Veridical Data Science

Bin Yu
Statistics and EECS, UC Berkeley

Breiman Lecture, NeurIPS
Vancouver, Dec. 10, 2019
**veridical**

/ˈvɛrɪdɪk(ə)l/  

*adjective*  

FORMAL  

truthful.  

- coinciding with reality.  
  "such memories are not necessarily veridical"
Leo Breiman (1928-2005): a data scientist and a modern day polymath
Statistical Modeling: The Two Cultures

Leo Breiman

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from response variables $= f$(predictor variables, random noise, parameters)

The Algorithmic Modeling Culture

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

Alexa, Siri, ...
Wearable health devices
Streaming videos, on-line gaming, ...
On-line news
Self-driving cars
Election campaigns
Precision medicine
Biology
Neuroscience
Cosmology
Material science
Chemistry
Law
Political science
Economics
Sociology
...
Biomedical data problems are pressing

https://deepmind.com/blog/alphafold/

https://medium.com/
Data science is a key element of AI

Conway’s Venn Diagram

Goal:

combine data with domain knowledge to make decisions and generate new knowledge
DS Life Cycle (DSLC): a system

Image credits: R. Barter and toronto4kids.com
Veridical Data Science

Extracts reliable and reproducible information from data, with an enriched technical language to communicate and evaluate empirical evidence in the context of human decisions and domain knowledge.
Rest of the talk

• PCS framework for veridical data science

• Iterative random forests

• PDR framework for interpretable machine learning

• ACD for interpreting DNNs
PCS framework for veridical data science
PCS framework Y. and Kumbier (2019)

Three principles of data science: PCS

Predictability (P) (from ML)

Computability (C) (from ML)

Stability (S) (from statistics)

PCS bridges Breiman’s two cultures

Veridical Data Science

Predictability

Computability

Stability

Image credit: R. Barter
PCS connects science with engineering

• **Predictability** and **stability** embed two scientific principles: prediction and replication

• **Computability** is a necessity and includes data-inspired simulations

Image credits: nstat.org, hub.jhu.edu, vox.com, Andras Libal
Stability is robustness for all parts of DSLC

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

Stability

BIN YU

It unifies and extends a myriad of works on “perturbation” analysis.

It is a minimum requirement for interpretability, reproducibility, and scientific hypothesis generation or intervention design.
Stability tests DSLC by “shaking” every part. Shakes come from human decisions.
PCS workflow

• Workflow incorporates P, C, S into each step of the DSLC

• In particular, basic PCS inference applies PCS through data and model perturbations at the modeling stage (with P as a first screening step before perturbation intervals are made)
Data perturbations (existing)

• Cross-validation
• Bootstrap
• Subsampling
• Adding small noise to data
• Bootstrapping residuals
• Block-bootstrap
Data perturbations (recent)

- Data modality choices
- Synthetic data (mechanistic PDE models)
- Data under different environments (invariance)
- Differential Privacy (DP) (2020 US census)
- Adversarial attacks to deep learning algorithms
Data perturbations (new)

- Data pre-processing (cleaning) matters

Covered widely in popular media, often as “high debt/GDP ratio is bad for growth”.

It was used to support austerity policies in UK and Europe.
Data perturbations (new)

• Data cleaning versions: stability principle calls for replication

Herdon, Ash and Pollin (2014) was a replication and found that RR had exclusive data selection (cleaning), coding errors, and unconventional weighting. When corrected by Herdon, Ash and Pollin (2014), RR’s conclusion fails to hold.

Image credit:: New Yorker
Model/algorithm perturbations (existing)

- Robust statistics
- Semi-parametric
- Lasso and Ridge
- Modes of a non-convex empirical minimization
- Kernel machines
- Sensitivity analysis in Bayesian modeling
Model/algorithm perturbations (new)

- Researcher to researcher (or team to team) perturbation

Example: 9 climate models

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

- Which problem to work on
- Which data sets to use
- How to clean
- What plots
- What data perturbations
- What algorithm perturbations
- What post-hoc plots/results
- What interpretations
- What conclusions

Image credits: toronto4kids.com
PCS doc. bridges reality and models on [GitHub](https://github.com).

**Reality**

![Neuron image](image-credit: Rebecca Barter)

**Models**

```
# Bootstrap sampling is a widely accepted perturbation understanding of the dependencies. However, sequence behavior that is possible to account for is particularly critical to regulatory processes (Hsing, 2012). In that over 75% of the time examined, new responses are anywhere from 0.01 to 0.05 to account for this potential dependency along the gene. We define the stability of an interaction to be the property of bootstrap samples using the 3 proposed perturbation.
```

**Stability formulation**

```
# Bootstrap sampling is a widely accepted perturbation understanding of the dependencies. However, sequence behavior that is possible to account for is particularly critical to regulatory processes (Hsing, 2012). In that over 75% of the time examined, new responses are anywhere from 0.01 to 0.05 to account for this potential dependency along the gene. We define the stability of an interaction to be the property of bootstrap samples using the 3 proposed perturbation.
```

**Dangerous inference & conclusions**

```
# Bootstrap sampling is a widely accepted perturbation understanding of the dependencies. However, sequence behavior that is possible to account for is particularly critical to regulatory processes (Hsing, 2012). In that over 75% of the time examined, new responses are anywhere from 0.01 to 0.05 to account for this potential dependency along the gene. We define the stability of an interaction to be the property of bootstrap samples using the 3 proposed perturbation.
```

