# Veridical Data Science
# Towards Trustworthy AI


**Bin Yu**

**Statistics, EECS, Center for Comp. Bio., Simons Institute**


Dean's Distinguished Women in Mathematics, Statistics and Computer Science Lecture Series
David Sprott Distinguished Lecture Series, Statistics
University of Waterloo, Canada

Sept. 23, 2025

# Words by Prof. David Sprott

*"L. You then returned to Toronto.*

S. Yes, it was through Penrose and Sheppard, the actuarial professor, who knew Stokes, the head of psychiatry, that I came back in the position of clinical teacher of psychiatry at the University of Toronto. I think t**his experience must have contributed a great deal to my outlook on science and statistics**, because I could see what their researchers were doing, and I had a feeling for what they ought to be doing, and it seemed to me that the **statisticians weren't answering the right questions**."

— "a conversation recorded December 19, 1988 at the University of Waterloo originally appeared in *Liaison* Vol. 3, No. 2, February 1989."

https://ssc.ca/en/profile/a-conversation-david-a-sprott

# The in-context approach for stats/DS/ML research

*Start with a* ***"practical problem"****,*

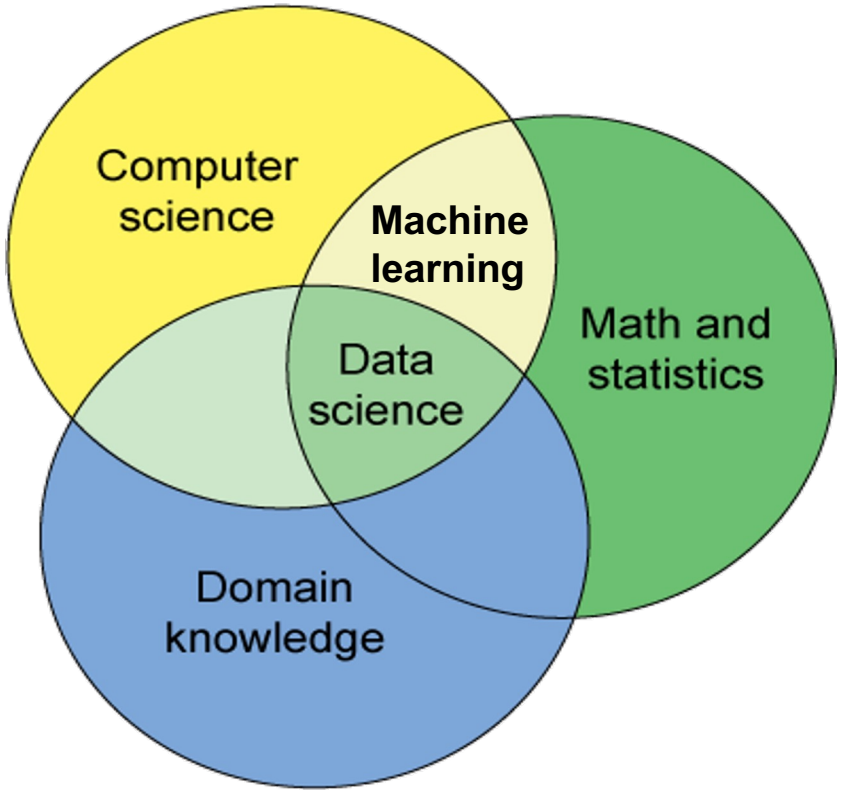
*and develop* ***"a true feeling for, and insight into"*** *it*

*before bringing in*

***"a high level of mathematical talent of the most abstract sort"****.*

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

# Data science (DS) is a pillar of ML & AI




*Conway's Venn Diagram*

# *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 in Spanish "**verified truth**".

- In VDS, "**veridical**" hinges on both definitions:

    1. Seeking truthful data-driven conclusions.
    2. Transparent (truthful) PCS-driven data-science life cycle

# Veridical data science (VDS)

**Veridical Data Science (VDS)** aims at building a realistic philosophical and conceptual framework for practicing reproducible data science, including a rigorous documentation in context.

It is built on three first principles: predictability, computability, stability (PCS).

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

# Part 1: Why do we need VDS?

# Reproducibility crisis (early 2010's)

"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

# PNAS article in 2022

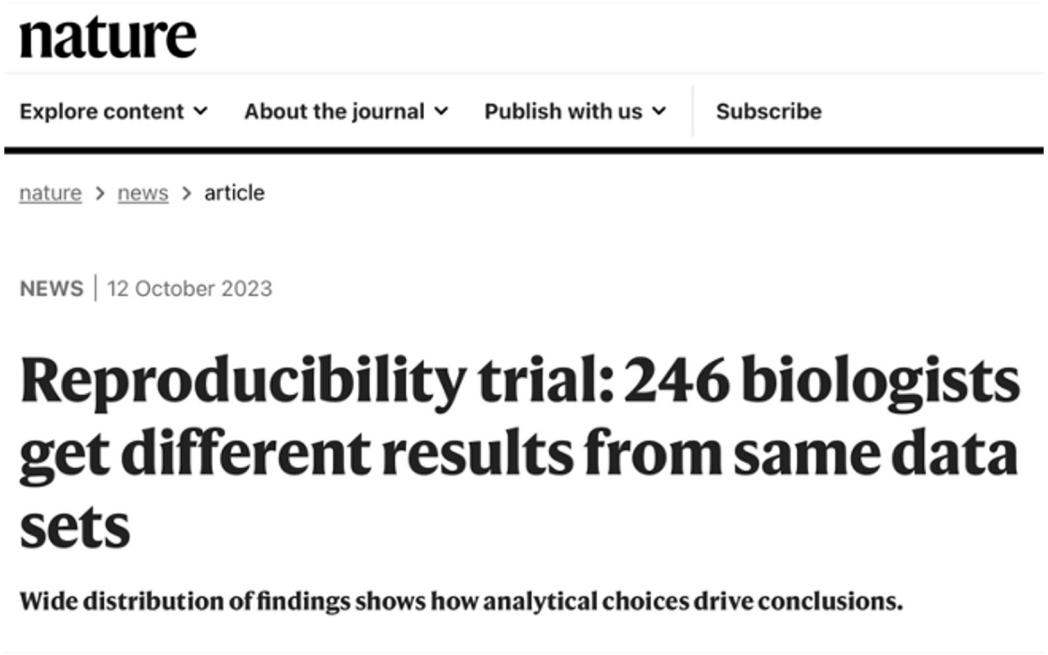
## Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

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.**"

# Nature article in 2023



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › news › article

NEWS | 12 October 2023

## Reproducibility trial: 246 biologists get different results from same data sets

**Wide distribution of findings shows how analytical choices drive conclusions.**

**Gould et al (2023)**: "Same data, **different analysts: variation in effect sizes due to analytical decisions** in ecology and evolutionary biology

# Uncertainty from data cleaning

TA: O.Ronen

- Students developed models to predict the risk of Traumatic Brain Injuries (TBI) for pediatric patients (in Stat 215A, Fall 2021)

- Three groups of students, each team with a UCSF medical doctor, worked on the problem independently, using the same raw data and with the same data cleaning guidelines

In terms of sensitivity, uncertainty (10%) from data cleaning choices is similar to uncertainty from bootstrap samples from each cleaned dataset.

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

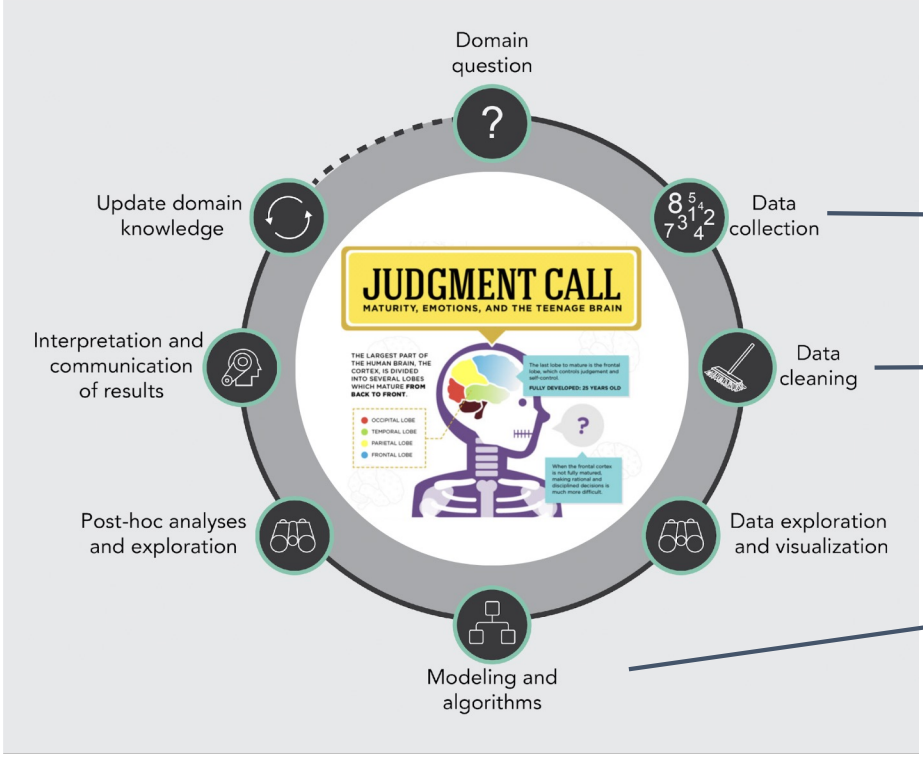
C. Singh A. Kornblith

# Uncertainty quantification (UQ) builds trust in DS/AI

- Current stats/DS practice only considers uncertainty from generative stochastic models that undergo often very **limited empirical checking.**
- In a DSLC, human judgement calls are an important source of uncertainty
- **Current stats/DS practice often significantly underestimates uncertainty, leading to unnecessary false discoveries with wasted downstream resources and possible harms.**

**Trustworthy uncertainty quantification is indispensable.**

# Data Science Life Cycle (DSLC)




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

Image credits: R. Barter and toronto4kids.com

Uncertainty across the DSLC

What choices were made while collecting data?

How was the data cleaned?

