Bridging the Gap: A Comprehensive Approach to Responsible Data Science Education

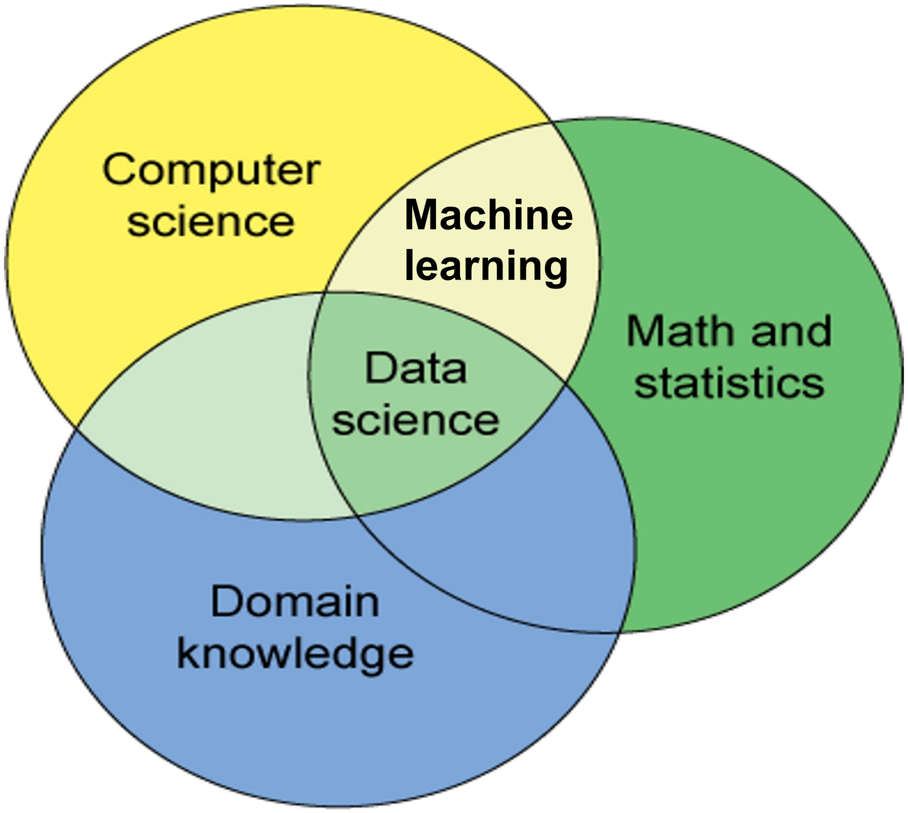
Bin Yu
Statistics, EECS, Comp. Bio.

binyu@berkeley.edu

Berkeley
UNIVERSITY OF CALIFORNIA

AI and Data Science: Integrating Artificial and Human Ecosystems
Harvard Data Science Initiative (HDSI)
June 13, 2024

Data science (DS)



Conway's Venn Diagram

Goal:

Leverage **algorithms**
to combine **data**
with **domain knowledge**
to make decisions
and generate
new knowledge

What is the goal of DS education?

(A)

To prove mathematical results about *data*

or

(B)

To solve real-world domain (e.g., scientific) problems

What is the goal of DS education?

(A)

To prove mathematical results about *data*

or

(B)

To solve real-world domain (e.g., scientific) problems

The overall goal of data science/statistics/ML education is to train students to data (and domain knowledge) to answer questions about the real world.

While ***relevant* theory** underlies practice when assumptions can be checked (e..g using data or domain knowledge), this is not the case with *all* theory.

Marvin Zelen in 1965 on statistics education

“...scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one’s lifetime in drawing inferences from data.”

– Quoted by S. Goodman (2019)

“Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics” in *American Statistician*, which cited Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman, M. A. (1966), “The Role of Hypothesis Testing in Clinical Trials,” *J Chron Disease*.

Most textbooks

1. Start with a data set

2. Exploratory data analysis

3. Devise optimal mathematical solutions under a **“true”** probabilistic model about data

4. Model checking deals with uncertainty under a probabilistic model

Most textbooks don’t have sections on running simulations or coding exercises

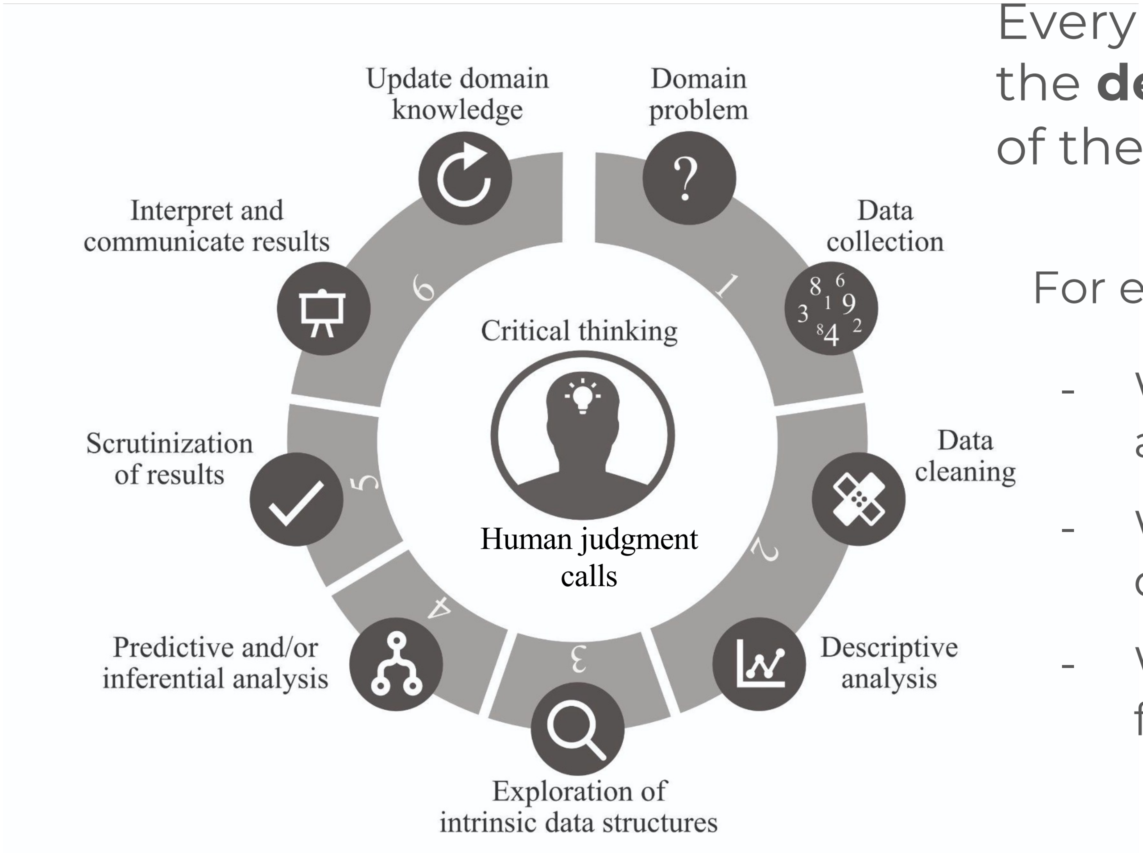
Gaps

1. In practice, data scientists try **many models** and often cherry-pick the best one using data possibly to align with the “true” model framing.

2. Data cleaning takes up a substantial amount of human time and **many choices** – missing in traditional textbooks and hiding under the rug in practice.

3. Coding is a must in practice, but is not an integral part of textbooks.

The Data Science Life Cycle (DSLC)



Every data-driven result is a result of the **decisions made at *every* stage** of the Data Science Life Cycle (DSLC)

For example:

- What if a researcher uses data from a different database?
- What if the data had been cleaned differently, or by a different person?
- What if a different model had been fit to generate a predictive algorithm

Uncertainty and judgment calls

Even if you can assume a probabilistic model... **uncertainties** arise from other sources:

Every data-driven result is a function of

- The human analyst
- The data cleaning protocol
- The assumptions made
- The modeling choices
- ...

Every data-driven result is based on a series of human **judgment calls**

Uncertainty quantification (UQ) is central for building trust in DS and AI

Current stats/DS practice considers only

uncertainty from a generative stochastic model,

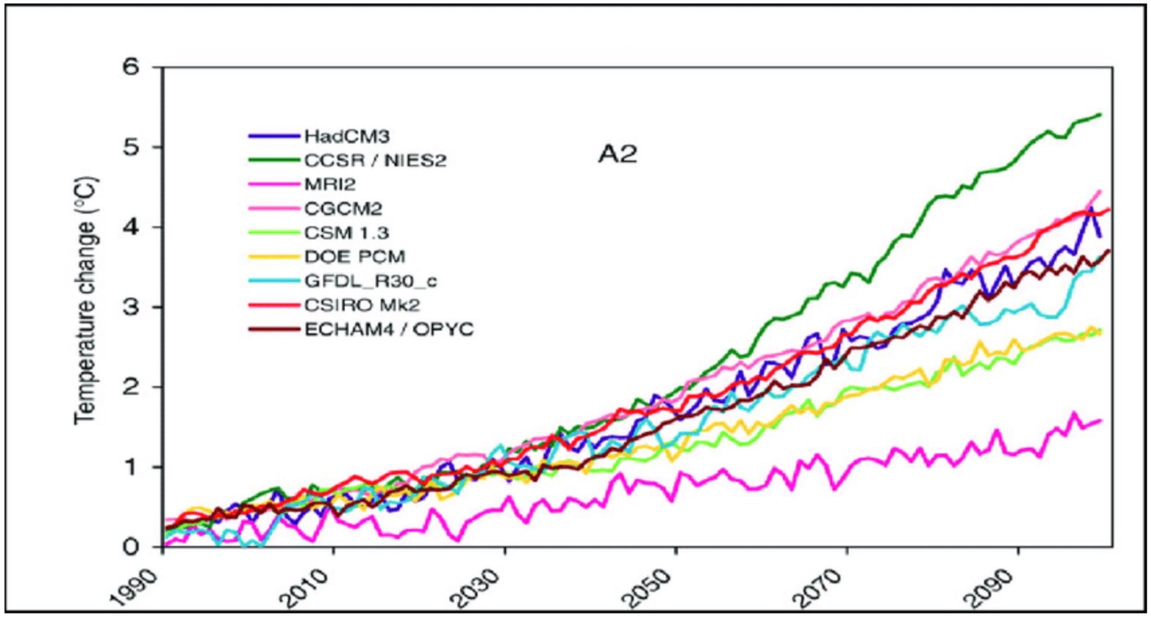
which is often assumed, with limited empirical checking.

In a data science life cycle (DSLC), there are many other important sources of uncertainty, due to human judgment calls.

Trustworthy uncertainty quantification is a must.

There are multiple “plausible” versions of almost every data-driven result

Climate scientists have generated a range of different projections of mean global temperature change:



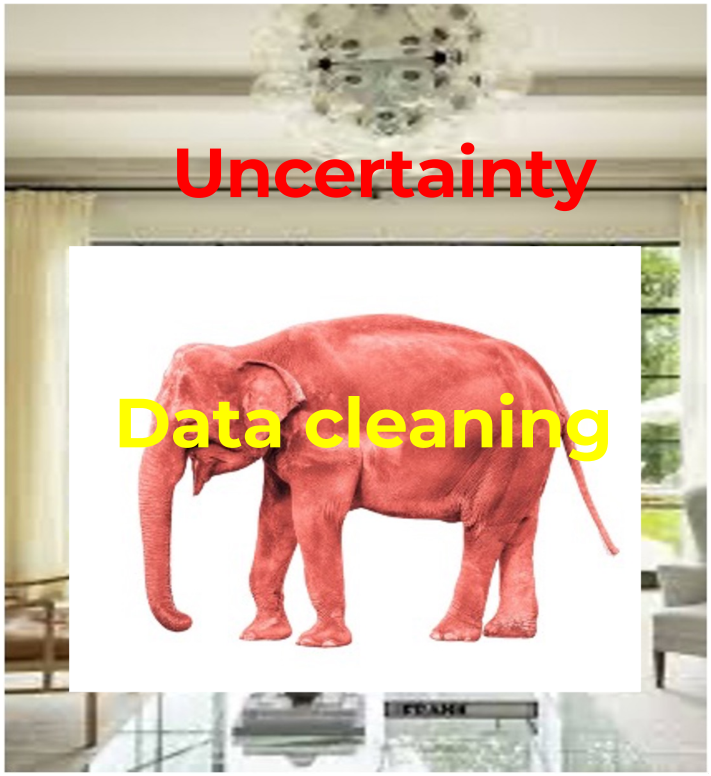
The change in global-mean temperature estimated by nine climate models forced by the SRES A2 emission scenario. (Source: IPCC TAR, Chapter 9)

Questions:

What are some reasons for why these models’ projections differ?

How should we account for different sources of uncertainty?

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/

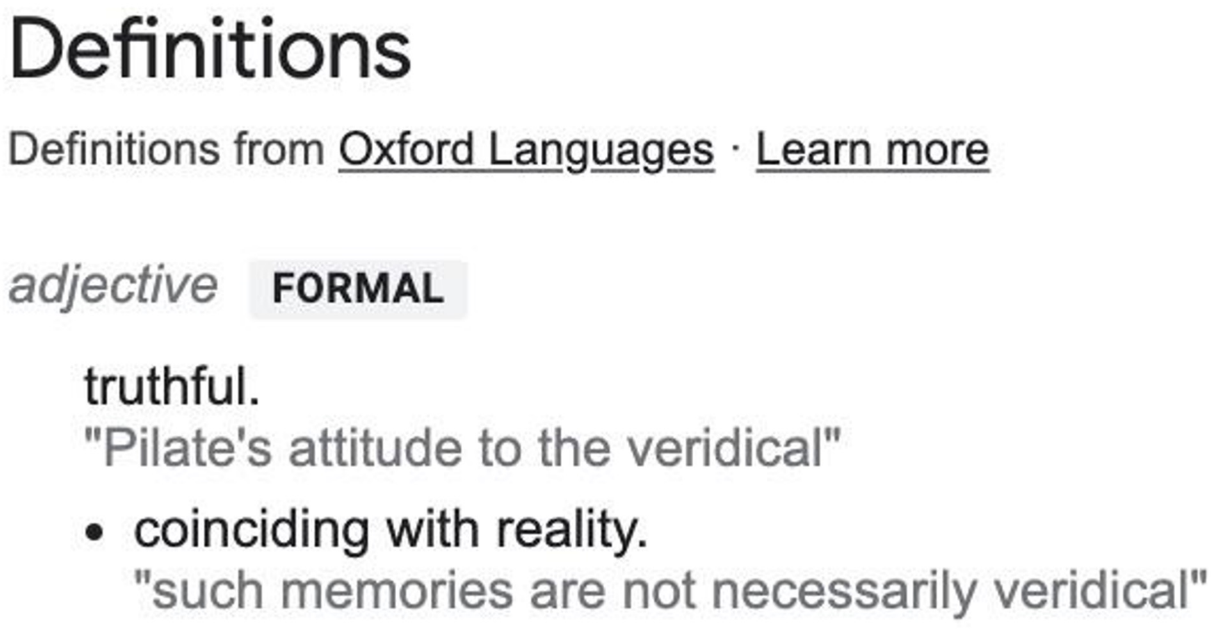
Veridical data science

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

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

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

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"

Motivating Veridical Data Science

Case study: prostate cancer detection

- A **leading cause of cancer death** in the developed world

- **Unclear benefits of current screening procedures** via prostate-specific antigen (PSA, protein from KLK3 gene) (AUC 59%: high-rate of false positives and invasive biopsies)

- Can one do better than PSA?

