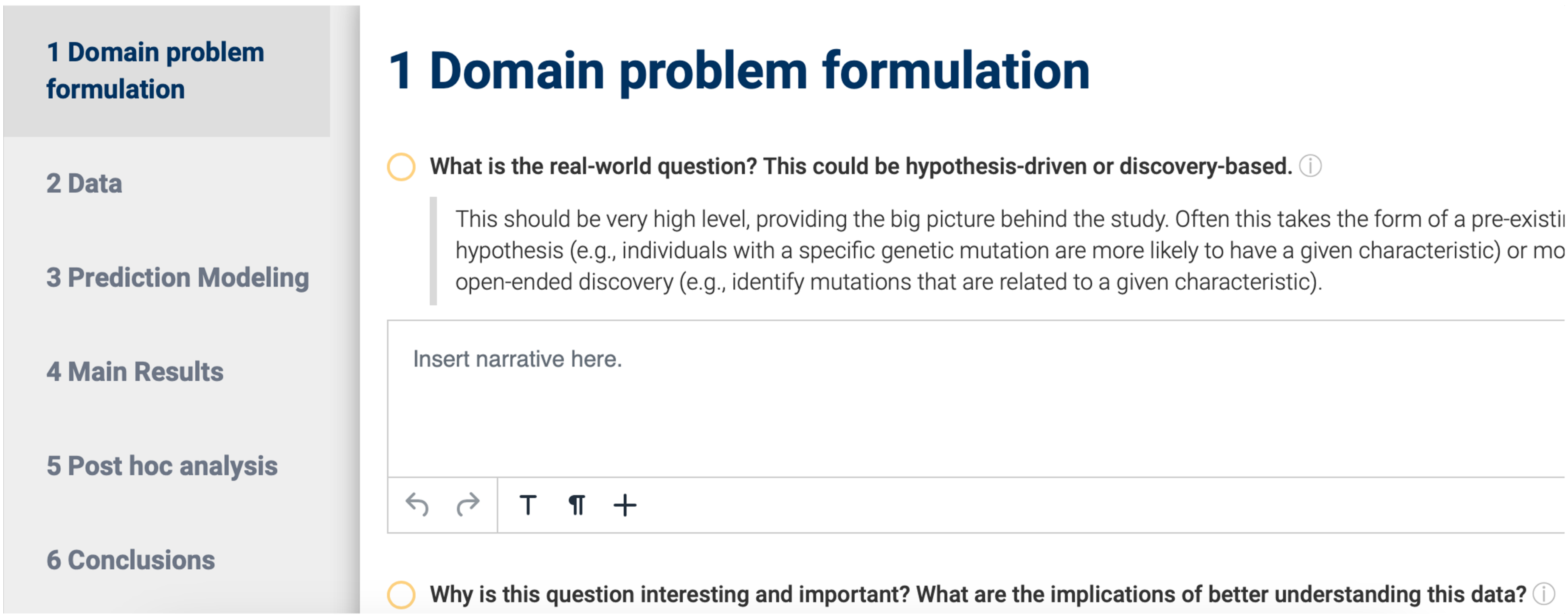
M. Behr

K. Kumbier

PCS documentation

T. Tang A. Kenney

Template at my website: [\https://yu-group.github.io/vdocs/PCSDoc-Template.html](https://yu-group.github.io/vdocs/PCSDoc-Template.html)



PCS workflow paper makes practicing VDS easy

Accepted by a special issue on *Workflow for Applied Data Analysis* edited by A. Gelman

PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

PCS Workflow for Veridical Data Science in the Age of AI

Zachary T. Rewolinski[1] and Bin Yu[1,2,3]

[1] Department of Statistics, UC Berkeley
[2] Department of EECS, UC Berkeley
[3] Center for Computational Biology, UC Berkeley

First link at Bin's website front page https://www.stat.berkeley.edu/~yugroup/pubs/pcs_workflow-6.pdf

Future VDS workshops: check Bin's website

Recent and upcoming workshops:

May 29, 2026, Veridical AI Workshop in Paris (upcoming)