Modeling choices

**A DSLC creates uncertainty in every step!**

# Uncertainty Quantification

via

# Predictability-Computability-Stability (PCS)

# Rest of the talk

- Brief intro to PCS framework and documentation
- PCS Uncertainty Quantification
- Experimental Evaluation
- PCS Current Directions

  Book, softwares, document template, on-going projects…

# Part 2: PCS Framework

# PCS framework: one culture

Yu and Kumbier (PNAS, 2020)

Three principles of data science:

(**P**)redictability [ML and Stats]

(**C**)omputability [ML]

(**S**)tability [Stats, control theory, numerical analysis]

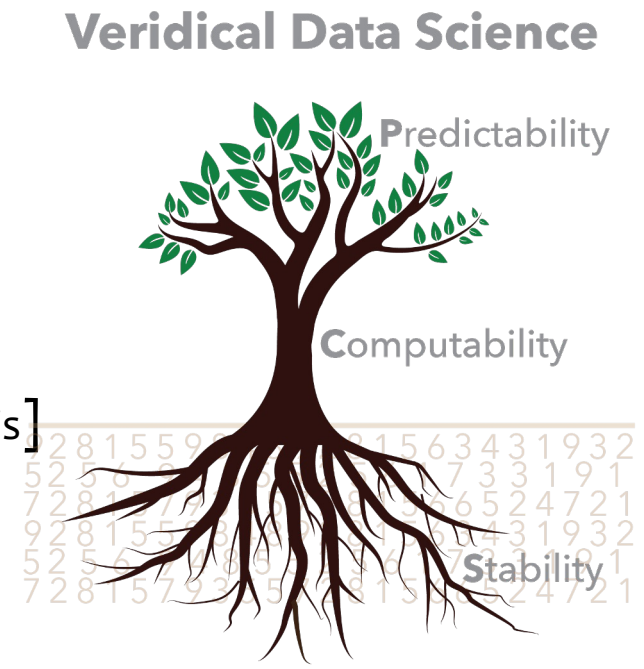



Image credit: R. Barter

# PCS

**Unifies, synthesizes,** and **expands** on ideas and best practices in **both** ML & Stats

Builds a **platform** for further developments to assess and improve stability/robustness for the entire DSLC

# PCS

**Unifies, synthesizes,** and **expands** on ideas and best practices in **both** ML & Stats

Builds a **platform** for further developments to assess and improve stability/robustness for the entire DSLC

Its principles are common-sense principles:

"Pred-check" is about general reality check, model checking…

"S" covers new sources of uncertainty in a DSLC

"C" is indispensable and includes data-inspired simulations

# PCS in a nutshell

"P" for Pred-check: reality-check.

"S" **"shakes"** every step of DSLC via **reasonable** perturbation(s), assesses the effect via **stability metric(s),** and **aggregates results,** ideally after Pred-checks.

Perturbations and stability metrics are chosen by user ``**in-context**'' and **documented**.

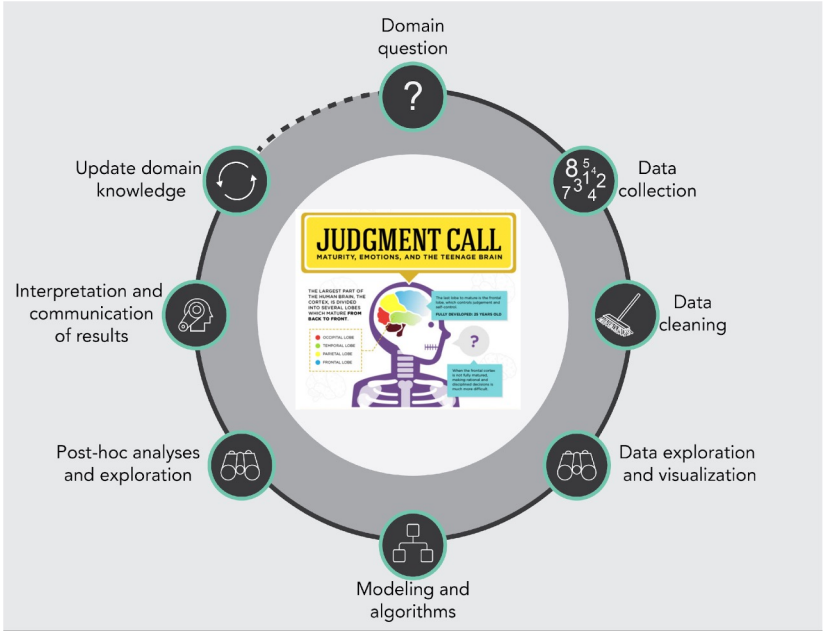



Image credits: R. Barter and toronto4kids.com

# Stability Principle

"At a minimum, reproducibility manifests itself in the stability of statistical results relative to "reasonable" perturbations to data and to the method or model used."

– Yu (2013) "Stability"

# Stability Principle

In PCS for VDS,  the stability principle is for the entire DSLC. It is a significant expansion on the sample-to-sample variability in statistics to consider **reasonable perturbations** such as above and data augmentation and include the numerical stability and stability from control theory.

# Stability Principle

In PCS for VDS,  the stability principle is for the entire DSLC. It is a significant expansion on the sample-to-sample variability in statistics to consider **reasonable perturbations** such as above and data augmentation and include the numerical stability and stability from control theory.

On the theory side, CLT and concentration results are stability results. There is uniform stability in ML theory and new works in ML are appearing related to stability.

# Stability Principle

In PCS for VDS, the stability principle is for the entire DSLC. It is a significant expansion on the sample-to-sample variability in statistics to consider **reasonable perturbations** such as above and data augmentation and include the numerical stability and stability from control theory.

On the theory side, CLT and concentration results are stability results. There is uniform stability in ML theory and new works in ML are appearing related to stability.

On the ground, after prediction check, stability has been crucial for finding genetic drivers for a heart disease, simplifying prostate cancer detection, understanding human brain, …

# PCS principles unify many aspects of Trustworthy AI

- Reliability
- Robustness
- Transfer learning
- Fairness
- Differential Privacy (DP)
- …

PCS is a practical prerequisite for interpretable ML/AI, trustworthy AI, and AI alignment, causal conclusions, …

# Stability Principle has two roles in PCS

- Assessing stability of results relative to reasonable perturbations and corresponding metrics in a DSCL (documented in context) (including PCS-UQ, which is quantitative)

- Improving stability of a DSLC, including the data cleaning step (standardize one data cleaning protocol) and modeling step (leading to new ML algorithms such as iterative random forests (iRF), staNMF, PCS ranking), which are **quantitative** PCS-guided methods.

# How to choose perturbations in PCS?

For **each step** of DSLC, there are **multiple reasonable choices**, possibly favored with different weights based on prior knowledge, and subject to resource constraints.

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

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

# Data and model perturbations worked in our projects

- Data cleaning schemes (e.g. 4) in prostate cancer prediction and classes
- Bootstrap samples
- Data split choices
- Randomized initializations (NESS) improving t-SNE and UMAP
- Linear and exponential prediction models for covid death count
- Supervised algorithms choicesfor tabular data (LS, Lasso, Ridge, ElasticNet, RF, RF+, XDBoost, MLPs)
- Decision tree choices in RFs and iRFs
- Perturbed DL models (drop-out, added noise, random embeddings)
- Importance measure choices

…

**DOCUMENT Judgment calls**

# Data and model perturbations: other forms

- Reality-checked simulation based data or synthetic data
- Augmented data
- Group-based split, time-based split, …
- Color, font, point size, and other graphic parameter choices in EDA in the data cleaning, model understanding and evaluation, and final report stages
- Unsupervised algorithm choices (e.g. cluster alg. choices)
- Architecture choices in training DL models
- Optimization algorithm choices in training and fine-tuning DL models
- Distillation choices for DL models

…

**DOCUMENT Judgment calls**

# Stability metrics worked in our projects

- Correlation
- L2
- 0-1 loss
- Sensitivity, specificity, AUC
- Interaction stability
- Importance ranking
- Absolute count error, relative count error, abs. error of sqr. count

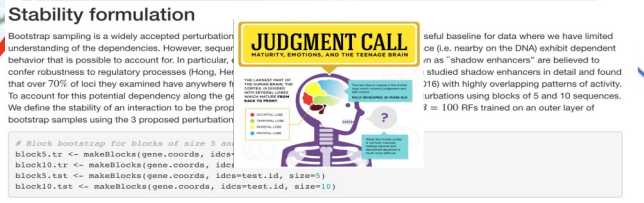
**DOCUMENT judgment calls**

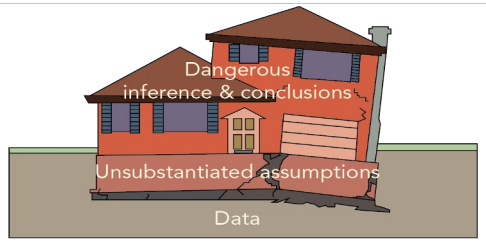
# How to aggregate pred-checked algorithms?

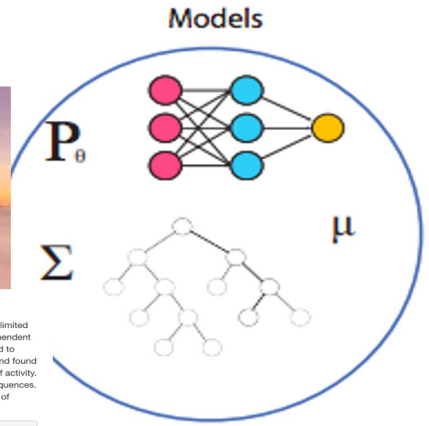
- In early stages of a project, taking union
- In later stages, taking intersection, weighted average, majority vote
- In PCS-UQ, keeping  pred-checked predictions

# PCS documentation [on GitHub ( JupyterNotebook / Quarto )]

Reality


Mental Construct

Image credits: Rebecca Barter

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

# PCS' multiple roles in past projects

- **Internal validity for rigorous DS/AI algorithm development**
- Recommendation for external causality validation
- Evaluating or stress-testing existing CDRs
- **New algorithm developments in context** to add (appropriate) stability (e.g. iterative random forests (iRF), lo-siRF, staNMF, staDISC, staDRIP, MDI+, PCS importance ranking, …)
- **Extensions** to veridical spatial data science, veridical network analysis, and reinforcement learning by others, and PCS-guided LLM development, …

# **PCS case studies:** modern experimental design

A different perturbation and related stability notion is used in context for each case study, after pred-check:

- StaDISC: finding stably interpretable and calibrated subgroups from RCT **(medicine)**
(Dwivedi and Tan, …, Madigan, Yu (2020) *International Statistics Review*).