American Cancer Society (2023): https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html

Reproducibility crisis (early 2010’s)

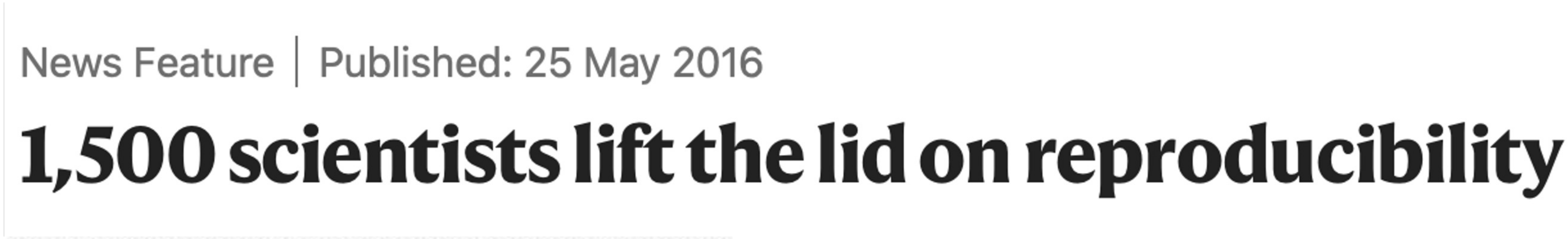
“Scientists from biotech companies Amgen and Bayer Healthcare reported alarmingly **low replication rates (11–20%)** of landmark findings in preclinical oncological research.”

-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

Nature article in 2016



News Feature | Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

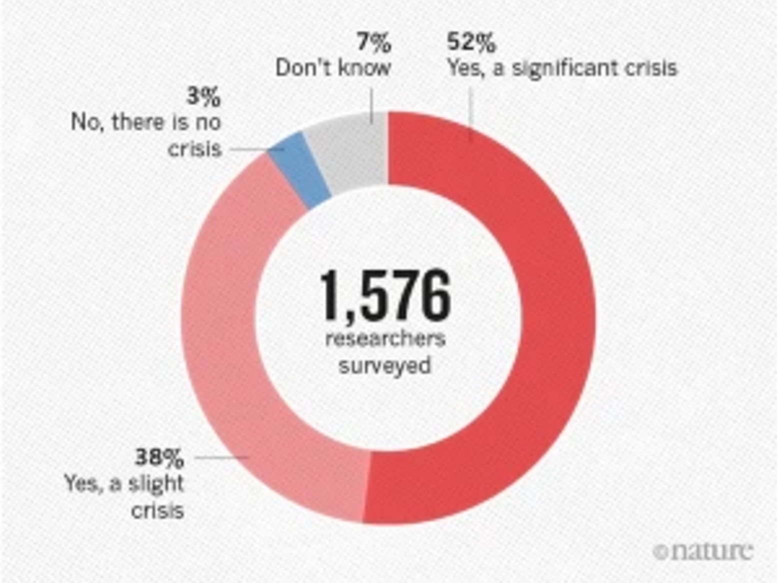
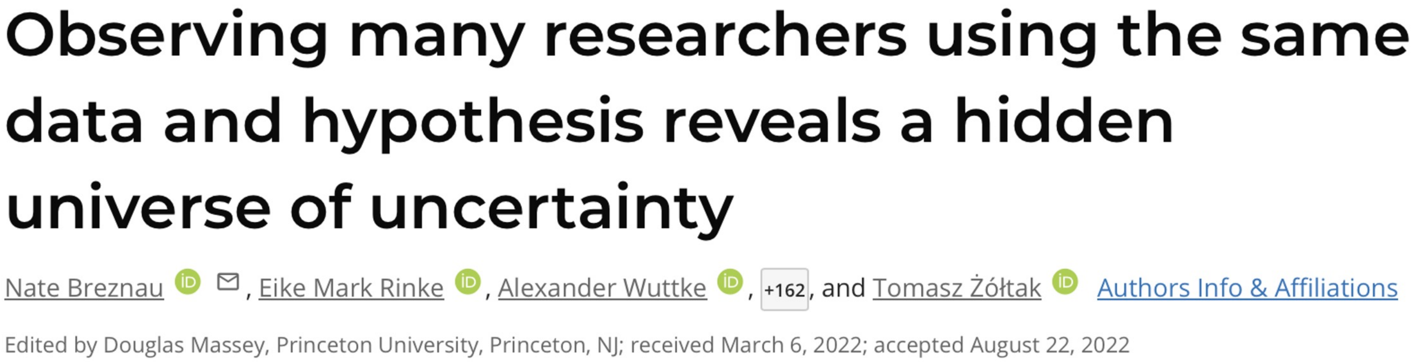


Image from https://www.nature.com/articles/533452a

“More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.”

PNAS article in 2022



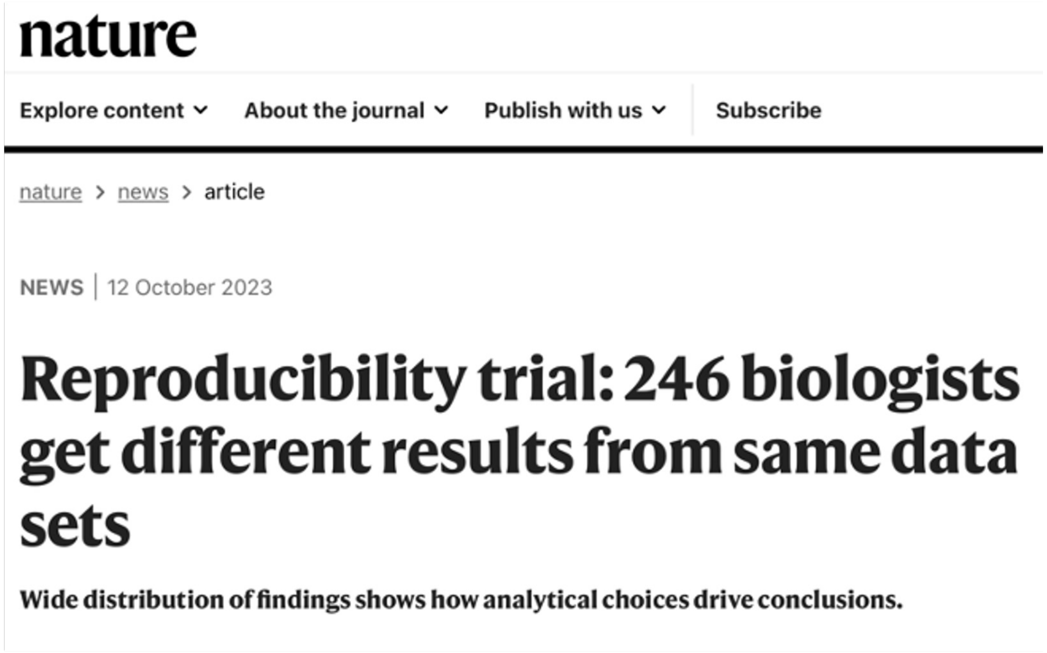
Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

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



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.”

VDS helps improve data science

As statisticians/data scientists/machine learners, our job is to (help) **formulate domain problems** and use **data** together with **domain knowledge and relevant theory** to collect **evidence** to postulate answers to **domain questions**

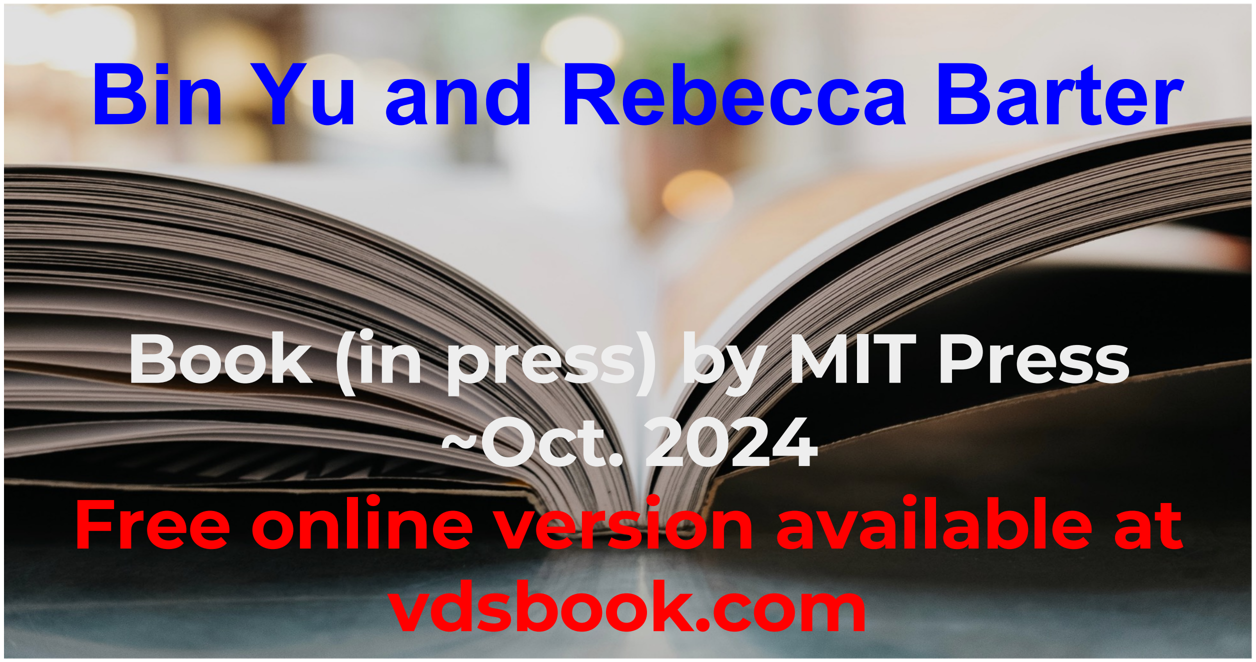
Evidence needs to be documented!

Book by Yu and Barter:

Veridical data science

Veridical Data Science:

The Practice of Responsible Data Analysis and Decision Making



Bin Yu and Rebecca Barter

Book (in press) by MIT Press ~Oct. 2024

Free online version available at vdsbook.com

Intended audience

Through first principles embedded in narratives and graphs, we aim at

- Upper div and beginning grad Stats/ML/DS courses, as primary or secondary textbook

- Domain experts including government officials

Image credits:
https://www.campussafetymagazine.com/university/help-new-college-students-feel-safe/
https://prettybrainy.com/7-women-scientists-who-broke-barriers/

Co-author: Rebecca Barter

@rlbarter

- Former statistics PhD student and postdoc (now at Univ. of Utah) of me at UC Berkeley
- Data Science Educator and Communicator & Data Science applications in healthcare
- Website: www.rebeccabarter.com (Blogs about statistics and R)
- Co-author of Superheat R package
- Co-author of upcoming *Veridical Data Science* book

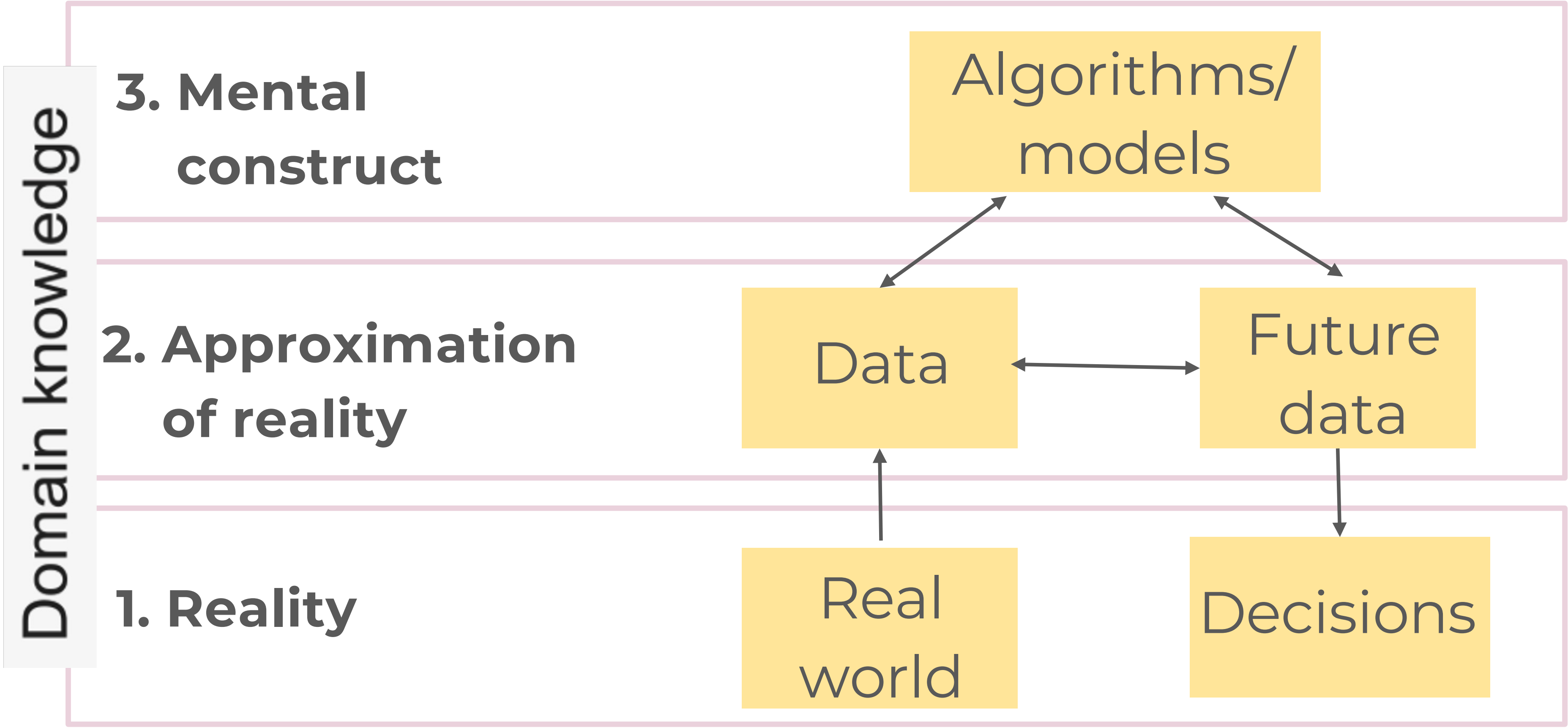
Real data imperfectly reflects reality

The real world is **too complex** to be entirely captured within a dataset

The extent that **data-driven conclusions reflect the real world** is upper-bounded by how well the underlying **data** reflects the real world.