July 11, 2025, QB3, UC Berkeley

Veridical Data Science in Biology

June 20, 2025, Sapienza University, Rome

Rome Workshop on Veridical Data Science

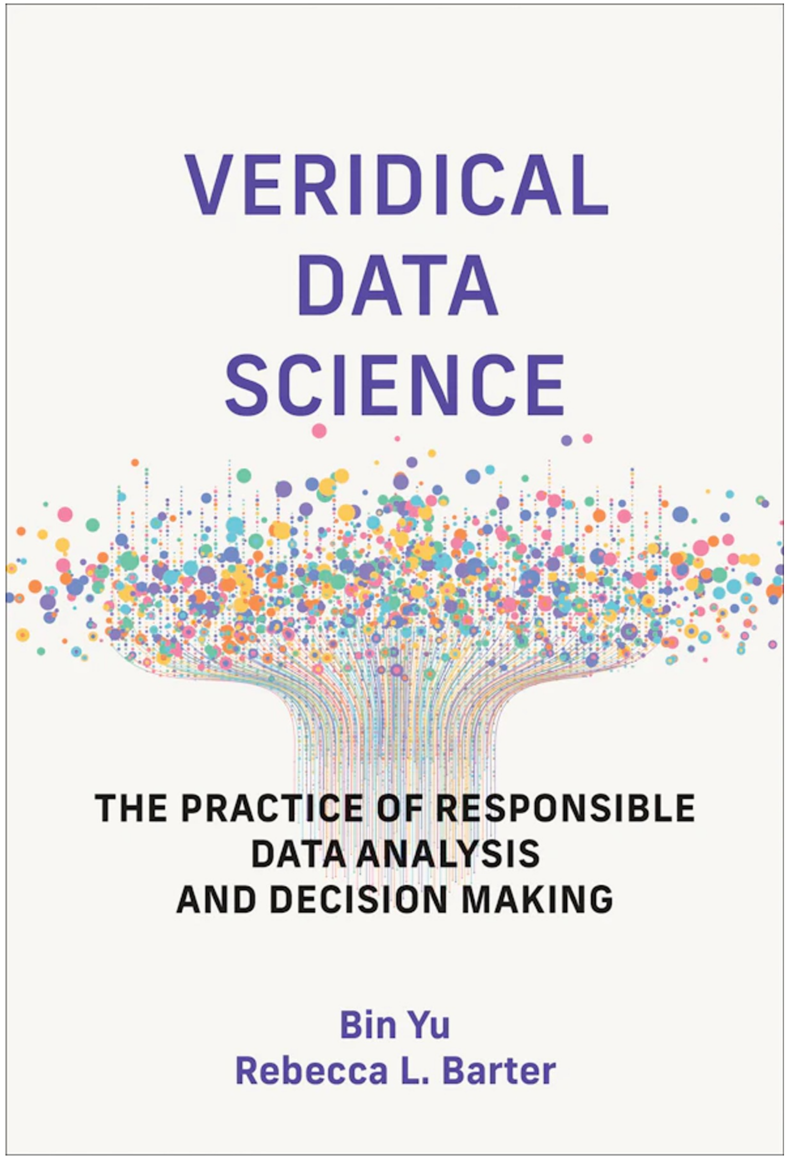
May 31, 2024, at UC Berkeley

Inaugural Berkeley-Stanford Workshop on Veridical Data Science at UC Berkeley (May 31, 2024) (talk videos available)

Bin Yu   Rebecca Barter

Free version
vdsbook.com



MIT Press · '24
(ML Series)
Paper Book

PCS recommendations for any data project

Concretely and "**in-context**",

- For problem formulation, try more than one
- For data cleaning, at least keep two copies of cleaned data
- For EDA and reporting, try different graphics parameters, and data perturbations
- For modeling step, it values multiple good or pred-checked models, moving away from "true" model framing; and it provides principled ways to aggregate good models (incl. those from data perturbations)

In a nutshell, PCS is an evolving frontier of statistics

- PCS principles work for a DSLC or AI workflow and embrace pluralism

- Documentation is an integral part

- Many open areas: multi-modality, dynamic data, PCS theory, …

PCS is essential to AI startup Traversal

"**Traversal** builds an AI Site Reliability Engineer (AI-SRE) that helps companies like **American Express, Pepsi, and DigitalOcean** diagnose outages by searching petabytes of telemetry and code to deliver fast, trustworthy root causes and resolutions.

In these high-pressure settings—where signals are very weak and the search space is enormous—the PCS framework is essential.

First get accuracy (P).

Then speed (C).

Then stability (S).

That's how we develop at Traversal AI."

– Raaz Dwivedi, co-founder, CTO

Assistant Prof. Cornell Tech

FORTUNE Article in 6/25

Traversal emerges from stealth with $48 million from Sequoia and Kleiner Perkins to reimagine site reliability in the AI era

By **Allie Garfinkle**

Senior Finance Reporter And Author Of Term Sheet

https://fortune.com/2025/06/18/traversal-emerges-from-stealth-with-48-million-from-sequoia-and-kleiner-perkins-to-reimagine-site-reliability-in-the-ai-era/

Parting message

Statistics was built to pursue truth under uncertainty.

VDS with PCS extends this mission to the complexity of the AI age as systems science.

You are invited to advance together VDS – a frontier of statistics!

Published papers

B. Yu (2013). Stability. *Bernoulli.*

S. Basu, K.Kumbier, B. Brown, B. Yu (2018), Iterative random forests. *PNAS.*

B. Yu and K. Kumbier (2020). Veridical data science. *PNAS.*

B. Yu (2023) What is uncertainty in today's practice of data science? *J.Econometrics.*

B. Yu and R. Barter (2024). Veridical data science: the practice of responsible data analysis and decision making. *MIT Press* (online free version at vdsbook.com).

T. Tang, Y. Zhang, A. Kenney, ..., B. Yu, Arul Chinnaiyan (2025). A simplified MyProstateScore2 for high-grade prostate cancer. *Cancer Biomarkers.*

Q. Wang*, T. M. Tang*, ..., B. Yu*, E. Ashley* (2025). Epistasis regulates genetic control of cardiac hypertrophy. *Nature Cardiovascular Research* (Code) (PCS documentation)

Recent papers

Z. Rewolinski and B. Yu (2025) PCS workflow for veridical data science in the age of AI . https://arxiv.org/abs/2508.00835 (accepted *Philosophical Trans. A*)

A. Alaa and B.Yu (2024) Veridical Data Science for Medical Foundation Models. https://arxiv.org/abs/2409.10580

A.Agarwal, M. Xiao, R. Barter, B. Fu, O. Ronnen and B. Yu (2025). PCS-UQ: Uncertainty Quantification via the Predictability-Computability-Stability Framework https://arxiv.org/abs/2505.08784 (submitted)

I.Azizi, J. Bodik, J. Heiss, B. Yu (2025). CLEAR: Calibrated Learning for Epistemic and Aleatoric Risk. (submitted)

R. Ma*, X. Li, J. Hu, B. Yu* (2025). Uncovering smooth structures in single-cell data with PCS-guided neighbor embeddings. https://arxiv.org/abs/2506.22228 (submitted)

Thanks to all my collaborators!

Thank you!

Questions?

NSF

CHAN ZUCKERBERG BIOHUB

NIH

SIMONS FOUNDATION