- Lo-siRF: finding genetic drivers of HCM (experimentally validated) (**genomics**)
(Wang and Tang, …, Yu, Ashley (2024) *Nature Cardiovascular Research*)

- PCS ranking: cutting cost by ½ of a prostate cancer detection algorithm (**cancer res**)
(Tang and Zhang, …, Chinnaiyan, Yu (2025), *Cancer Biomarkers*)

- GCT: finding new meaningful subareas of the brain related to speech (**comp. neuro.**)
(Antonello and Singh, …, Yu, Huth (2025), submitted)

# PCS outperforms conventional methods in two case studies

- PCS ranking reduces cost by 55% for prostate cancer detection relative to conventional approach using one method (joint patent application filed) (Tang et al, 2024, *Cancer Biomarkers*).

- PCS-guided lo-siRF led to many more heart disease HCM related genes than conventional methods, based on annotated databases. Most importantly, PCS-guided lo-siRF recommendations result in 80% success rate (4 out of 5) in follow-up gene-silencing experiments. (Wang et al, 2025, *Nature Cardiovascular Research*).

Reason for success:

realistic accounting of uncertainty hence fewer false positives.

# Stability vs Uncertainty Quantification

A stability analysis is defined for a relevant perturbation (on problem formulation, data or algorithm, etc) to solve a domain problem and with a stability metric(s) that measures performance in context.

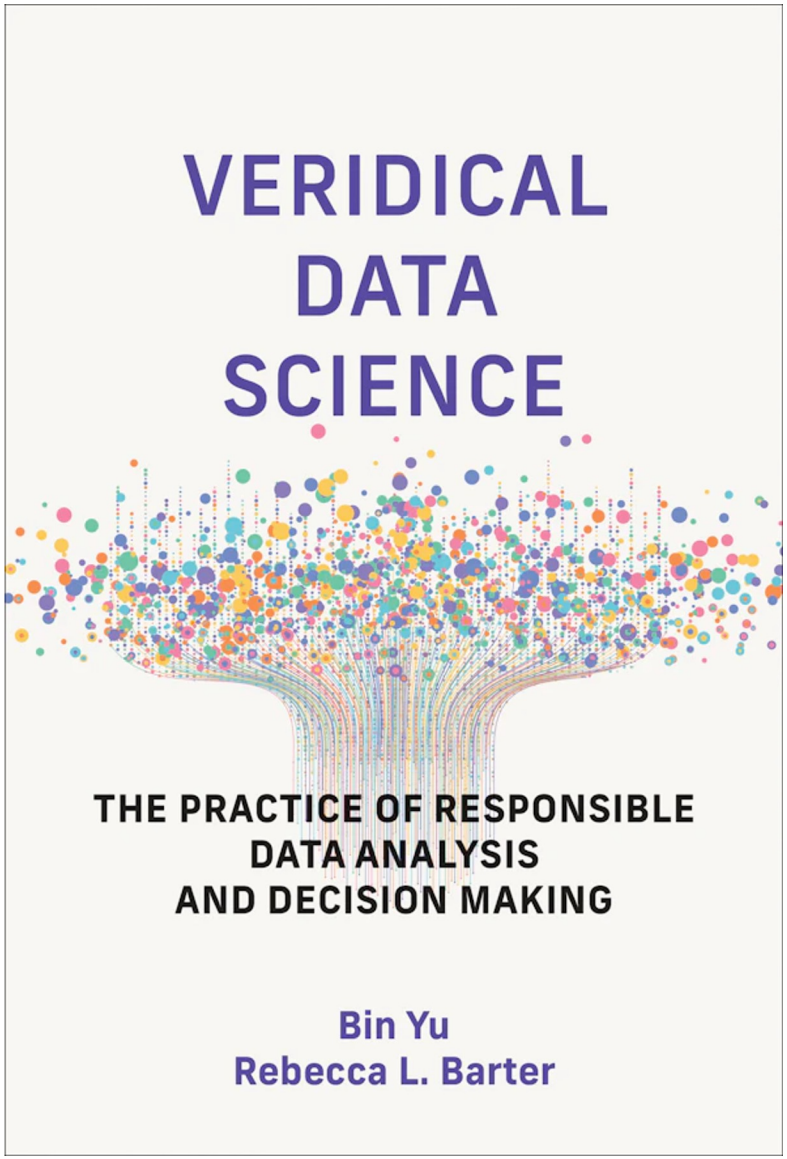
Uncertainty quantification is a special form of stability analysis when the perturbation is defined relative to a probability distribution (analytically or defined through a set of discrete datasets).