Photo by Faye Cornish on Unsplash

Algorithms are a mental construct built upon imperfect data



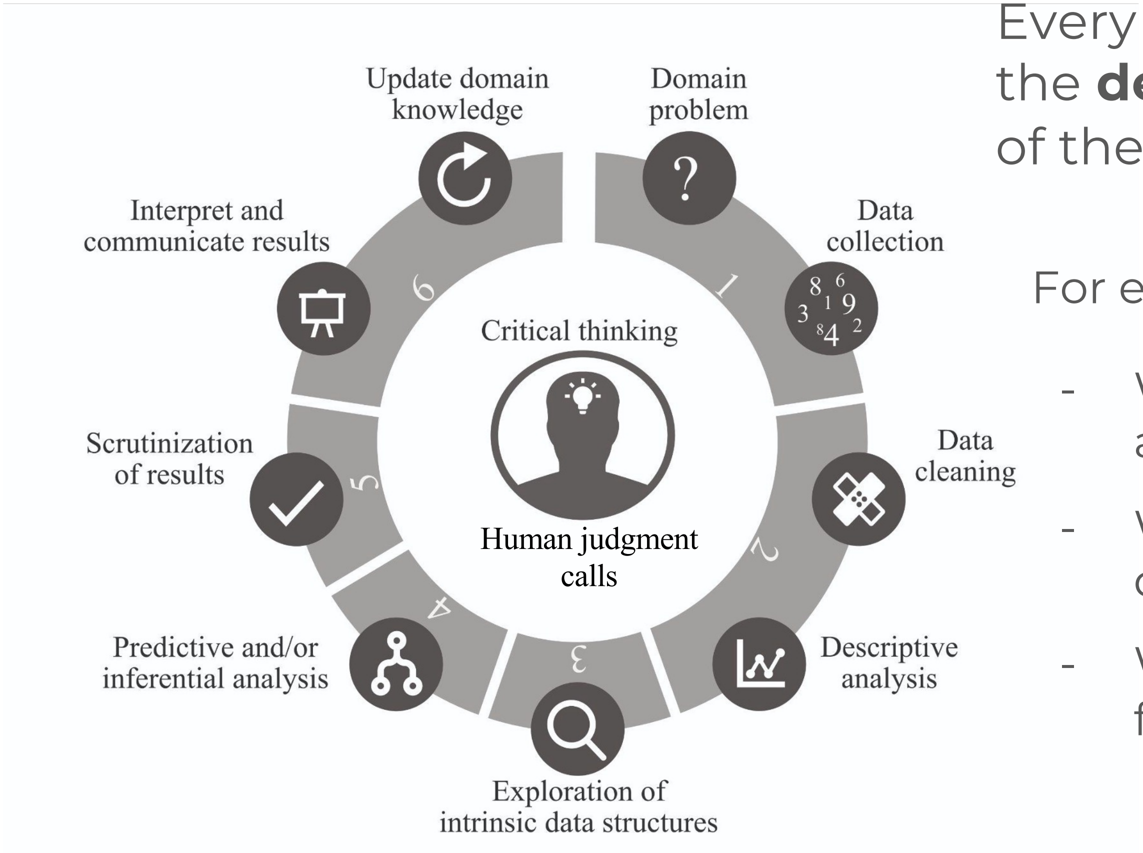
Domain knowledge

3. Mental construct

2. Approximation of reality

1. Reality

The Data Science Life Cycle (DSLC)

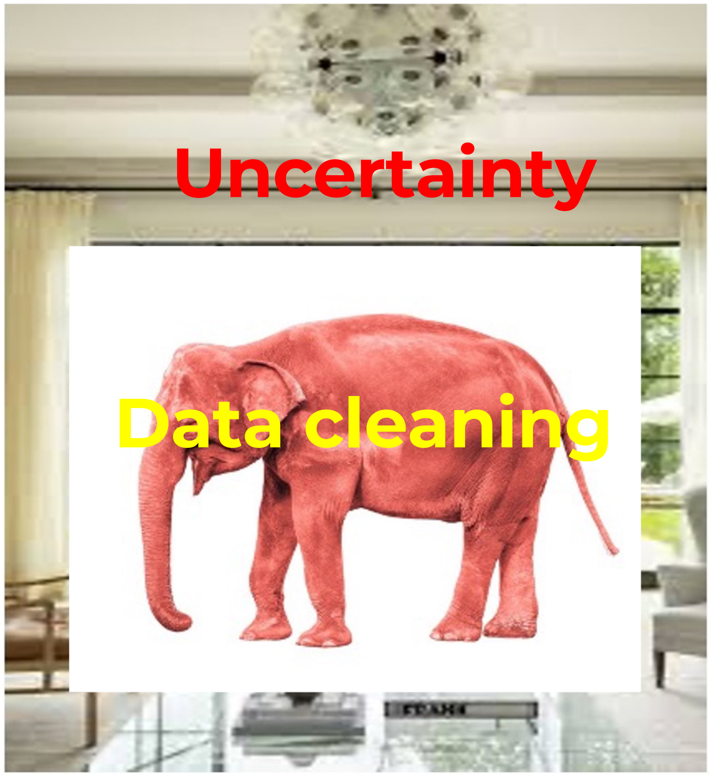


Every data-driven result is a result of the **decisions made at *every* stage** of the Data Science Life Cycle (DSLC)

For example:

- What if a researcher uses data from a different database?
- What if the data had been cleaned differently, or by a different person?
- What if a different model had been fit to generate a predictive algorithm

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/

Veridical Data Science

Veridical data science is the practice of extracting **reliable, reproducible and realistic information from data...**

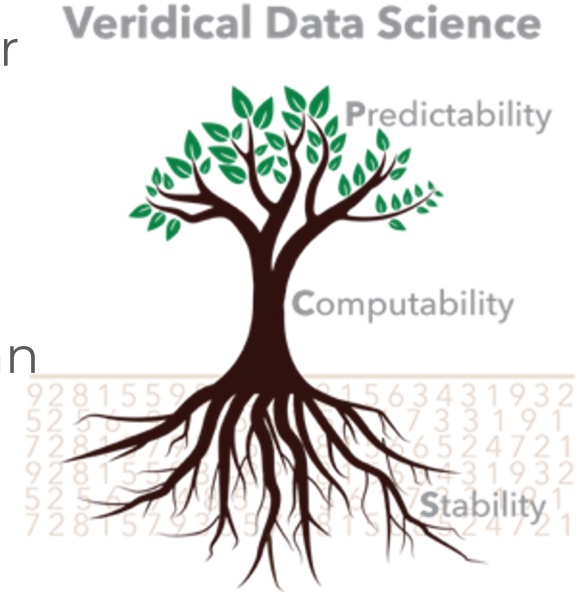
while recognizing that this information is a function of the **judgment calls** made throughout the entire DSLC...

and assessing critically the **impact** of these **judgment calls** on the conclusions made in the **context of a domain problem**

Practicing Veridical Data Science (tree) for Trustworthiness

Practicing veridical data science involves:

- Evaluating **Predictability** of data-driven results for reality check
- Using effective and efficient **Computational** techniques and data-inspired simulations
- Testing the **Stability** of data-driven results to relevant perturbations (data, data cleaning, human judgment calls, etc)



PCS (Yu and Kumbier, PNAS, 2020) for conducting data science **unifies and expands** on ideas and best practices from ML & stats.

PCS principle 1: predictability

Predictability stands in for **reality check (broader)**

Doesn't generate good predictions? ⟹ Probably not capturing real phenomena

Predictability is demonstrated by:

- Showing that a **result re-emerges in new/future data** (or a test set)
- Showing that your analysis/algorithm **uncovers known/expected relationships** based on domain knowledge

Photo from https://catbreedsfaq.com/why-do-cats-ignore-mirrors

PCS principle 2: computability

Computability refers to our ability to **create** and **use** the algorithms/results and to **simulate** using data.

Computable results:

- Are computationally **efficient**
- Are computationally **available** to practitioners
- Includes using data-inspired **simulations to check** relevant theory and algorithms

Photo by Irwan iwe on Unsplash

PCS principle 3: stability

Stability involves investigating how much your results change when:

- Making reasonable **alternate judgment calls**

- Choosing reasonable alternative **problem formulations, models**

- **Perturbing** reasonably your data (e.g. multiple versions of cleaned data, subsamples of data, etc)

PCS documentation

The key to conducting a thorough stability analysis is keeping **detailed documentation** about the different analyses you performed and judgment calls you made (which should be backed up by domain knowledge)

Narrative + code:

- RMarkdown/Quarto
- Jupyter notebook

DS Reproducibility as stability

Best (stable to human analyst)**:**

- Someone else independently replicates the final data results

Second best (stable to implementation of code and results)**:**

- You replicate your own results in the whole process

Third best (code and result review)**:**

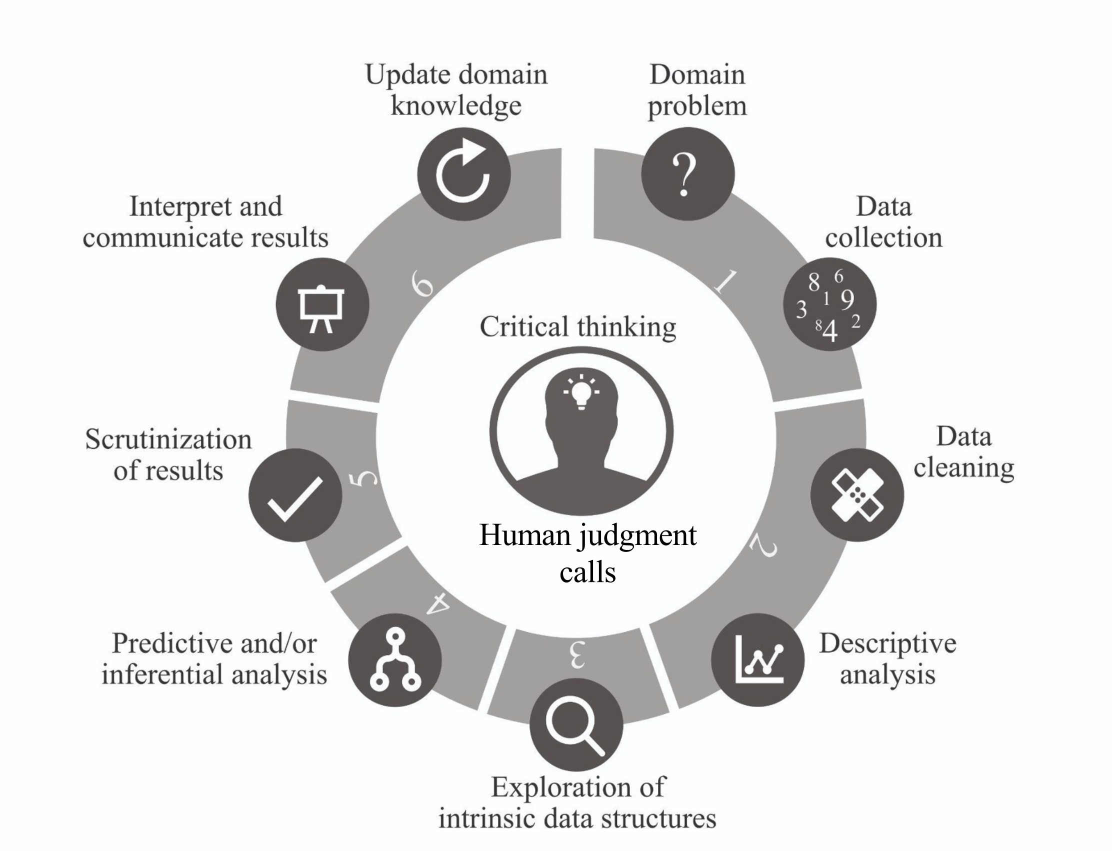
- You or someone else re-reads the code and results

What does “veridical” mean in VDS?

- Veridical means “truthful” in English, and is a common word in Spanish where it implies verified truth.
- A more precise articulation of “veridical” in VDS hinges on both:

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

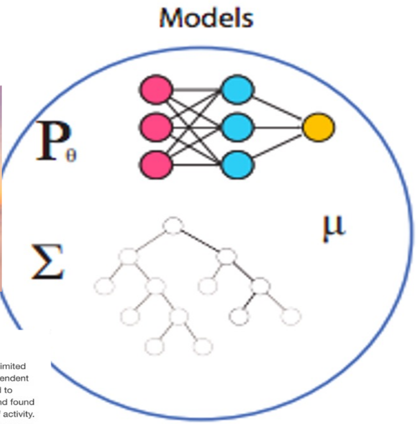
Veridical data science applies **PCS** to every stage of the **Data Science Life Cycle (DSLC)**

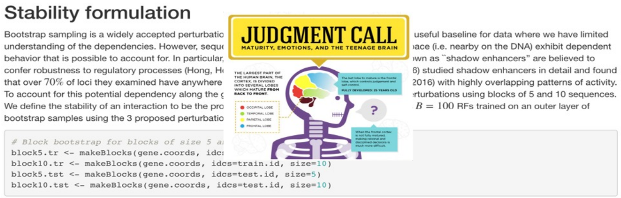

Update domain knowledge
Domain problem
Data collection
Data cleaning
Descriptive analysis
Exploration of intrinsic data structures
Predictive and/or inferential analysis
Scrutinization of results
Interpret and communicate results
Critical thinking
Human judgment calls
1
2
3
4
5
6

PCS **builds trust** in data results

PCS documentation [on GitHub (JupyterNotebook Quarto)]

Reality





Stability formulation

Bootstrap sampling is a widely accepted perturbatio useful baseline for data where we have limited
understanding of the dependencies. However, seque ›ace (i.e. nearby on the DNA) exhibit dependent
behavior that is possible to account for. In particular, ›wn as "shadow enhancers" are believed to
confer robustness to regulatory processes (Hong, H 6) studied shadow enhancers in detail and found
that over 70% of loci they examined have anywhere 2016) with highly overlapping patterns of activity.
To account for this potential dependency along the ‹ rturbations using blocks of 5 and 10 sequences.
We define the stability of an interaction to be the pro $B = 100$ RFs trained on an outer layer of
bootstrap samples using the 3 proposed perturbatio

```
# Block bootstrap for blocks of size 5 ar
block5.tr <- makeBlocks(gene.coords, idcs
block10.tr <- makeBlocks(gene.coords, idcs=train.id, size=10)
block5.tst <- makeBlocks(gene.coords, idcs=test.id, size=5)
block10.tst <- makeBlocks(gene.coords, idcs=test.id, size=10)
```

JUDGMENT CALL
MATURITY, EMOTIONS, AND THE TEENAGE BRAIN
THE LARGEST PART OF THE HUMAN BRAIN, THE CORTEX, IS DIVIDED INTO SEVERAL LOBES WHICH MATURE FROM BACK TO FRONT.
OCCIPITAL LOBE
TEMPORAL LOBE
PARIETAL LOBE
FRONTAL LOBE
The last lobe to mature is the frontal lobe, which controls judgement and self-control.
FULLY DEVELOPED: 25 YEARS OLD
?
When the frontal cortex is not fully matured, making rational and disciplined decisions is much more difficult.

Mental Construct

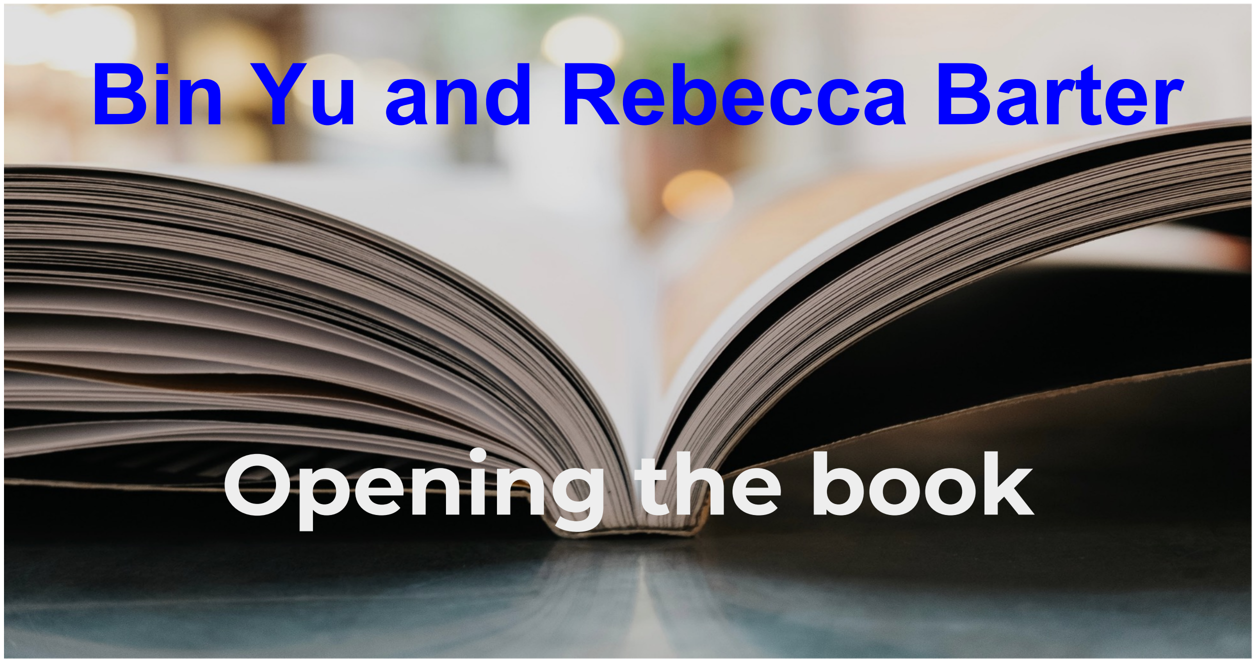
Dangerous
inference & conclusions
Unsubstantiated assumptions
Data

Image credits: Rebecca Barter

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

Veridical Data Science:

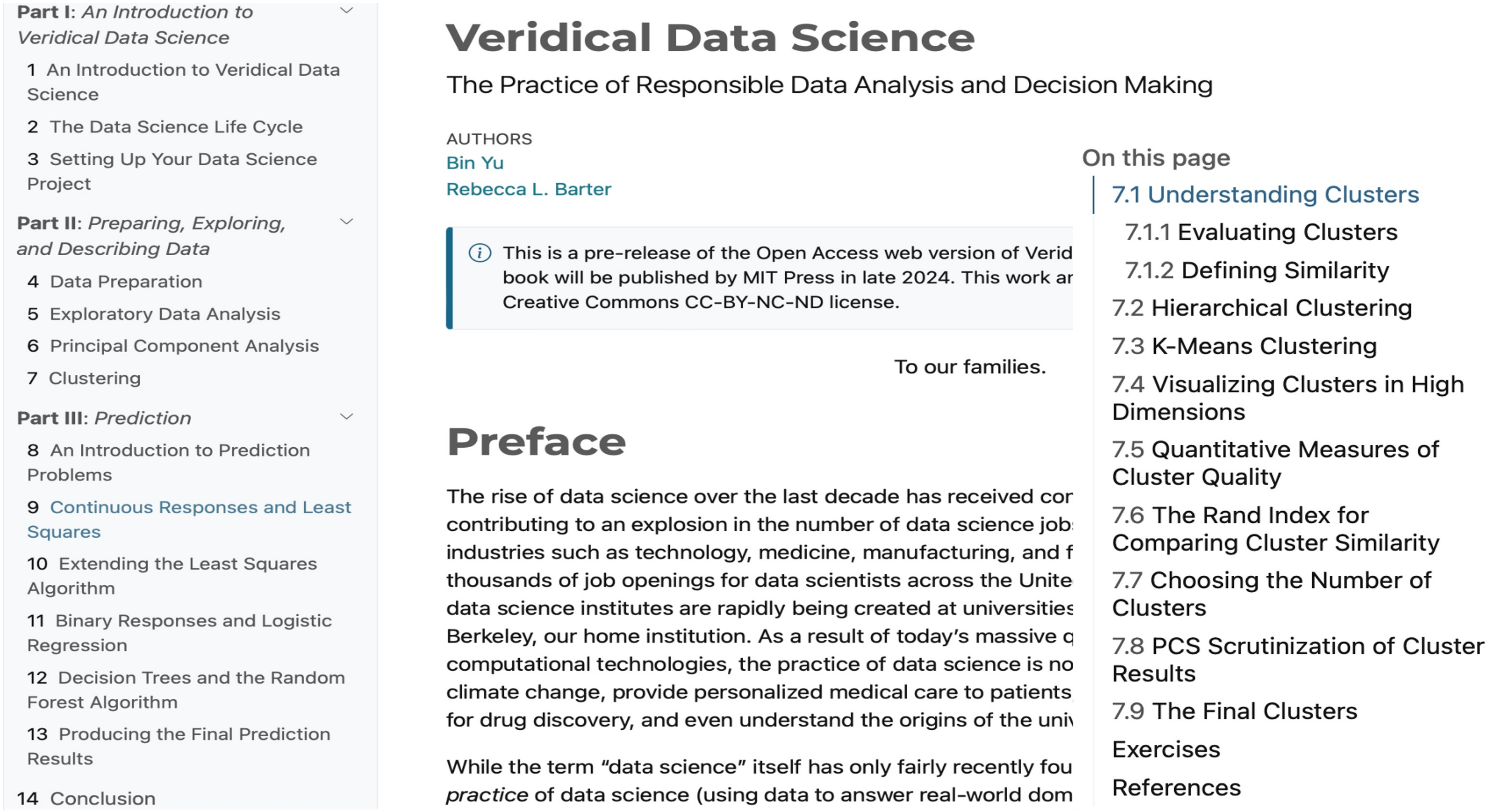
The Practice of Responsible Data Analysis and Decision Making



Bin Yu and Rebecca Barter

Opening the book

Free online version available at vdsbook.com



Veridical Data Science

The Practice of Responsible Data Analysis and Decision Making

AUTHORS

Bin Yu

Rebecca L. Barter

ⓘ This is a pre-release of the Open Access web version of Verid
book will be published by MIT Press in late 2024. This work ar
Creative Commons CC-BY-NC-ND license.

To our families.

Preface

The rise of data science over the last decade has received cor
contributing to an explosion in the number of data science job
industries such as technology, medicine, manufacturing, and f
thousands of job openings for data scientists across the Unite
data science institutes are rapidly being created at universities
Berkeley, our home institution. As a result of today's massive q
computational technologies, the practice of data science is no
climate change, provide personalized medical care to patients
for drug discovery, and even understand the origins of the univ

While the term "data science" itself has only fairly recently fou
practice of data science (using data to answer real-world dom

On this page

Goal of the book

"how to ask insightful data-driven domain questions; how to think critically about data in the context of the question being asked; how to clean and format your data; how to structure your analyses; and how to scrutinize the trustworthiness of your resulting data-driven results, while also providing an intuitive introduction to a range of common statistical and ML algorithms ..."

– Yu and Barter (2024)

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

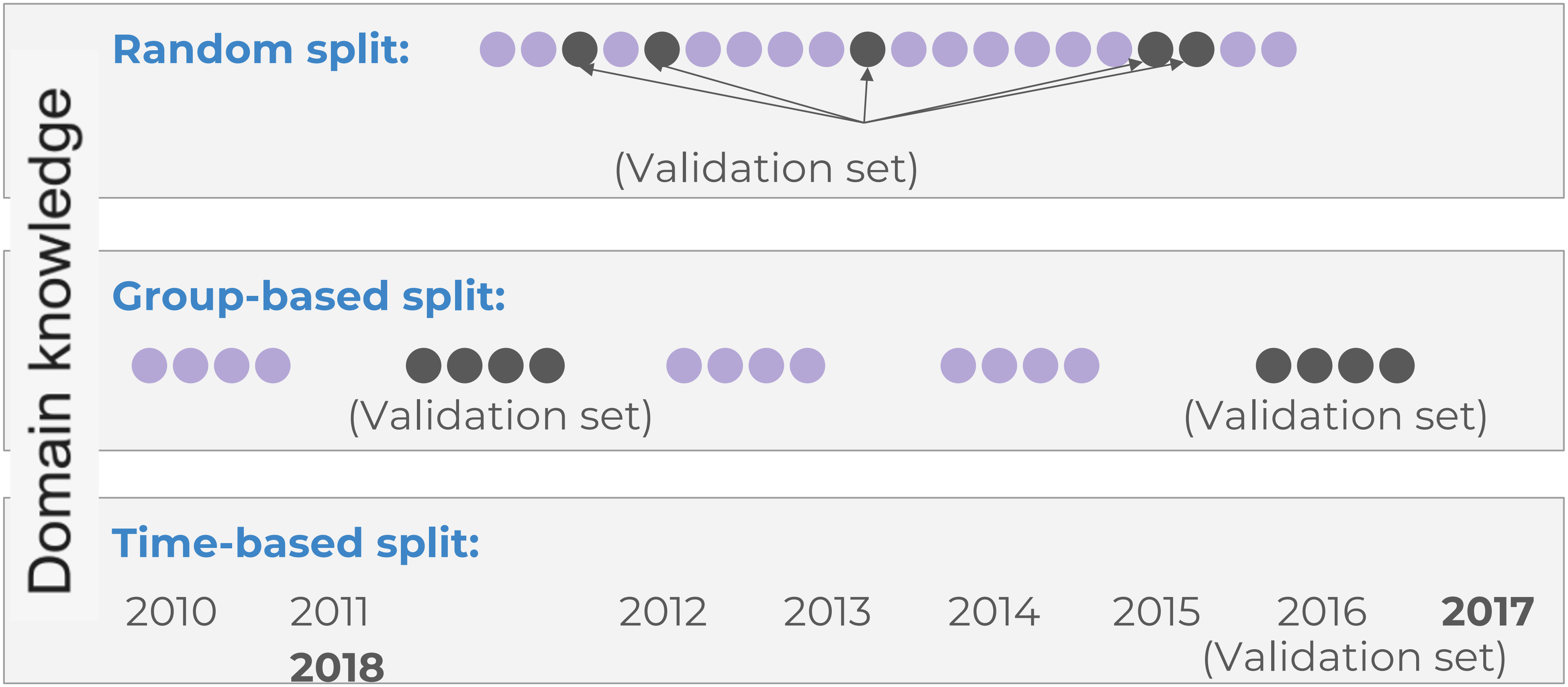
PCS is in every chapter, and so is documentation

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 pt 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)

Reality checking (P) via train/val/test splits



Domain knowledge

Random split:

(Validation set)

Group-based split:

(Validation set)

(Validation set)

Time-based split:

2010 2011 2012 2013 2014 2015 2016 **2017**

2018

(Validation set)

Case studies and project exercises

We use publicly available real-world datasets to build case studies and project exercises from areas that all students can relate to so they have some “domain knowledge”.