# Part 3: UQ Via PCS

@rlbarter

**Free online version now at**

**vdsbook.com**




**MIT Press (ML Series)**

**Oct. 15, 2024**

# 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***

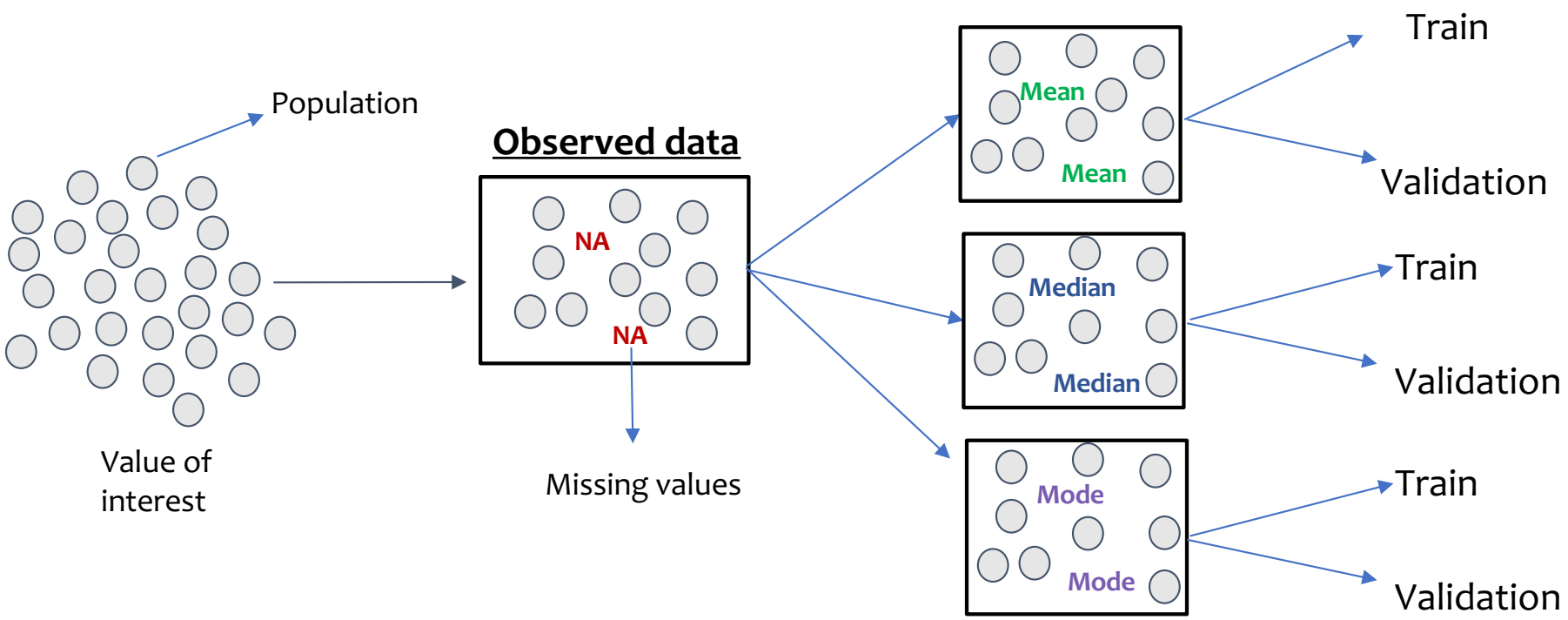
**Michael Xiao***

Boyu(Boris) Fan

Omer Ronen

* denotes equal contribution

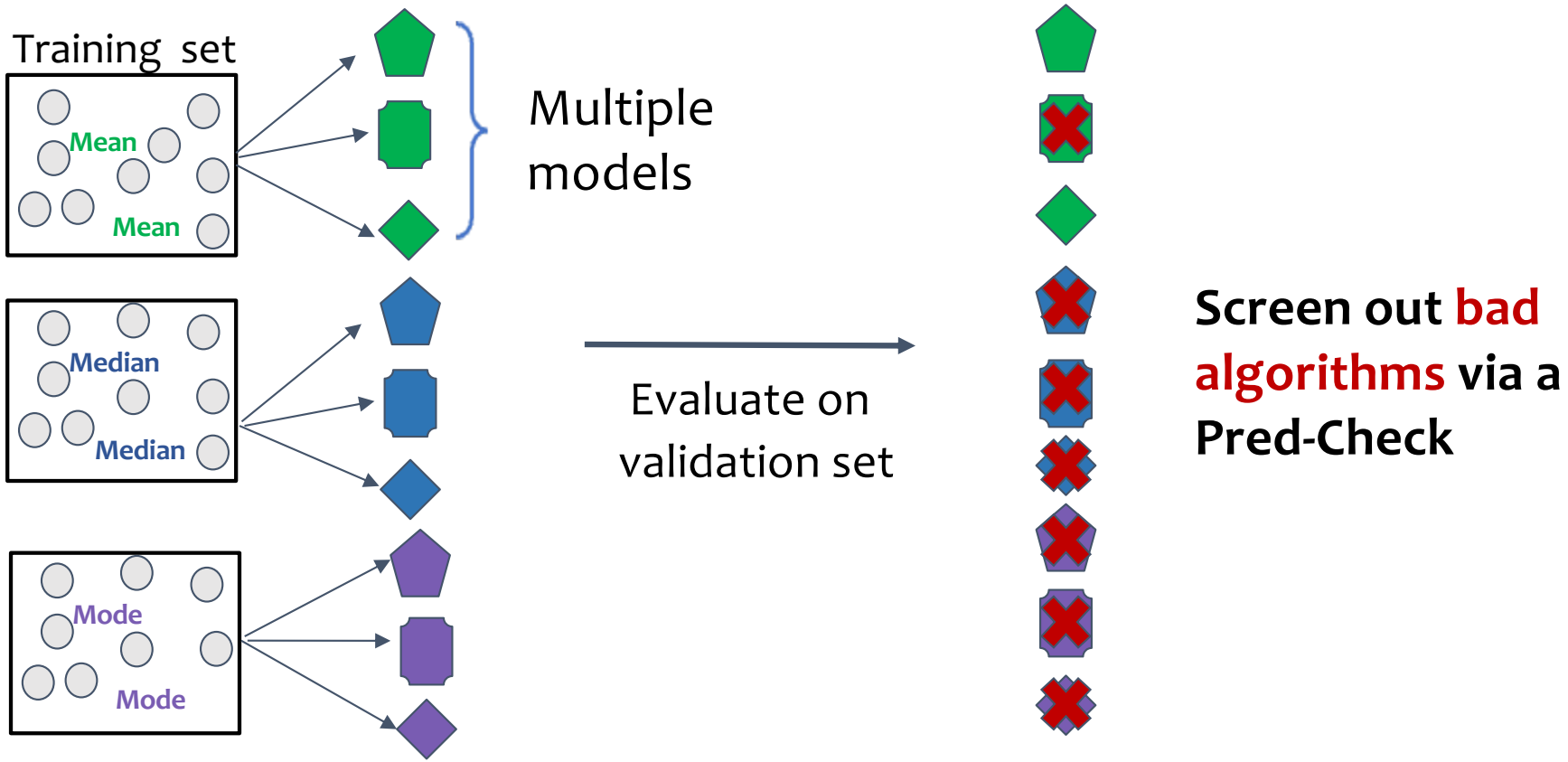
# PCS (Ch 13.) Perturbation Interval Step 1: Cleaning & Data Split