Organ donation data
Nutrition data
Ames house data

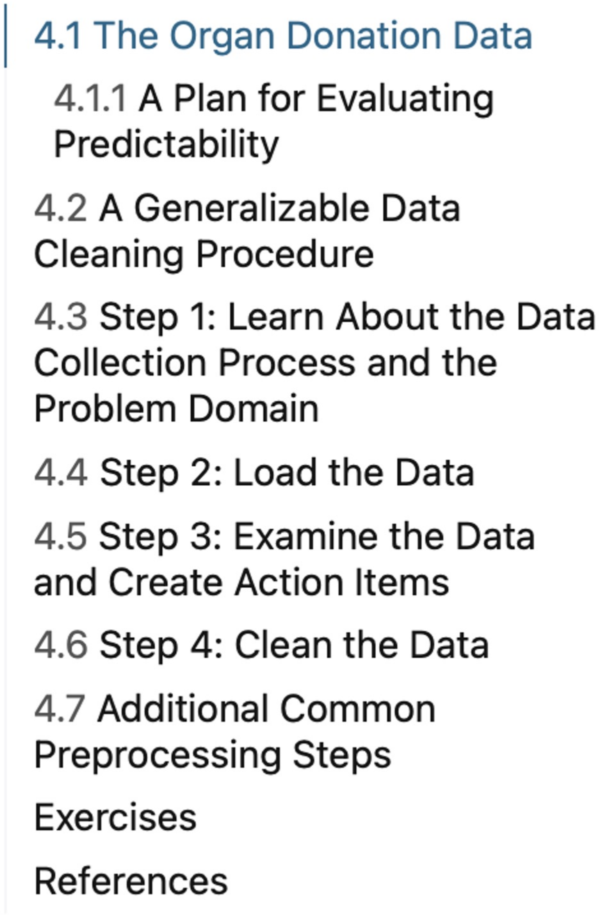
Codes and data available

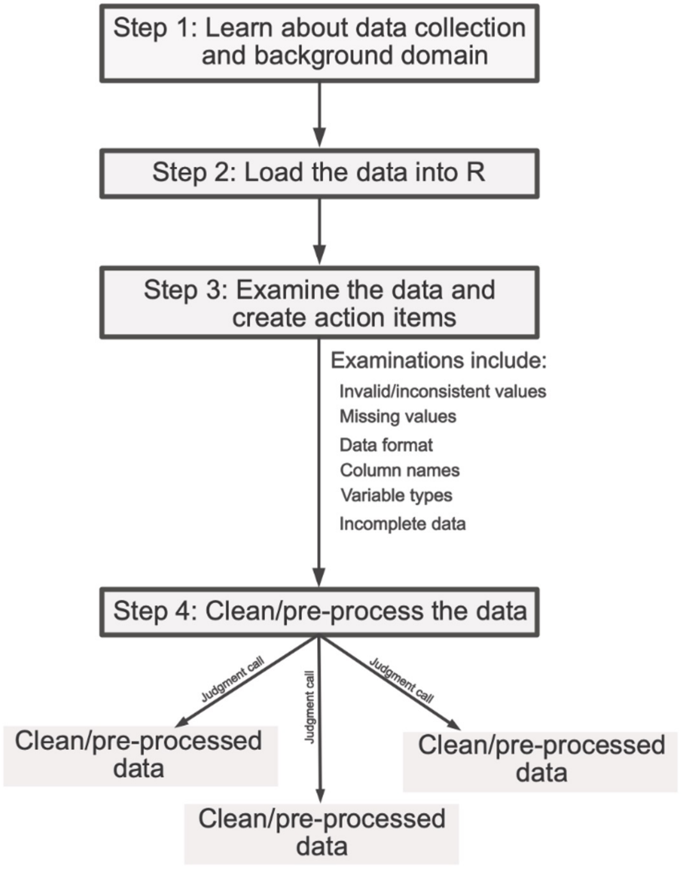
Supplementary R and Python code as well as the data that accompanies each case study in this book can be found on GitHub

(https://github.com/Yu-Group/vds-book-supplementary).

The exercises in Chapter 3 will walk you through “cloning” (downloading) this GitHub repository to your computer.

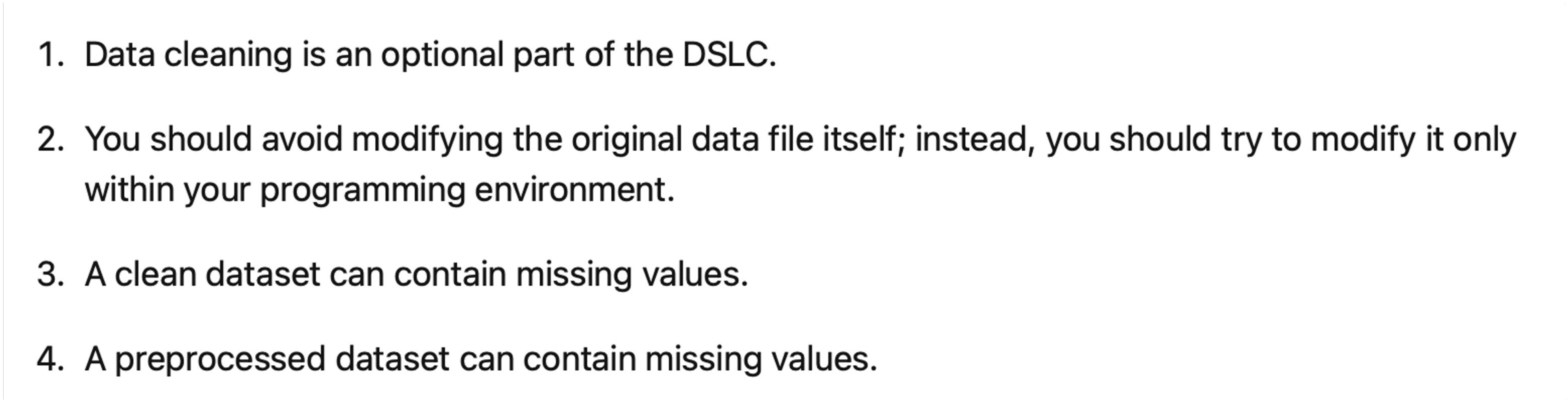
Chapter on data preparation





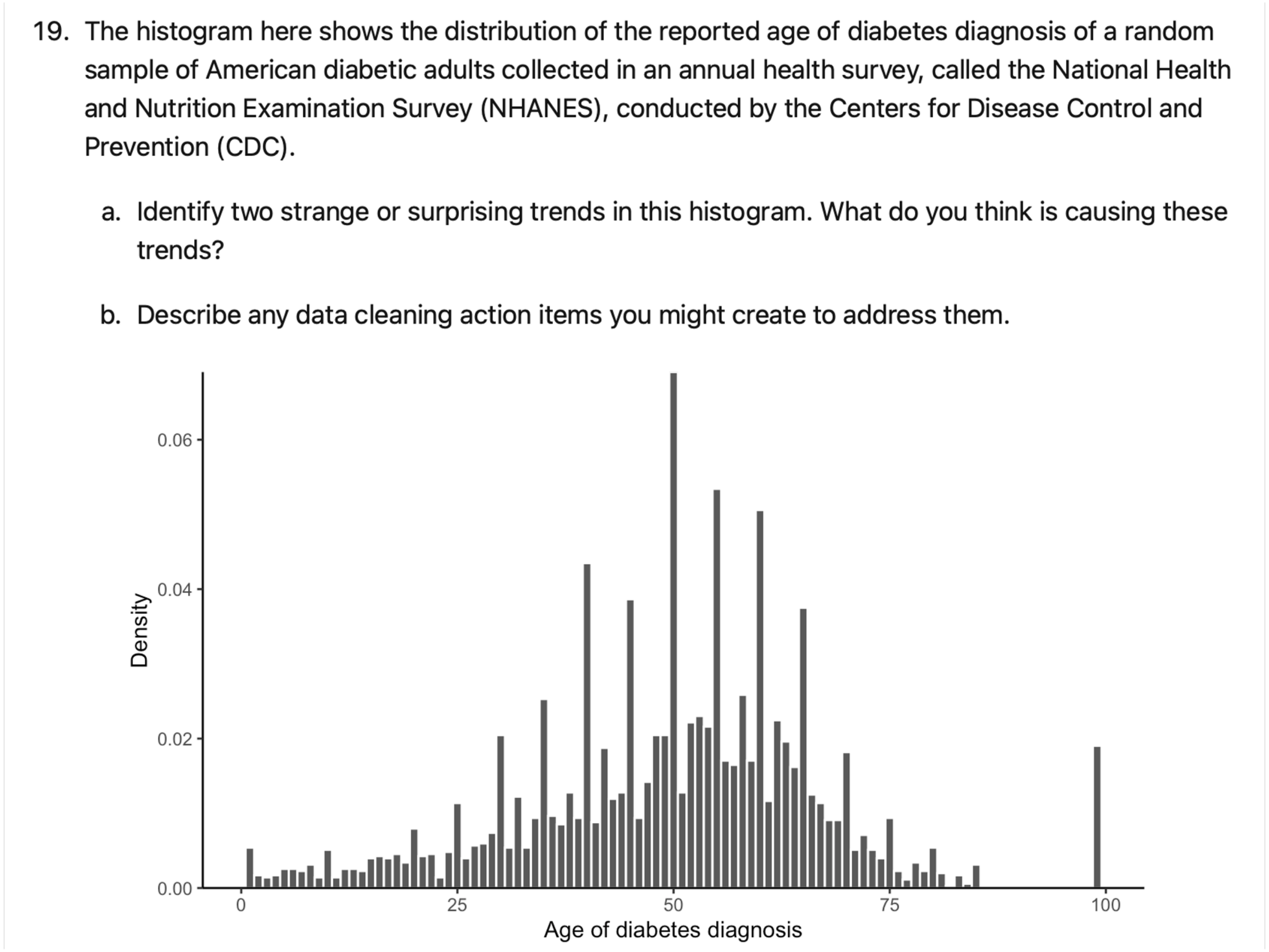
Multiple cleaned datasets to assess uncertainty arising from data leaning.

1. True and False Exercises



1. Data cleaning is an optional part of the DSLC.
2. You should avoid modifying the original data file itself; instead, you should try to modify it only within your programming environment.
3. A clean dataset can contain missing values.
4. A preprocessed dataset can contain missing values.

2. Conceptual Exercises

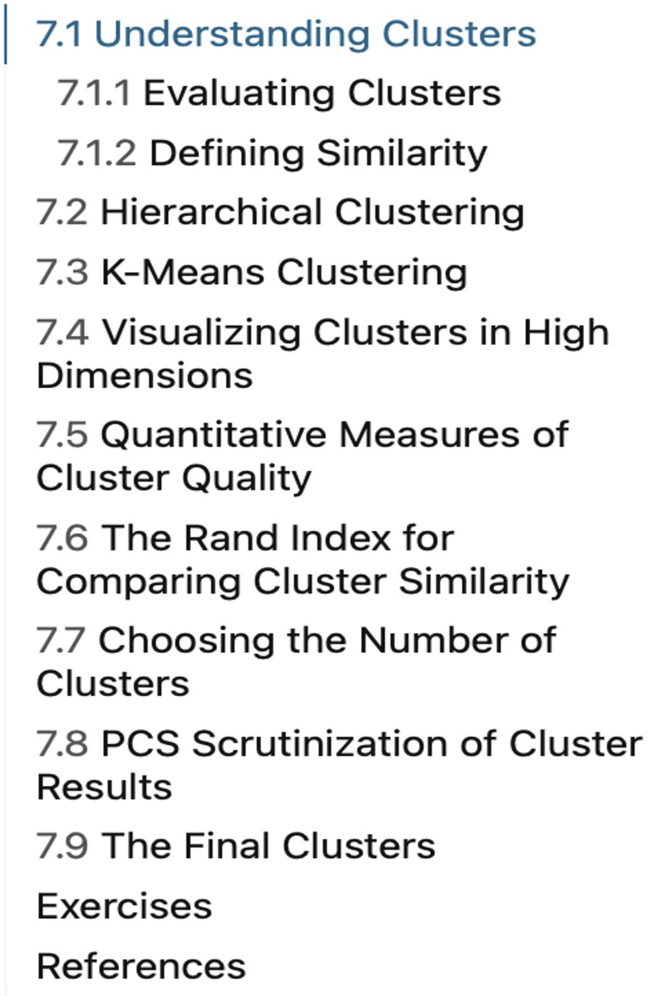


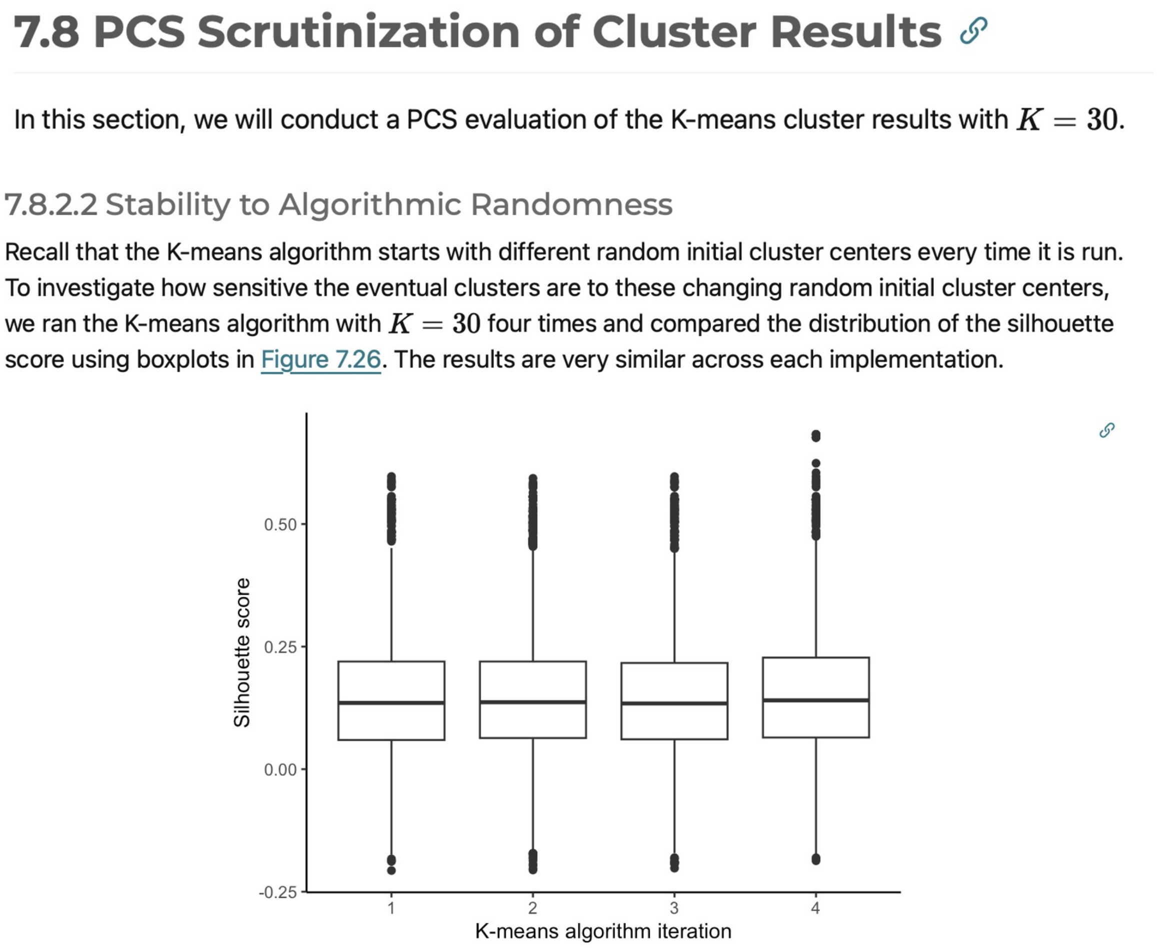
19. The histogram here shows the distribution of the reported age of diabetes diagnosis of a random sample of American diabetic adults collected in an annual health survey, called the National Health and Nutrition Examination Survey (NHANES), conducted by the Centers for Disease Control and Prevention (CDC).

    a. Identify two strange or surprising trends in this histogram. What do you think is causing these trends?

    b. Describe any data cleaning action items you might create to address them.

Chapter on clustering





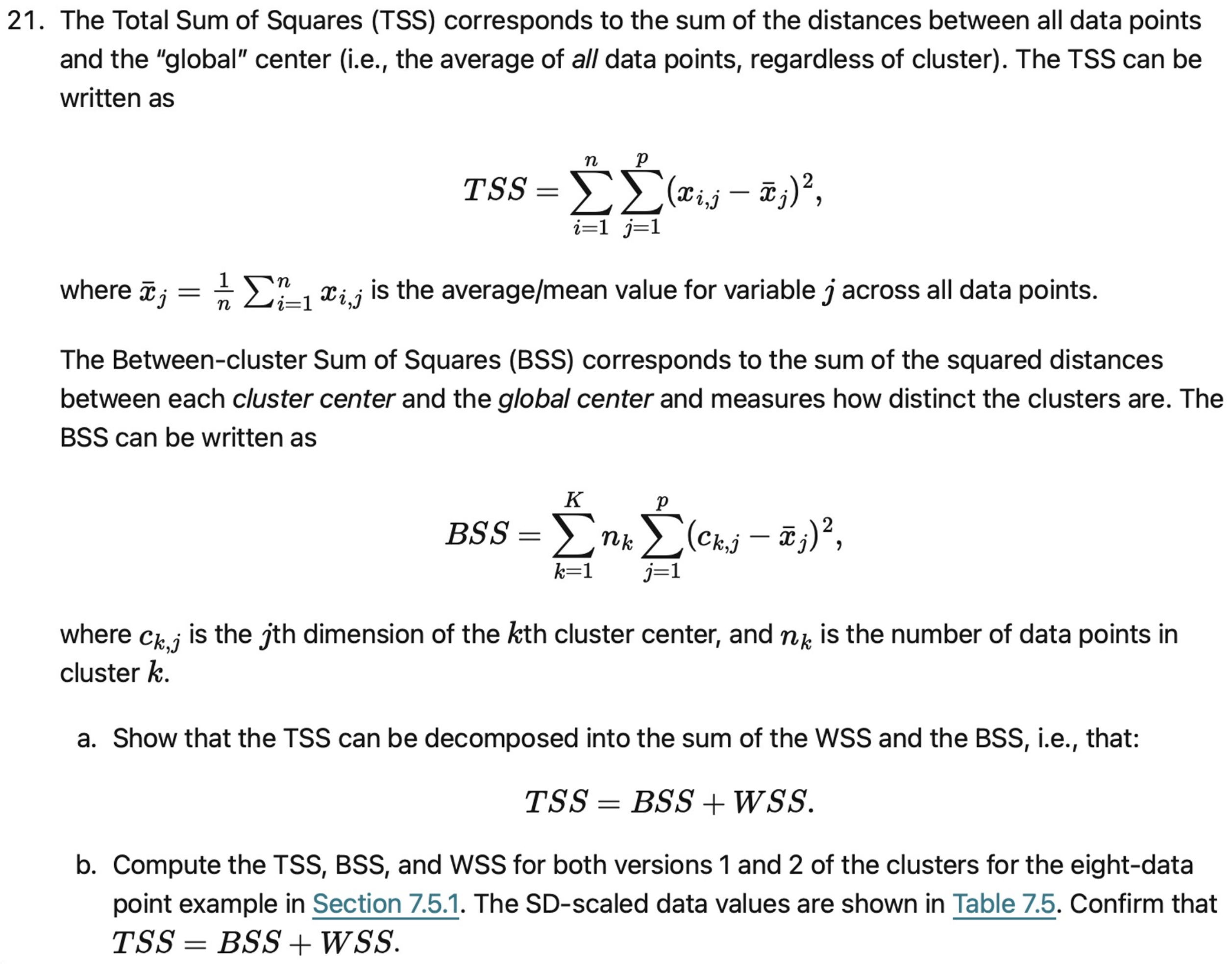
7.8 PCS Scrutinization of Cluster Results

In this section, we will conduct a PCS evaluation of the K-means cluster results with $K = 30$.

7.8.2.2 Stability to Algorithmic Randomness

Recall that the K-means algorithm starts with different random initial cluster centers every time it is run. To investigate how sensitive the eventual clusters are to these changing random initial cluster centers, we ran the K-means algorithm with $K = 30$ four times and compared the distribution of the silhouette score using boxplots in Figure 7.26. The results are very similar across each implementation.

3. Mathematical Exercises



21. The Total Sum of Squares (TSS) corresponds to the sum of the distances between all data points and the "global" center (i.e., the average of *all* data points, regardless of cluster). The TSS can be written as

$$TSS = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{i,j} - \bar{x}_j)^2,$$

where $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}$ is the average/mean value for variable j across all data points.

The Between-cluster Sum of Squares (BSS) corresponds to the sum of the squared distances between each *cluster center* and the *global center* and measures how distinct the clusters are. The BSS can be written as

$$BSS = \sum_{k=1}^{K} n_k \sum_{j=1}^{p} (c_{k,j} - \bar{x}_j)^2,$$

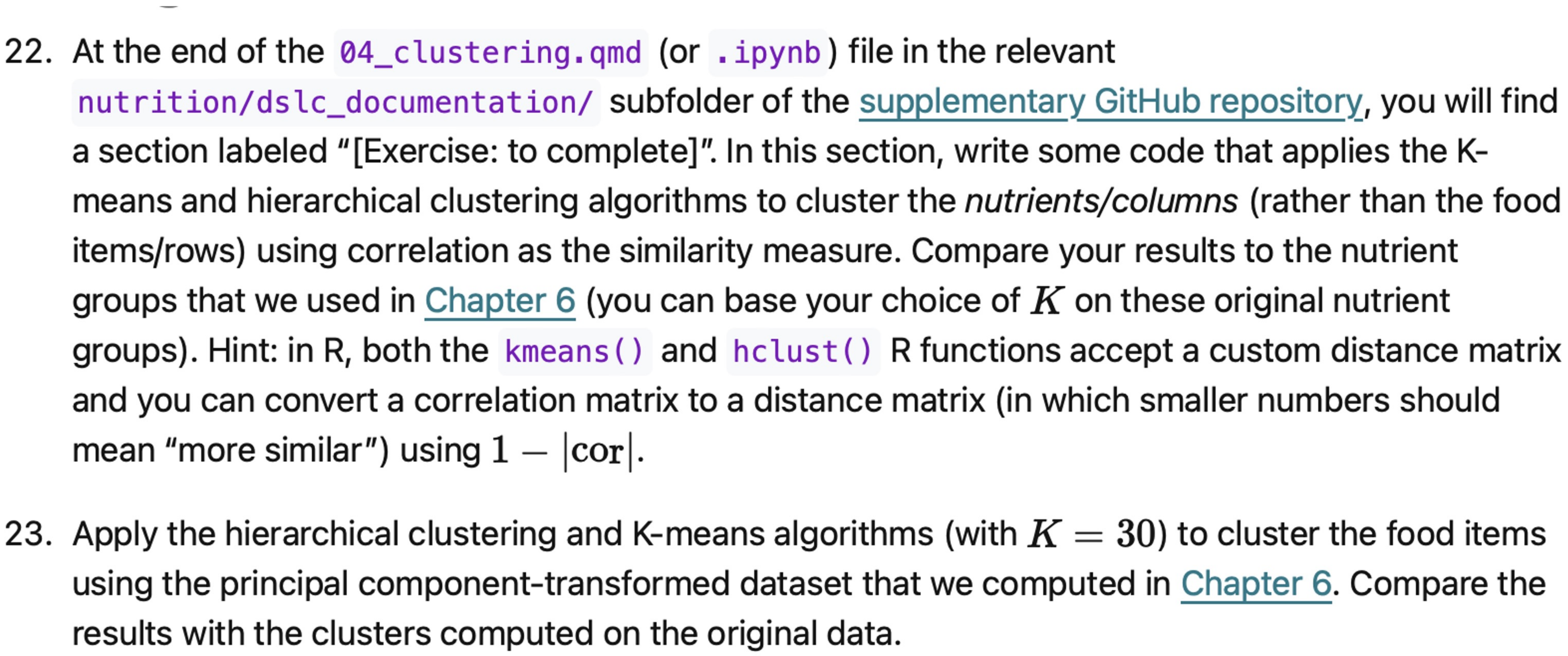
where $c_{k,j}$ is the jth dimension of the kth cluster center, and n_k is the number of data points in cluster k.

a. Show that the TSS can be decomposed into the sum of the WSS and the BSS, i.e., that:

$$TSS = BSS + WSS.$$

b. Compute the TSS, BSS, and WSS for both versions 1 and 2 of the clusters for the eight-data point example in Section 7.5.1. The SD-scaled data values are shown in Table 7.5. Confirm that $TSS = BSS + WSS.$

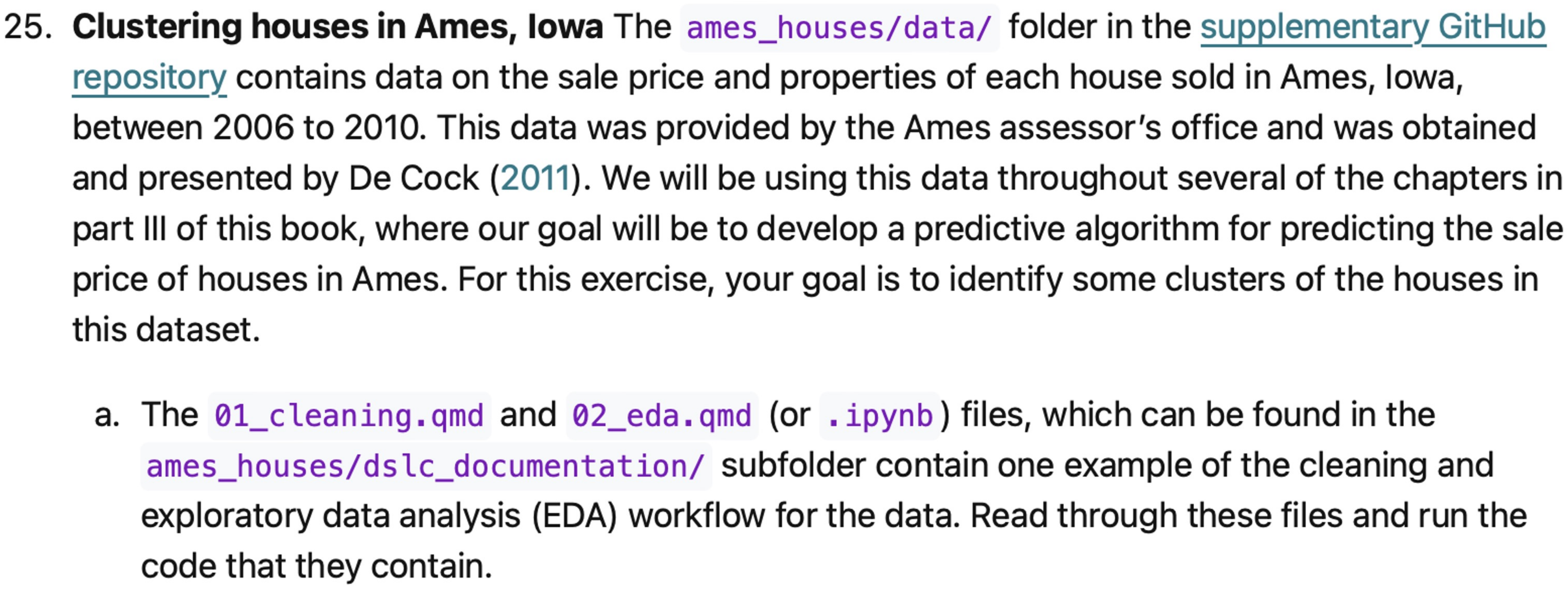
4. Coding Exercises



22. At the end of the `04_clustering.qmd` (or `.ipynb`) file in the relevant `nutrition/dslc_documentation/` subfolder of the supplementary GitHub repository, you will find a section labeled "[Exercise: to complete]". In this section, write some code that applies the K-means and hierarchical clustering algorithms to cluster the *nutrients/columns* (rather than the food items/rows) using correlation as the similarity measure. Compare your results to the nutrient groups that we used in Chapter 6 (you can base your choice of K on these original nutrient groups). Hint: in R, both the `kmeans()` and `hclust()` R functions accept a custom distance matrix and you can convert a correlation matrix to a distance matrix (in which smaller numbers should mean "more similar") using $1 - |\text{cor}|$.