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

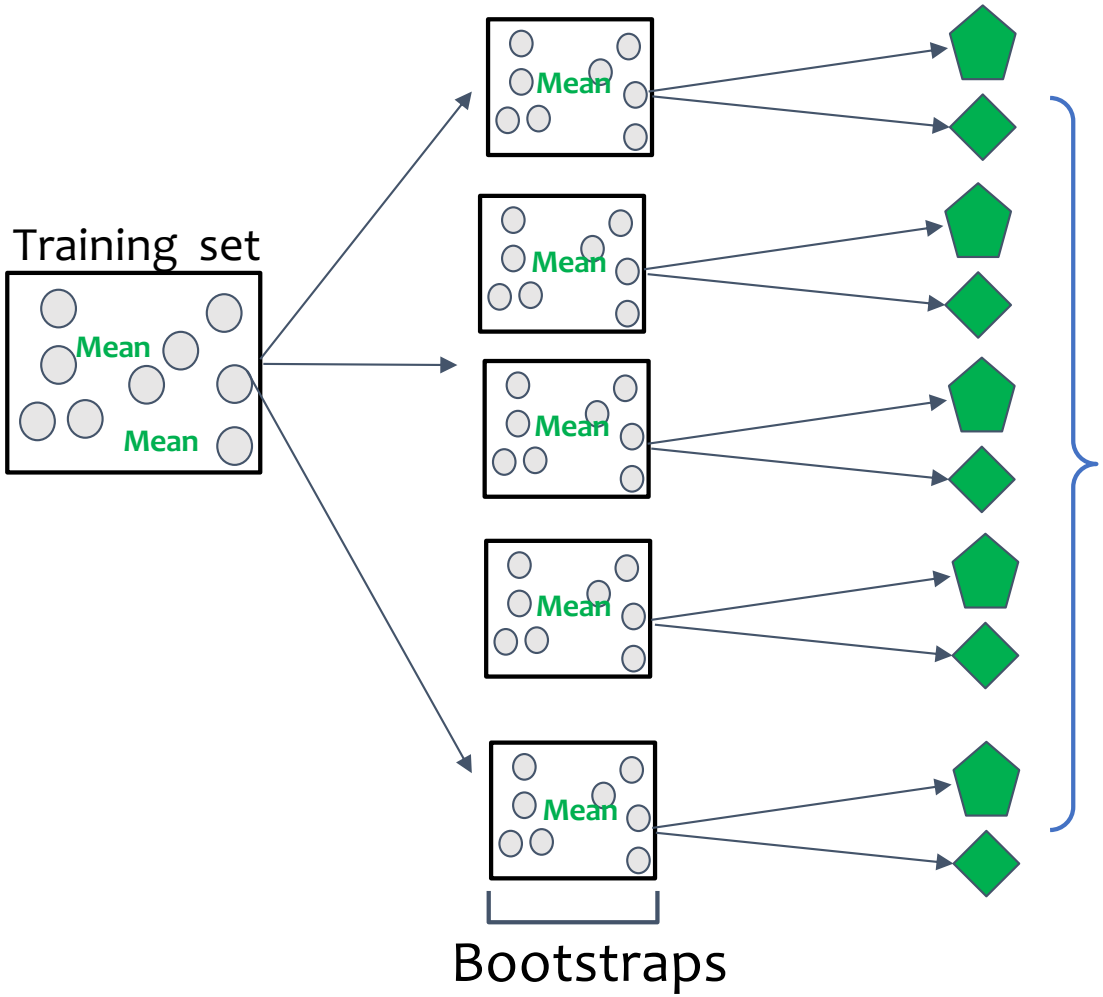
**Held-out Test Set.**

# Step 2: Pred-Check



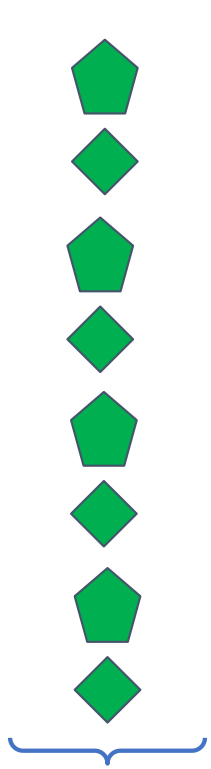

**Screen out bad algorithms via a Pred-Check**

# Step 3: Bootstrap

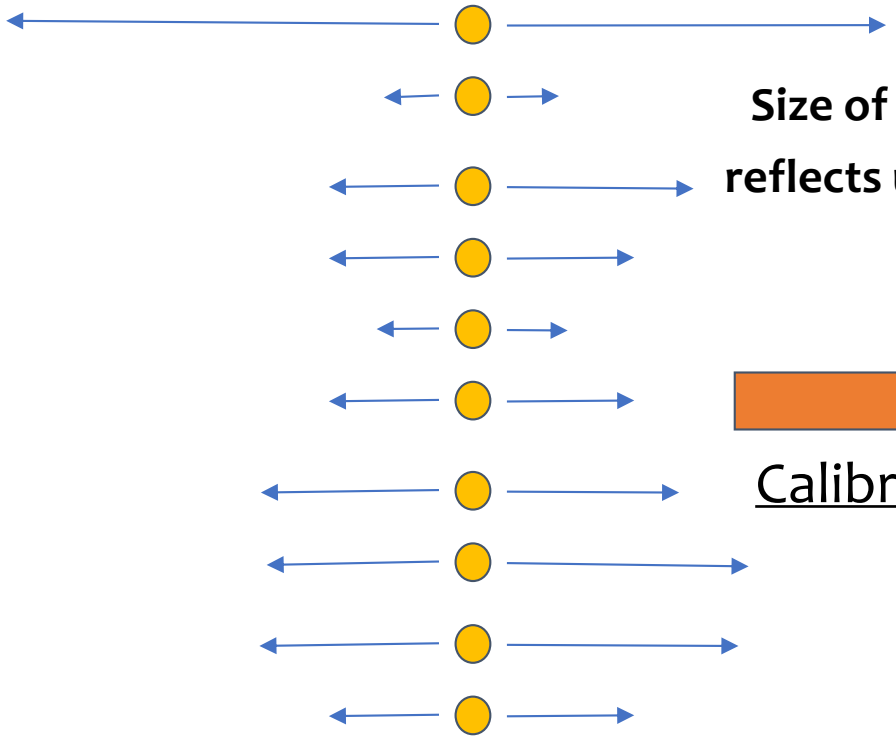



Bootstrapped **Pred-checked** models serve as uncertainty

# Step 4: Intervals & Calibration



Bootstrapped
**Pred-checked** models



**Size of interval**
**reflects uncertainty**

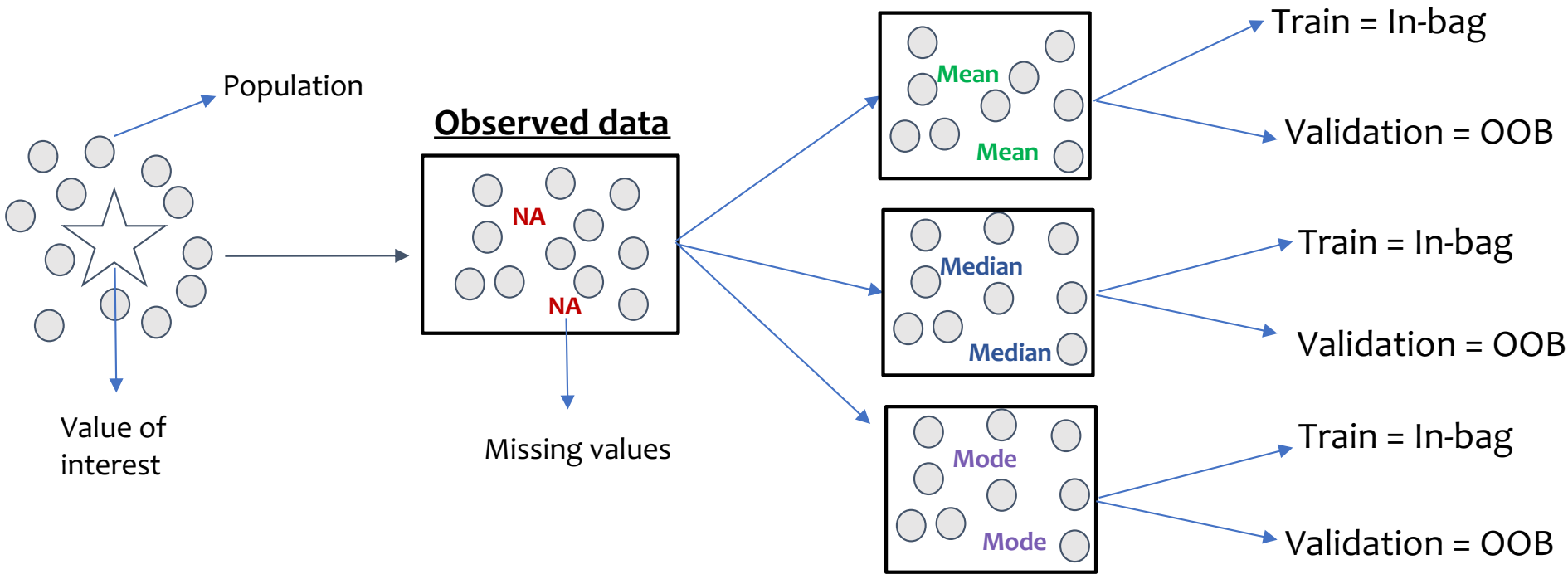
Construct "uncalibrated"
intervals on validation
via bootstrap predictions

Calibration

Find **multiplicative**
factor to achieve
desired coverage

# PCS (OOB)

For each bootstrap sample, use OOB (out of bootstrap) set as validation instead of a set-aside validation set as in Ch. 13 of VDS book.



# PCS Perturbation Interval

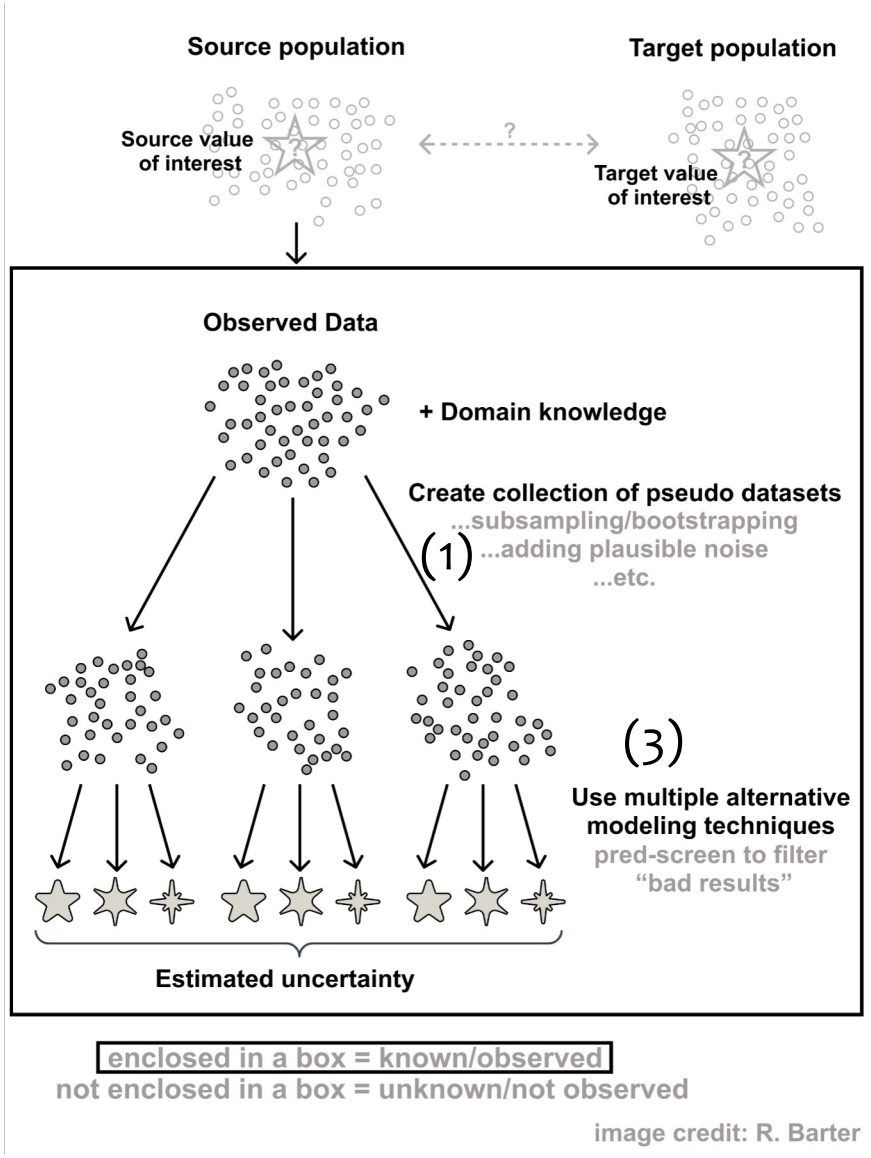
VDS book considers three sources of uncertainty in DSLC from

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 Conformal

## PCS

- Data cleaning uncertainty allowed
- Uses multiple ML algorithms
- "Pred-check" to screen bad algorithms
- "Local" calibration via stability (bootstrap, multiple p-checked algorithms, data cleaning)
- Scales length 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
- Constant length adjustment, to achieve coverage if exchangeability assumption holds between future data and current data

# Conformal Methods for Regression

- **(Split) 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)*

- **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)*

- **Majority Vote (Ensemble):** Merging Uncertainty Sets via Majority Vote
  - **Authors**: Gasparin M., & Ramdas A. .
  - ***Journal***: ArXiv (2024)

# Part 4: Experiments

# Overview of Experimental Set-Up

- Regression (17 Datasets)
  - Measure coverage & width per dataset across UQ methods
- Regression Subgroup Coverage & Width
  - Investigate coverage & width across natural subgroups per dataset
- Multi-Class Classification
- Deep-learning
  - Approximation schemes to reduce computational complexity

# 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-1 best** performing models across **1000** bootstraps

<u>How were hyper-parameters chosen?</u>

- 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**

# Real-World Regression Datasets

**(no data cleaning uncertainty)**

| Data | Num Observations | Num Features | Source |
|---|---|---|---|
| Airfoil | 1503 | 5 | UCI ML Repo |
| CA Housing | 5000 | 8 | Kaggle |
| Computer | 8192 | 21 | openML |
| Concrete | 1030 | 8 | UCI ML Repo |
| Debutanizer | 2394 | 7 | openML |
| Diamond | 7000 | 23 | UCI ML Repo |
| Energy Efficiency | 768 | 10 | UCI ML Repo |
| Elevator | 16599 | 81 | openML |
| Insurance | 1338 | 6 | openML |
| Kin8nm | 8192 | 8 | openML |
| Miami Housing | 5000 | 15 | openML |
| Naval Propulsion | 11934 | 24 | Kaggle |
| Parkinsons | 5875 | 18 | UCI ML Repo |
| Powerplant | 9568 | 4 | UCI ML Repo |
| Protein Structure | 45370 | 9 | Kaggle |
| QSAR | 1000 | 266 | openML |
| Superconductor | 21263 | 79 | openML |