23. Apply the hierarchical clustering and K-means algorithms (with $K = 30$) to cluster the food items using the principal component-transformed dataset that we computed in Chapter 6. Compare the results with the clusters computed on the original data.

5. Project Exercises

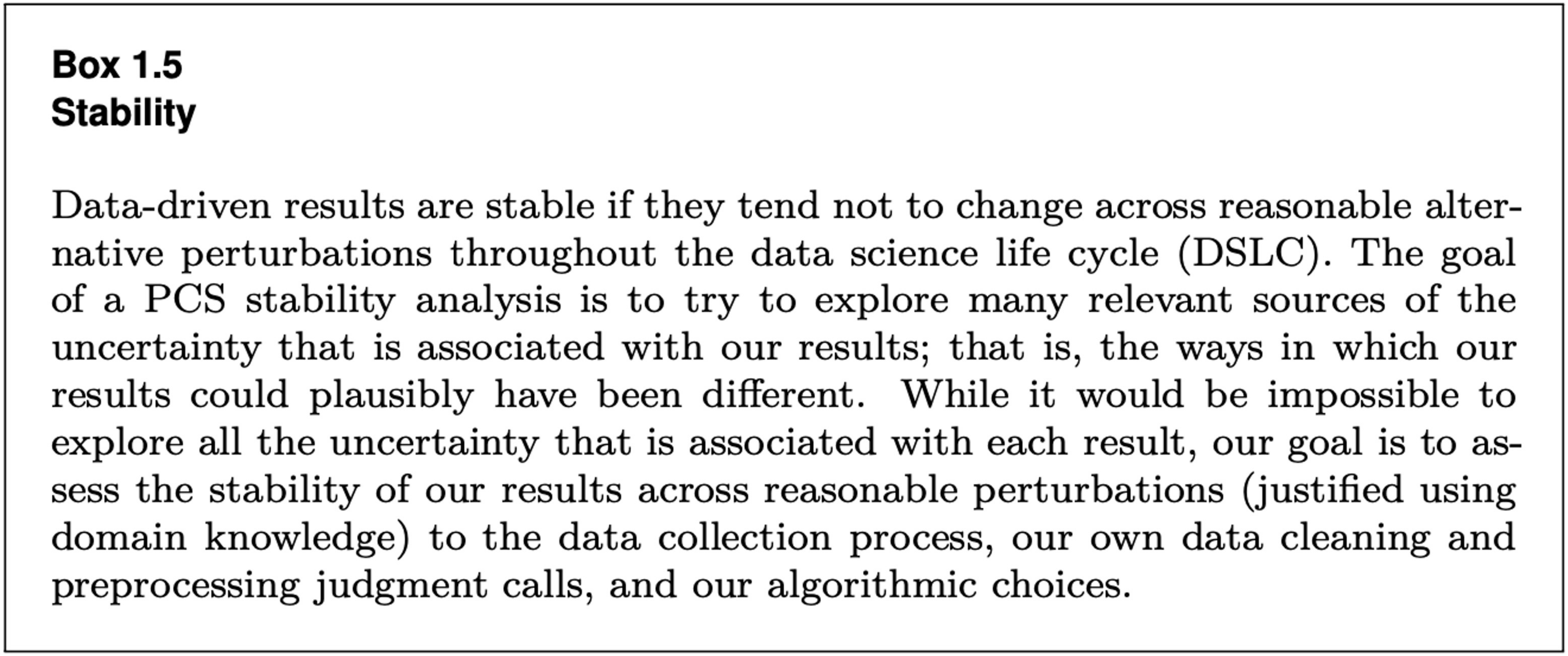


25. **Clustering houses in Ames, Iowa** The `ames_houses/data/` folder in the supplementary GitHub repository contains data on the sale price and properties of each house sold in Ames, Iowa, between 2006 to 2010. This data was provided by the Ames assessor's office and was obtained and presented by De Cock (2011). We will be using this data throughout several of the chapters in part III of this book, where our goal will be to develop a predictive algorithm for predicting the sale price of houses in Ames. For this exercise, your goal is to identify some clusters of the houses in this dataset.

    a. The `01_cleaning.qmd` and `02_eda.qmd` (or `.ipynb`) files, which can be found in the `ames_houses/dslc_documentation/` subfolder contain one example of the cleaning and exploratory data analysis (EDA) workflow for the data. Read through these files and run the code that they contain.

We use case studies and project exercises from areas that all students can relate to (e.g. organ donor, nutrients, house prices).

Use of boxes for emphasis



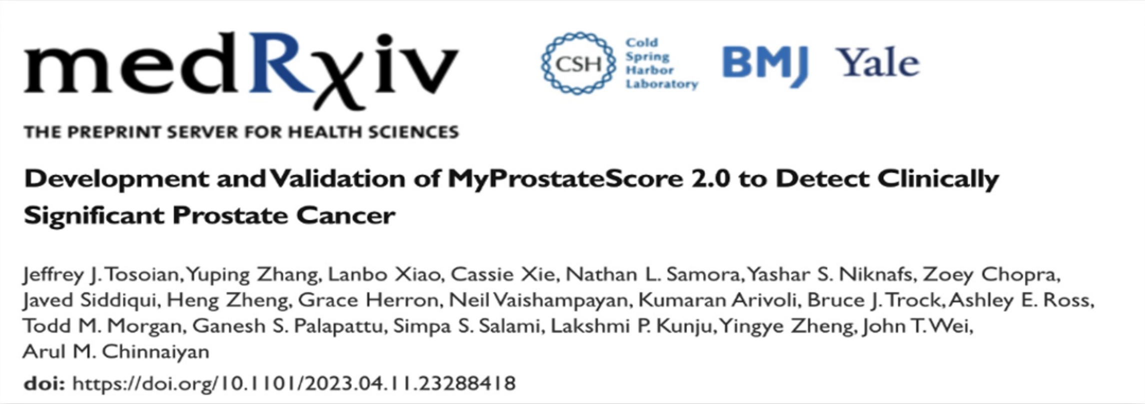
Box 1.5
Stability

Data-driven results are stable if they tend not to change across reasonable alternative perturbations throughout the data science life cycle (DSLC). The goal of a PCS stability analysis is to try to explore many relevant sources of the uncertainty that is associated with our results; that is, the ways in which our results could plausibly have been different. While it would be impossible to explore all the uncertainty that is associated with each result, our goal is to assess the stability of our results across reasonable perturbations (justified using domain knowledge) to the data collection process, our own data cleaning and preprocessing judgment calls, and our algorithmic choices.

Case Study:
Prostate cancer detection

Prostate cancer detection

MyProstateScore2.0 (MPS2)



medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

Development and Validation of MyProstateScore 2.0 to Detect Clinically Significant Prostate Cancer

Jeffrey J. Tosoian, Yuping Zhang, Lanbo Xiao, Cassie Xie, Nathan L. Samora, Yashar S. Niknafs, Zoey Chopra, Javed Siddiqui, Heng Zheng, Grace Herron, Neil Vaishampayan, Kumaran Arivoli, Bruce J. Trock, Ashley E. Ross, Todd M. Morgan, Ganesh S. Palapattu, Simpa S. Salami, Lakshmi P. Kunju, Yingye Zheng, John T. Wei, Arul M. Chinnaiyan

doi: https://doi.org/10.1101/2023.04.11.23288418

Yu Group (**Tiffany Tang** and **Ana Kenny**) has been collaborating with **Yuping Zhang** and **Arul Chinnaiyan.**

Tang

Kenney

Zhang

Chinnaiyan

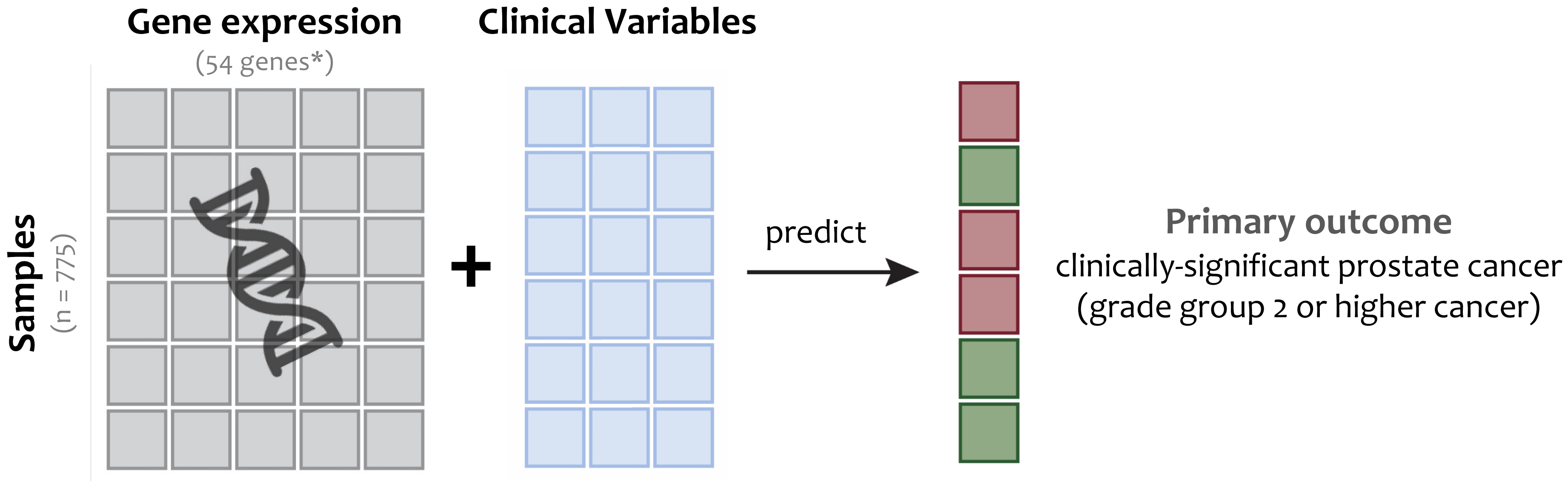
MICHIGAN

Berkeley
UNIVERSITY OF CALIFORNIA

Overview of data

Training / Development Cohort

(age, race, family history, abnormal DRE, prior negative biopsy, prostate volume)



Gene expression

(54 genes*)

Clinical Variables

Samples (n = 775)

+

predict

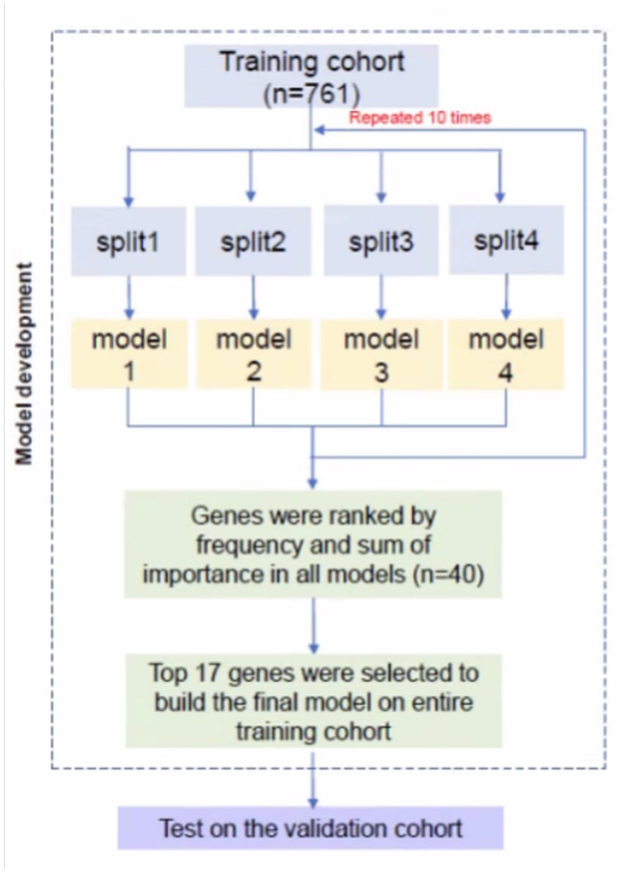
Primary outcome

clinically-significant prostate cancer
(grade group 2 or higher cancer)

Validation Cohorts: patients participating in the prospective NCI-EDRN PCA3 Evaluation Trial, UM-MRI cohort

MyProstateScore2.0 (MPS2) development

(Tosoian and Zhang, ..., Wei and Chinnaiyan, 2023; JAMA 2024)



Model development

Training cohort (n=761)

Repeated 10 times

split1

split2

split3

split4

model 1

model 2

model 3

model 4

Genes were ranked by frequency and sum of importance in all models (n=40)

Top 17 genes were selected to build the final model on entire training cohort

Test on the validation cohort

Prediction Model: **Logistic elastic net**

Chose **17 genes (+ *KLK3*)** due to technical features of the OpenArray™ platform, and use clinical variables (excluding prostate vol.)