# Part 4.1: Results

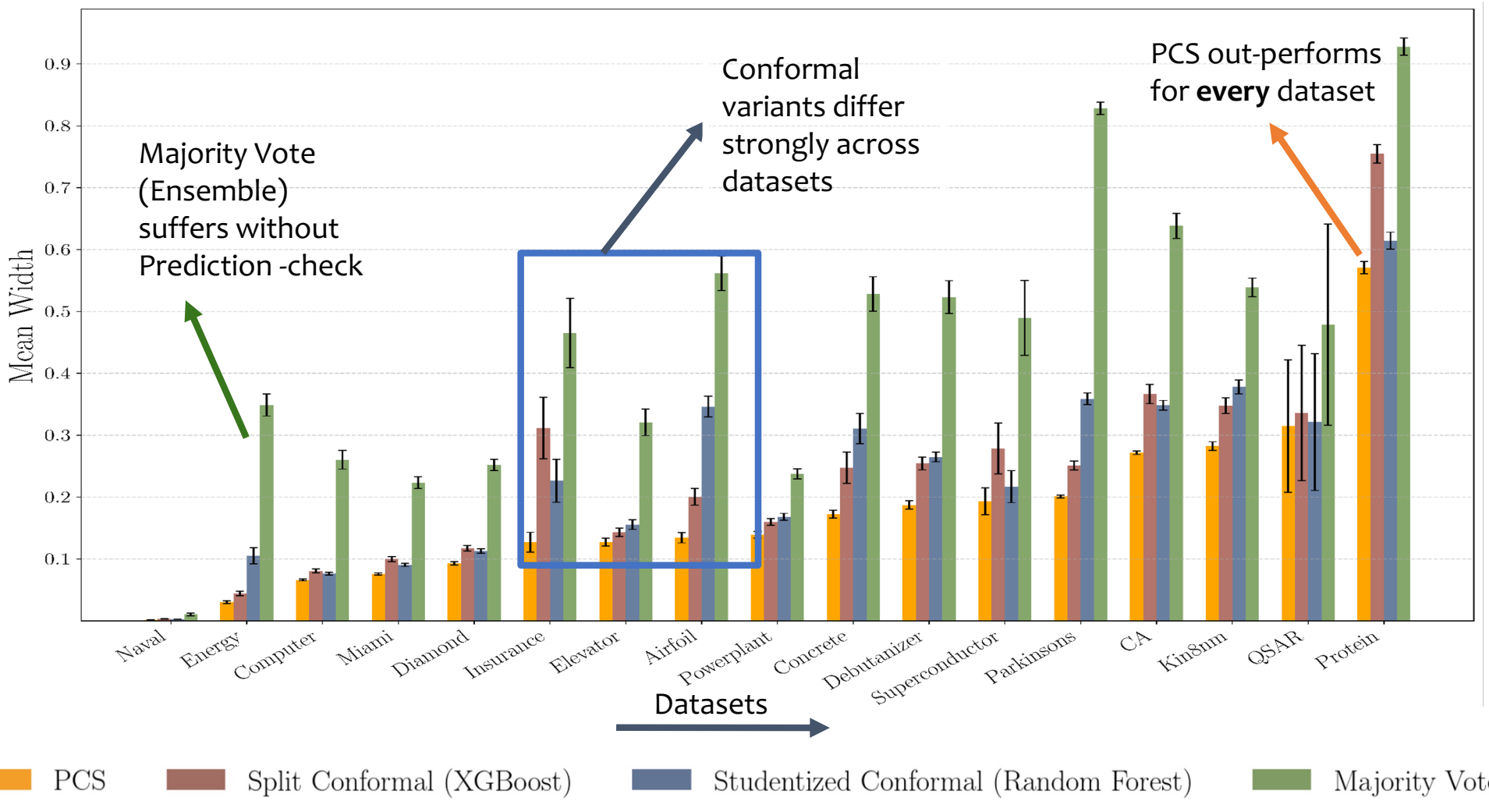
# Takeaways across 17 datasets

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

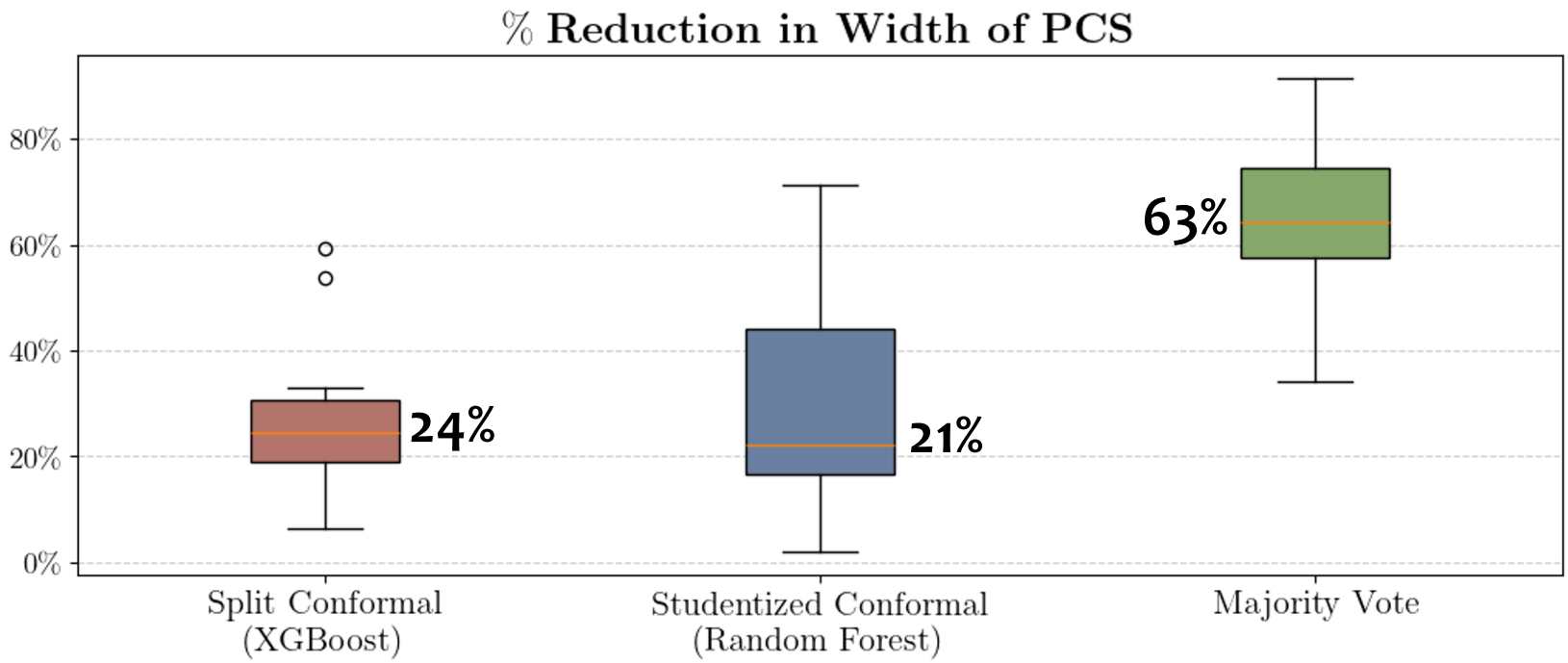
**Subgroups**

- **PCS** & **Studentized Conformal** adapt width to achieve subgroup-specific coverage; **Split conformal** does not
- **PCS** smaller width than studentized

# PCS Comparison to **Best** Conformal Methods



# Distribution of PCS Reduction in Width

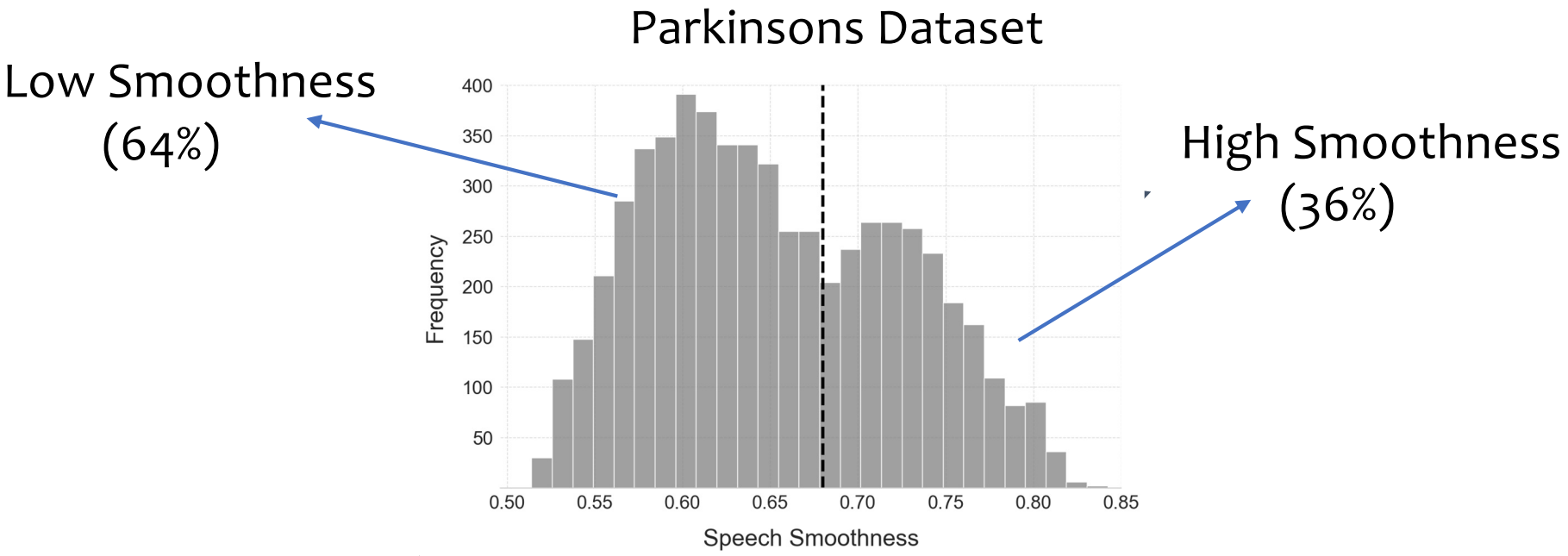



PCS UQ significantly reduces width

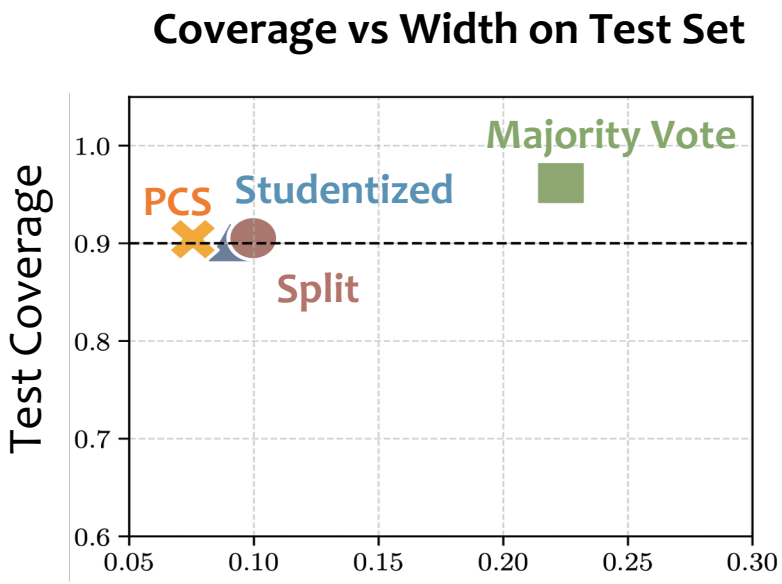
# Part 4.1: Subgroup comparisons

# Overview of Experimental Set-Up

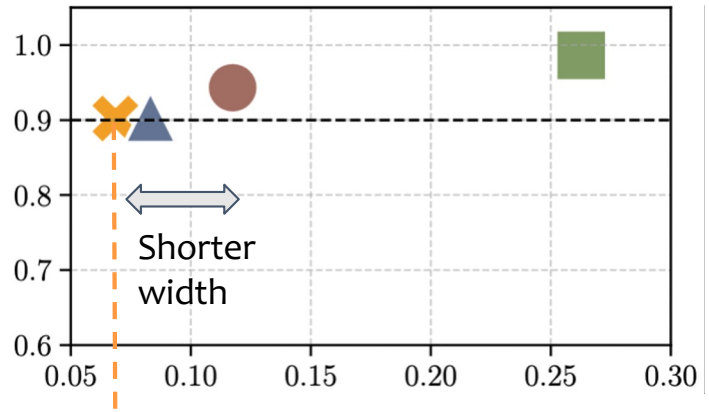
- Subgroup Coverage & Width
  - Investigate coverage & width across natural subgroups per dataset



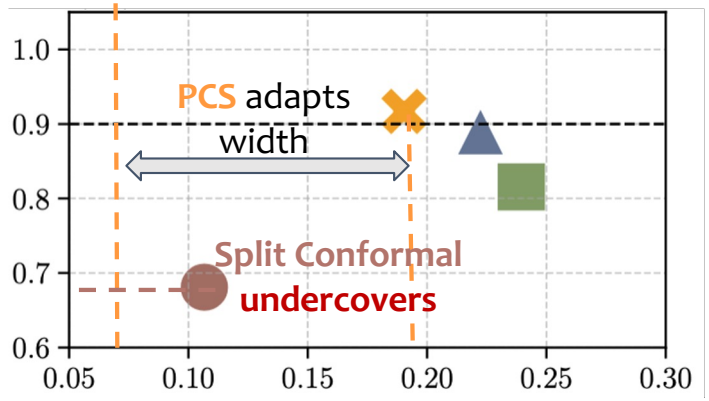
# Dataset: Miami Housing (n=13932, d=28)




Low income




High income




Takeaways:

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

# Part 4.3: Multi-Class & DL

# 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

# Conformal Methods for Classification

- **TopK:** Classification with valid and adaptive coverage
  - **Authors**: Romano Y., Sesia M., Candes E. .
  - ***Journal***: NIPS (2020)

- **Adaptive Prediction Sets:** Classification with valid and adaptive coverage
  - **Authors**: Romano Y., Sesia M., Candes E. .
  - ***Journal***: NIPS (2020)

- **RAPS Conformal:** Uncertainty Sets for Image Classifiers using Conformal Prediction
  - **Authors**: Angelopoulos A., Bates S., Malik J, Jordan M.
  - ***Journal***: ICLR (2021)

- **Majority Vote (Ensemble):** Merging Uncertainty Sets via Majority Vote
  - **Authors**: Gasparin M., & Ramdas A. .
  - ***Journal***: ArXiv (2024)

# Real-World Classification Datasets

**(no data cleaning uncertainty)**

| Data | Num Observations | Num Features | Num Classes | Source |
|---|---|---|---|---|
| Language | 1000 | 19 | 30 | OpenML |
| Yeast | 1484 | 8 | 10 | OpenML |
| Isolet | 7797 | 613 | 26 | UCI ML Repo |
| Cover Type | 10000 | 13 | 100 | UCI ML Repo |
| Chess | 28056 | 34 | 18 | OpenML |
| Dionis | 30000 | 60 | 355 | OpenML |
| CIFAR-100 **(DL)** | 60,000 | | 100 | HuggingFace |
| Tiny Image-Net **(DL)** | 100,000 | | 200 | HuggingFace |
| Places 365 Small **(DL)** | 100,000 | | 365 | HuggingFace |

# PCS Candidate Models for Classification

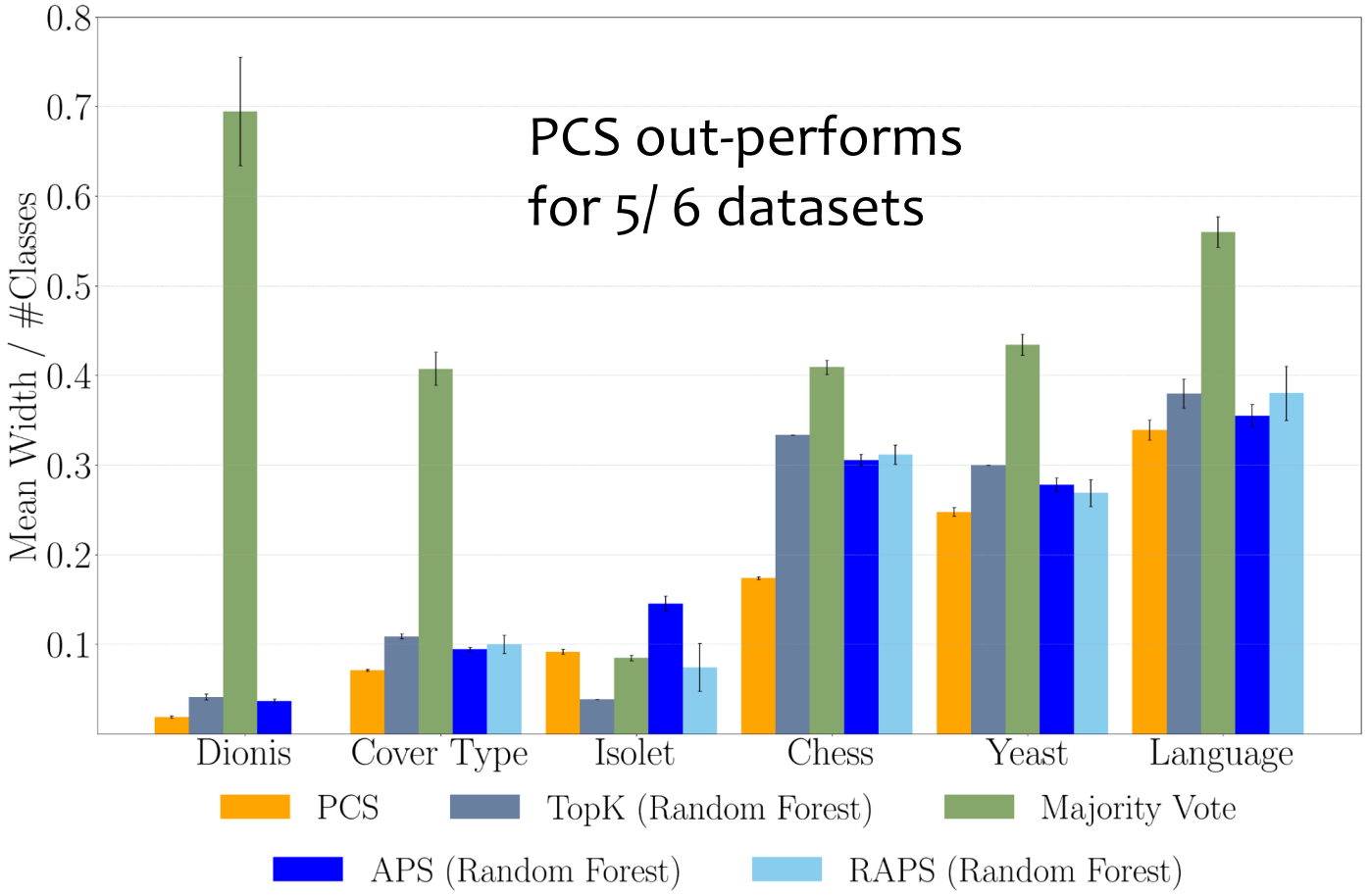
- For tabular datasets:
  - Linear: Logistic Regression with L2-regularization
  - Bagging: Random Forests (RFs), ExtraTrees
  - Boosting: XGBoost, AdaBoost,
  - Multi-layer Perceptrons (1 hidden layer)

- For Deep-learning:
  - Res-net 18

# PCS Candidate Models for Classification

- For tabular datasets:
  - Linear: Logistic Regression with L2-regularization
  - Bagging: Random Forests (RFs), ExtraTrees
  - Boosting: XGBoost, AdaBoost,
  - Multi-layer Perceptrons (1 hidden layer)

# Tabular Classification: PCS out-performs Conformal by ~20%



## PCS UQ Deep Learning: bootstrap too expensive

Perturbations to improve computational efficiency (compared to bootstrap):

- *Weighted dropout*: randomly remove nodes based on magnitude of activation
- *Additive noise*: add i.i.d. Gaussian noise to weights
- *Randomized embedding*: obtain embeddings from randomly sampled layers & train linear classifier on embeddings

## 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**

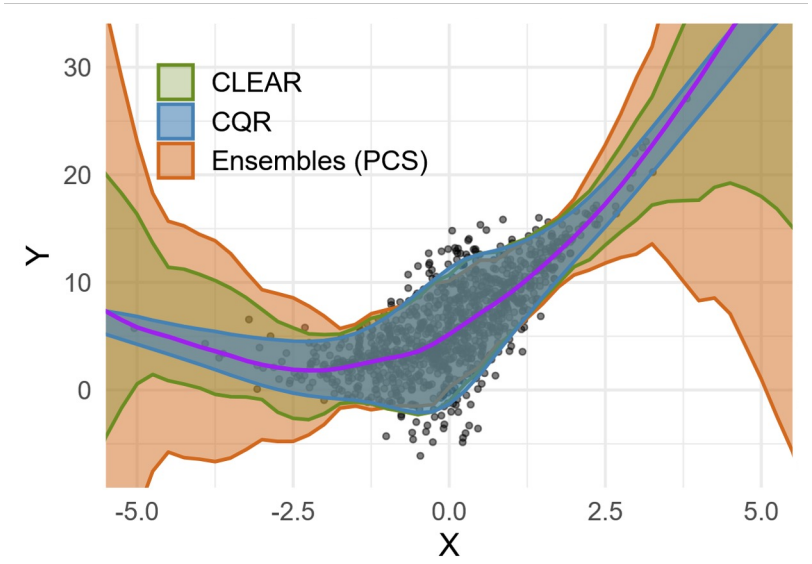
# Part 5: Theory

# 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.

# Part 6: PCS  Current Directions

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




Combine PCS-UQ and Quantile Regression:

multiplicative scaling with 2 calibration parameters

➔ width reduction of (averaged over 17 datasets)