MPS2 results on EDRN external validation cohort

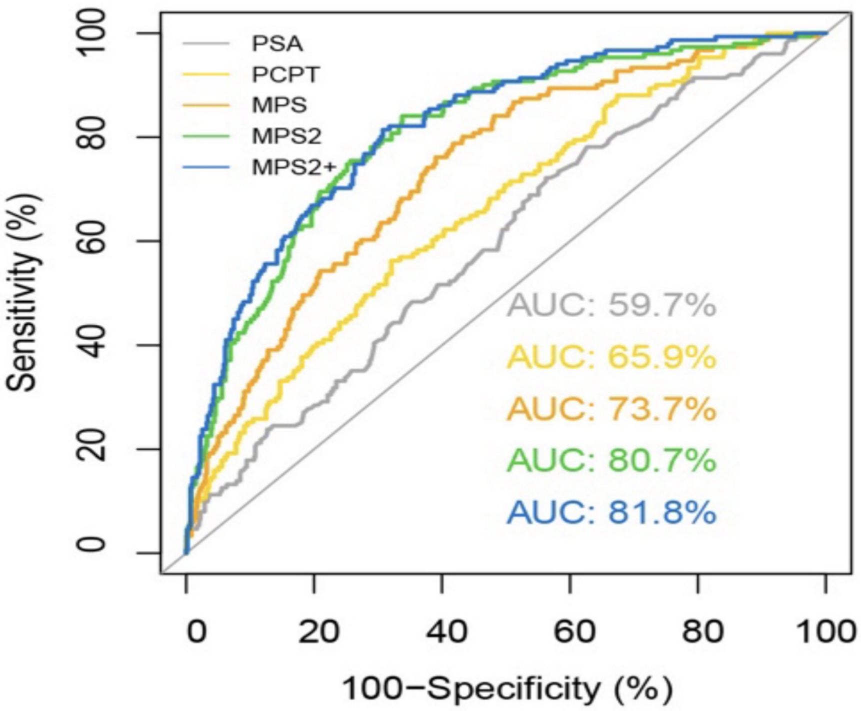
(Tosoian and Zhang, ..., Wei and Chinnaiyan, 2023)

On EDRN external validation cohort (n=859)

MPS2 and MPS2+ yield the highest AUC of 80.7% and 81.8%, respectively

This is significantly higher, by 21, 15, 7%, than the AUCs of

- PSA (59.7%) (current standard)
- PCPT (65.9%) (only clinical variables)
- MPS (73.7%) (only two genes+clinical)



(false positive rate)

PCS stress-test recipe

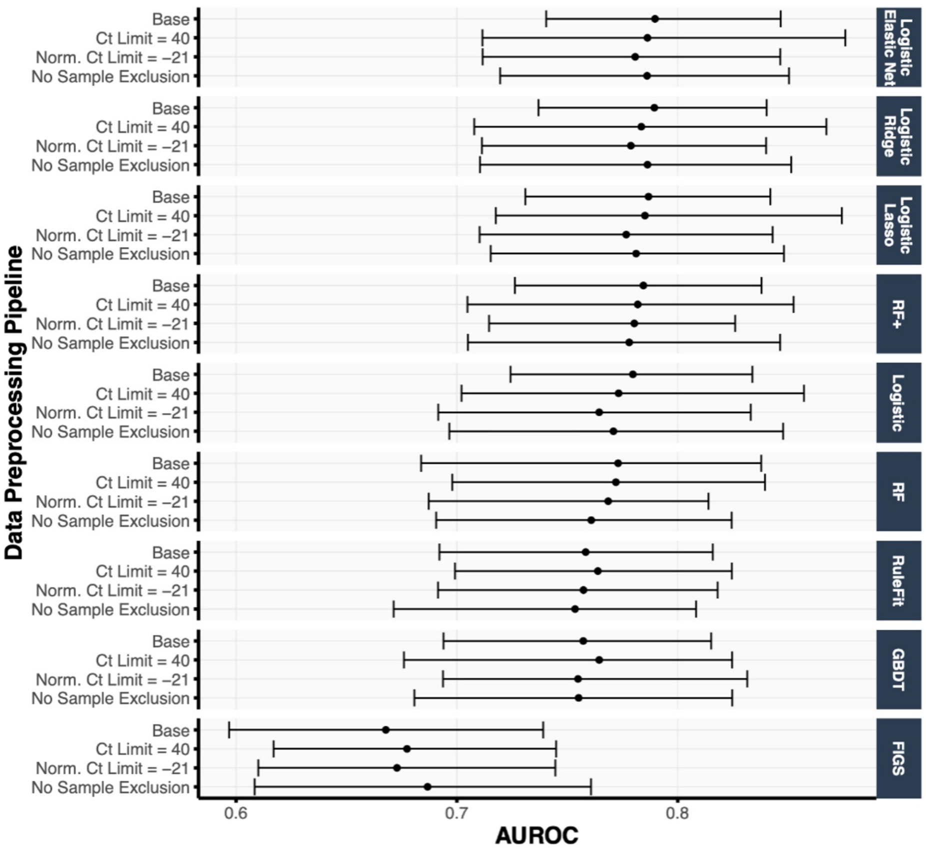
1. **Identify (reasonable) human judgment calls** and **perturbations** in the pipeline

2. **Perform perturbations** within resource constraints

3. For each perturbation, **evaluate performance metrics**
   a. “P” or Pred-check
   b. Gene importance ranking (to seek “S” or stable ranking after Pred-check)
   c. Others (e.g., subgroup accuracies, …)

Pred-check accuracies of MPS2 are stable

(data cleaning uncertainty 1-2%, compared to 2-10% improvements)



Removed RuleFit, GBDT, and FIGS due to prediction check

Seeking stable genes after Pred-check

For each feature, compute the **mean feature importance ranking** across all Pred-checked models (i.e., RF, RF+, logistic, logistic L1/L2/elastic net)

Rank features by this mean (stable) feature importance ranking

Use top-k-ranked features (+ clinical variables) to make predictions

Using these stability-driven genes, test prediction performance is similar and interpretability/sparsity and cost improve significantly.

Simplified MPS2 or sMPS2 (Tang et al, 2024)

7 genes (+KLK3) based on stable ranking + clinical variables

T2ERG, SCHLAP1, OR51E2, PCAT14, TFF3, PCA3, APOC1

All known prostate cancer genes.

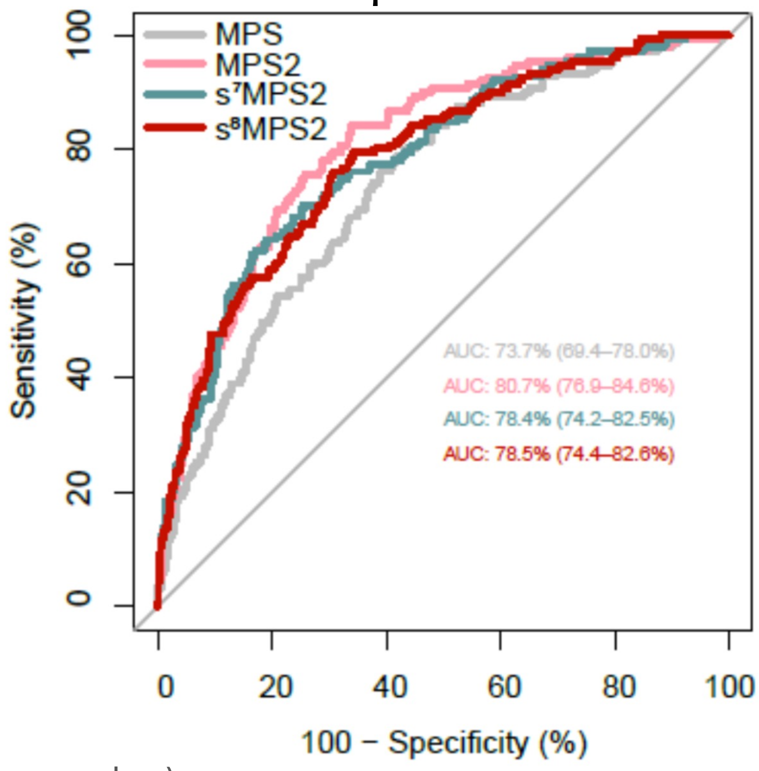
Similar performance on development cohort with 17-gene (+KLK3) and clinical variables for MPS2.

Bin’s note on July 19, 2025: Tang et al (2024) has been published in Cancer Biomarkers in 2025:
https://pubmed.ncbi.nlm.nih.gov/40109218/

Proxy to future data: EDRN external validation cohort

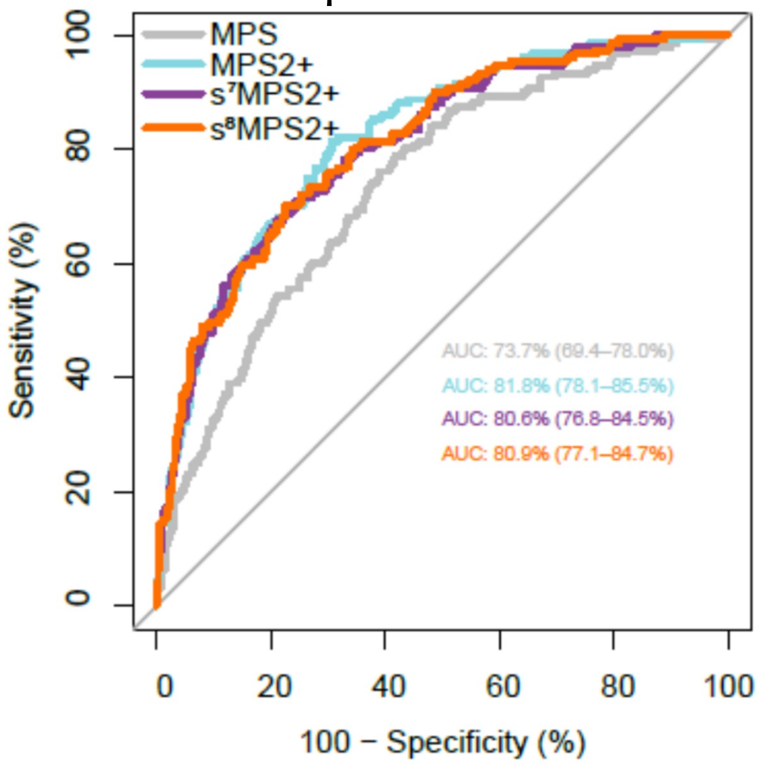
Comparable results (1-2% diff within 1-2% data cleaning uncertainty)

Without prostate volume



(false positive rate)

With prostate volume



(false positive rate)

(n = 743 samples)

Declaration of potential COI

A patent will be filed on sMPS2 model to facilitate its potential commercial and clinical development.

I am named as a co-inventor on the disclosure (as a step towards the patent application).

“Secret sauce” of PCS-guided DS or VDS

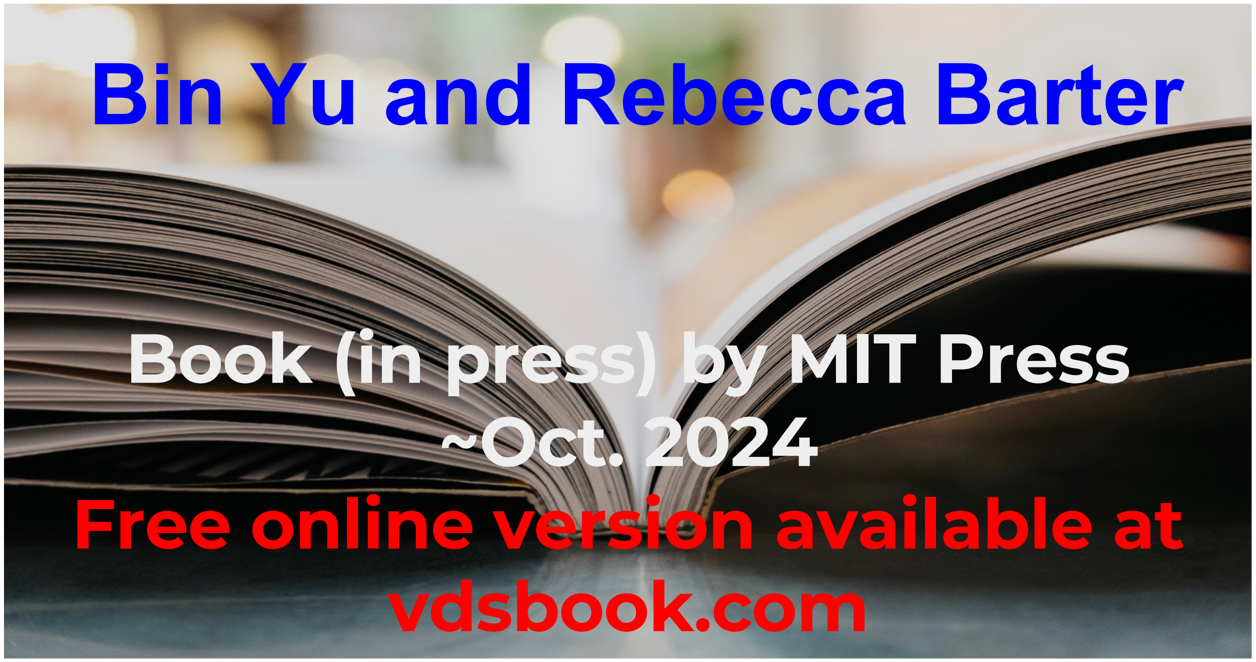
- Data science hygiene for the DSLC

  (for each step: critical thinking, UQ when possible, and documentation)

- Modeling step: “Get *multiple* opinions from **good** doctors and integrate”

  Sensible aggregations in context of results from *multiple* **Pred-screened** algorithms and over data perturbations such as bootstrap when deemed reasonable in context.

Veridical Data Science:
The Practice of Responsible Data Analysis and Decision Making

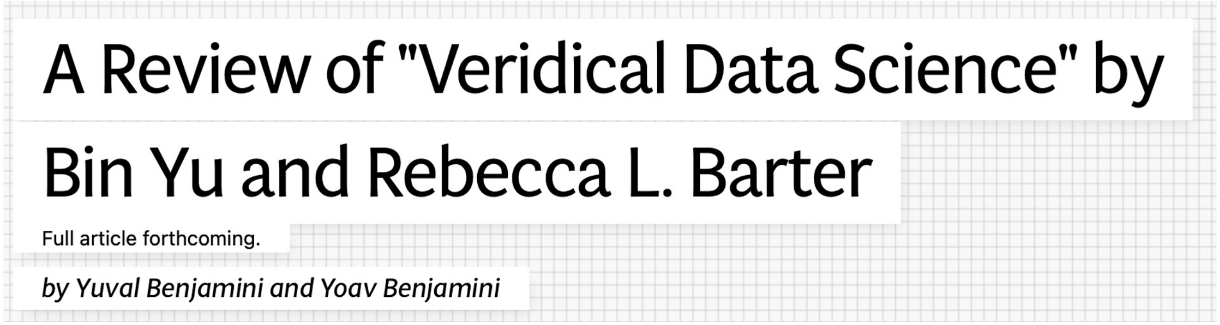


Bin Yu and Rebecca Barter

Book (in press) by MIT Press ~Oct. 2024

Free online version available at vdsbook.com

Thanks to Yuval and Yoav Benjamini
Thanks to Xiao-li and HDSR



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

Full article forthcoming.

by Yuval Benjamini and Yoav Benjamini

*“**Editor-in-Chief’s Note:** In this **inaugural book review** for* Harvard Data Science Review, *Yuval Benjamini and Yoav Benjamini provide a succinct summary and insightful reflection on* Veridical Data Science *by Bin Yu and Rebecca Barter (2024). … 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.”*

Thank you all!

NSF

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951

Office of Naval Research
ONR
Science & Technology

CHAN ZUCKERBERG
BIOHUB