- 15% compared to PCS-UQ
- 28% compared to CQR


I. Azizi*


J. Bodik*


J. Heiss*


B. Yu

* denotes equal contribution

# PCS Research directions

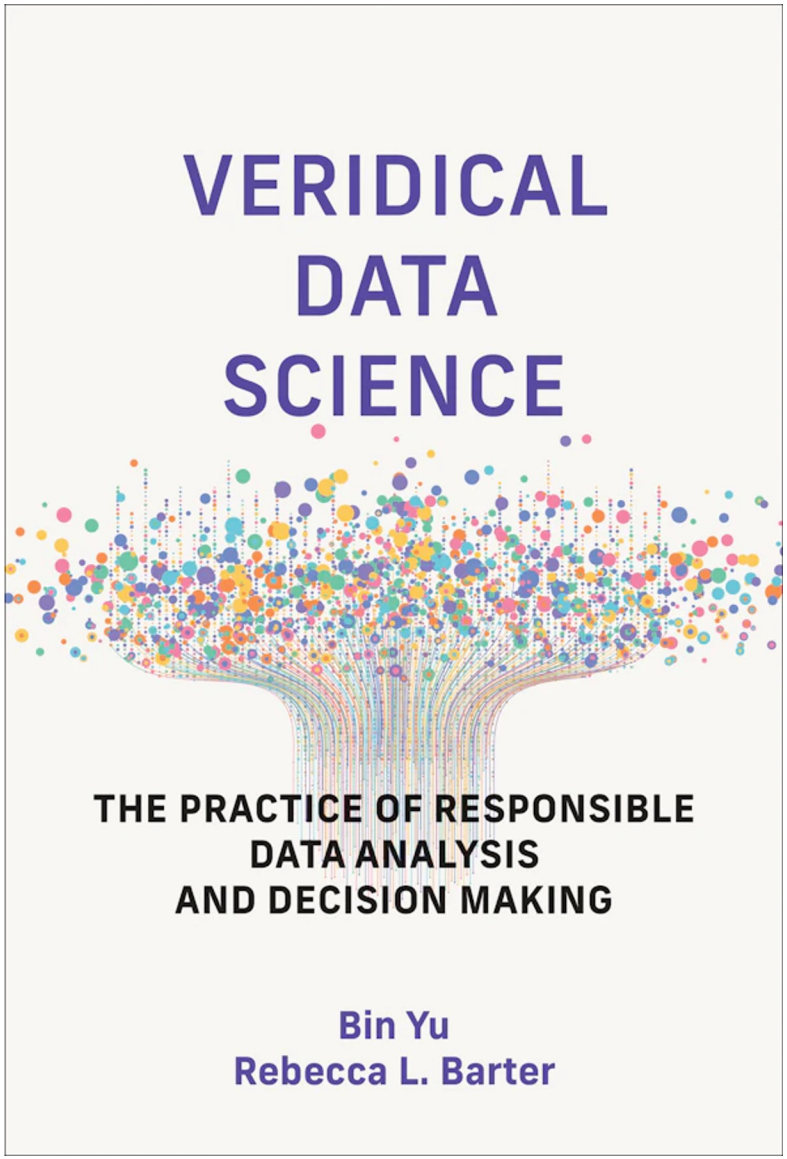
- Clear-UQ combines PCS-UQ with CQR to achieve improvements of 17% and 28% over PCS-UQ and CQR respectively (Azizi, Bodik, Heiss and Y., 2025)
- Protein fitness prediction (Ronen, Zhao, Ye, and Y., 2025)
- Enhancement of t-SNE and UMAP (Ma, Li, Hu, and Y., 2025)
- Experimental design of promoters in genomics
- …

@rlbarter

**Free online at**

**vdsbook.com**

Intended audience:
upper div and grad.
domain experts




**MIT Press**
**(ML Series)**

**Oct. 15, 2024**

Book review by
Benjaminis in HDSR

# Distinctive features of the book

It **mirrors practice** or follows the **data science life cycle** with chapters on **problem formulation** and **data preparation,** on stats/ML methods, and on communication

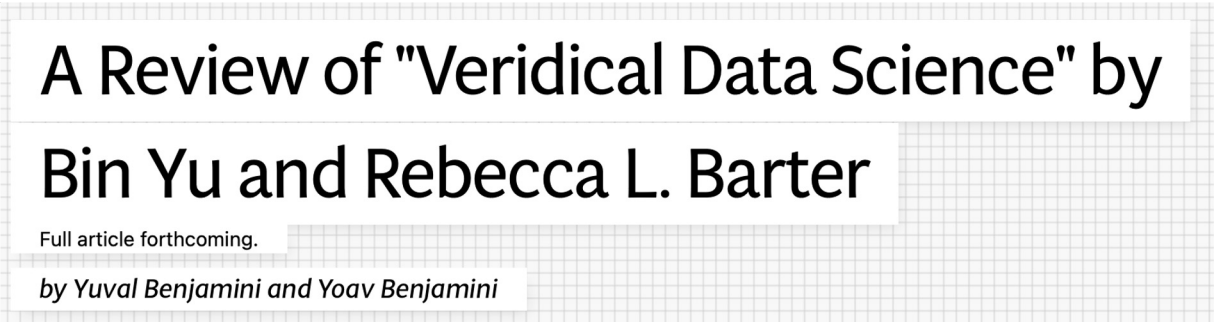
**PCS** is in every chapter, and so is documentation

It is **comprehensive** and coaches **critical thinking**

# Detailed differences from traditional books

- Moves away from “true-model” framing
- Fills gaps between domain problem and X, Y, …
- Teaches Stats/DS/ML methods through case studies with a PCS overlay from the user point of view
- Addresses two new sources of uncertainty arising from choices of data cleaning schemes and models
- Five kinds of exercises
  (T/F, conceptual, math, coding, project)
- Codes on github

# Book review by Yuval 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."*

# PCS for VDS is a research program for DS and AI

- Philosophical, conceptual, practical, and systems approach, standing on basic principles PCS. It embraces pluralism.

- Indispensable PCS documentation to build trust and to encourage qualitative and quantitative critical thinking in context.

Necessarily vague to allow domain knowledge and critical thinking to devise Pred-checks (i.e. reality checks) and **reasonable perturbations** for S-checks "in-context".

# “Veridical data science for medical foundation models”

( Alaa and Yu, 2024)

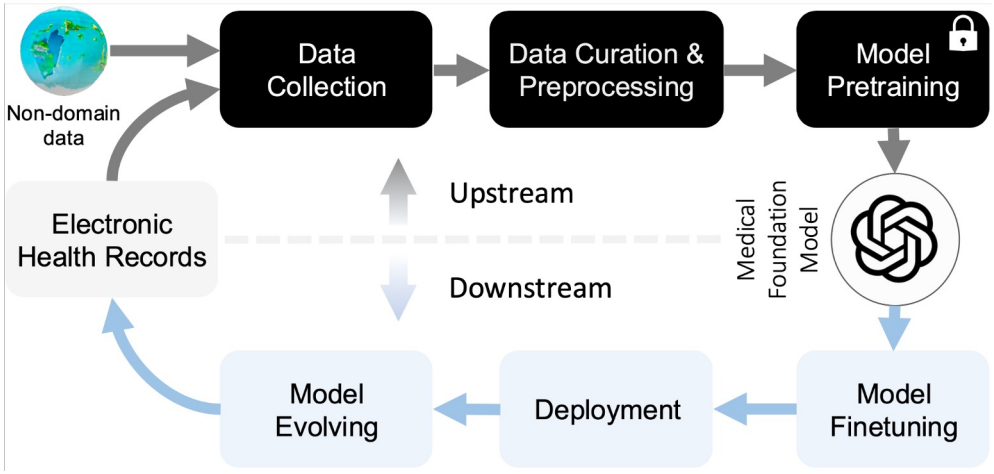
How is the foundation model life cycle (FMLC) different?

Ahmed Alaa

“Black-box” upstream process

Upstream:

Downstream:




Downstream process constrained by the black-box upstream process

# Software to address “C” in PCS

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


A. Agarwal J. Duncan R. Kapoor C. Singh

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

**Merits:** simulation guidelines (Elliott et al, 2024)


J. Duncan C. F. Elliott T. Tang M. Behr K. Kumbier

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

# MERITS of a high-quality simulation study

(Elliott et al, 2024): PCS-inspired simulation guidelines to address "C"

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

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

(Predictability) **Realistic:** Faithful to the physical world.

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

**Transparent:** Documented thoroughly and candidly.

(Stability) **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



# 1 Domain problem formulation

**What is the real-world question? This could be hypothesis-driven or discovery-based.**

> This should be very high level, providing the big picture behind the study. Often this takes the form of a pre-existi
> hypothesis (e.g., individuals with a specific genetic mutation are more likely to have a given characteristic) or mo
> open-ended discovery (e.g., identify mutations that are related to a given characteristic).

Insert narrative here.

**Why is this question interesting and important? What are the implications of better understanding this data?**

# VDS workshops: check Bin's website

Upcoming:

Jan., 2026, On-line biweekly seminar on VDS in Biology (details to come)
Early May, 2026, Paris, VDS workshop (details to come)

Past:

**July 11 2025,** at UC Berkeley
VDS in Biology
https://www.eventbrite.com/e/veridical-data-science-for-biology-2025-tickets-1384456339179l

**June 20, 2025**, at University of Sapienza, Rome
https://www.integreat.no/events/public-events/workshops/veridical-data-science.html

**May 31, 2024,** at UC Berkeley

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

# The in-context approach for stats/DS/ML research

*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"****.*

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

# Parting thoughts

- Solve problems of today
- Engage in ML, Deep Learning, and AI (esp. AI safety): one culture through VDS
- Do relevant theory
- Make impact in society

# Thank you!


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

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

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

T. Tang, Y. Zhang, A. Kenney, …., B. Yu, Arul Chinnaiyan (2024). A simplified MyProstateScore2 for high-grade prostate cancer. *Cancer Biomarkers*.

Q. Wang*, T. M. Tang*, N. Youlton, C. S. Weldy, A. M. Kenney, O. Ronen, J. W. Hughes, E. T. Chin, S. C. Sutton. A. Agarwal, X. Li, M. Behr, K. Kumbier, C. S. Moravec, W. H. W. Tang, K. B. Margulies, T. P. Cappola, A. J. Buitte, R. Arnaout, J. B. Brown, J. R. Priest, V. N. Parikh, B. Yu*, E. Ashley* (2025). Epistasis regulates genetic control of cardiac hypertrophy. *Nature Cardiovascular Research* (Code) (PCS documentation)

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).

# Recent papers

A. Alaa and B.Yu (2014) 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 (submitted, https://arxiv.org/abs/2505.08784).

Z. Rewolinski and B. Yu (2025) PCS workflow for veridical data science in the age of AI (submitted; https://arxiv.org/abs/2508.00835).