**Unsubstantiated assumptions**

```
# Bootstrap sampling is a widely accepted perturbation understanding of the dependencies. However, sequence behavior that is possible to account for is particularly critical to regulatory processes (Hsing, 2012). In that over 75% of the time examined, new responses are anywhere from 0.01 to 0.05 to account for this potential dependency along the gene. We define the stability of an interaction to be the property of bootstrap samples using the 3 proposed perturbation.
```

**Data**

```
# Bootstrap sampling is a widely accepted perturbation understanding of the dependencies. However, sequence behavior that is possible to account for is particularly critical to regulatory processes (Hsing, 2012). In that over 75% of the time examined, new responses are anywhere from 0.01 to 0.05 to account for this potential dependency along the gene. We define the stability of an interaction to be the property of bootstrap samples using the 3 proposed perturbation.
```

**Image credit: Rebecca Barter**
How to choose **perturbations** in PCS?

• One can never consider all possible **perturbations**

• A pledge to the **stability** principle in PCS would lead to null results if too many **perturbations** were considered

• PCS requires documentation on the appropriateness of all the **perturbations**

• To avoid null results, PCS encourages careful and well-founded choices of the **perturbations** through PCS documentation
Expanding statistical inference under PCS

- Modern goal of statistical inference is to provide one source of evidence to domain experts for decision-making.

- The key is to provide data evidence in a transparent manner so that domain experts can understand as much as possible our evidence generation to evaluate the evidence strength.

Traditionally, p-value has been used as evidence for decisions, but its use has been problematic that psychology journals banned it.
“It is not p-value’s fault”

“The p-value is a very valuable tool, but when possible it should be complemented – not replaced - by confidence Intervals and effect size estimates” – Yoav Benjamini

For one thing, normal approximation can’t back up small p-values like $10^{-8}$, and there are other problems before normal approx. is used.
A critical examination of probabilistic statements in statistical inference

- Viewing data as a realization of a random process is an ASSUMPTION unless randomization is explicit.
- When not, using r.v. actually implicitly assumes “stability”.
- If this assumption is not substantiated, all probabilistic statements are questionable.
- Small p-values often measure model-bias.
- The use of “true” in the “true model” is misleading – we should use other words like approximate or postulated.
Inference beyond probabilistic models

Need trustworthiness measure of an estimated quantity of interest over multiple probabilistic models and/or without probabilistic models
Proposed PCS inference (basic)

1. **Problem formulation:** Translate the domain question to be answered by a model/algorithm (or multiple of them and seek stability). Specify a target of interest.

2. **Prediction screening for reality check:** Filter models/algorithms based on prediction accuracy on held out test data – a sample split approach (it helps assess model bias)

3. **Target value perturbation distribution:** Evaluate the target of interest across “appropriate” data and model perturbations

4. **Perturbation interval reporting:** Summarize the target value perturbation distribution.
Feature importance study: PCS performs well

Simulation results for lasso feature selection in linear model $n=1000$, $p=630$

Adding another method: Lasso (CV)+ asymptotic normal approx.
Climate scientists are practicing PCS inference

- 9 climate models provide a PCS perturbation range of $(1.5, 5.5)$ for global mean-temperature change by 2090

The change in global-mean temperature estimated by nine climate models forced by the SRES A2 emission scenario. (Source: IPCC TAR, Chapter 9)
Making Random Forests interpretable
by adding (more) stability
Iterative random forests to discover predictive and stable high-order interactions

Sumanta Basu\textsuperscript{a,b,c,1}, Karl Kumbier\textsuperscript{d,1}, James B. Brown\textsuperscript{c,d,e,f,2}, and Bin Yu\textsuperscript{c,d,g,2}

Co-authors

S. Basu \hspace{1cm} K. Kumbier \hspace{1cm} B. Brown

Culmination of 3+ years of work
Pattern Recognition vs. Pattern Discovery

Pattern Recognition:
Finding something for which you already know to look

Pattern Discovery:
Identifying structure that hasn’t been seen before
Iterative random forests (iRF) for pattern discovery in combinatorially vast systems

Classical statistical approaches are not sufficient:
Consider measurables: $x_1, \ldots, x_p$
We would like to identify relationships such as:
$y = g(x_j) + \text{noise}$, where $g$ depends only on a small subset of the $x$'s and is not too complex.
SOP is to leverage forward procedures:
$$y \approx \sum_{j \neq i} \alpha_j x_j + \sum_{k,l \neq i} \beta_{k,l} x_k x_l + \cdots$$

Polynomial interactions do not work well in genomics
Embryonic development in *Drosophila melanogaster*
Order-4 interaction regulate eve stripe 2

Goto et al. (1989), Harding et al. (1989), Small et al. (1992), Isley et al. (2013), Levine et al. (2013)
Regulatory interactions through predictability and stability or PCS

Natural phenomenon

Prediction

Interpretation

Transcription is initiated when activating transcription factors reach sufficient DNA occupancy
Capturing the form of genomic interactions

- Interactions are high-order and combinatorial in nature
- Interactions can vary across space and time as biomolecules carry out different roles in varied contexts
- **Interactions** exhibit **thresholding behavior**, requiring sufficient levels of constitutive elements before activating

(Wolpert, 1969; Jaeger and Reinitz, 2006)

(Hartenstein, 1993)

(Spitz and Furlong, 2006)
From genomic to statistical interactions

Transcription is initiated when a collection of activating TFs achieve sufficient DNA occupancy

\[
R(x) = \prod_{i \in S} 1\{x_i > t_i\}
\]

Order-\(S\) interaction,

\[
S \subseteq \{1, \ldots, p\}, |S| = s
\]
Random Forests (RFs)
Breiman (2001)

Draw $T$ bootstrap samples and fit a modified CART to each sample.

1. Grow CART trees to purity
2. When selecting splitting feature, choose a subset of $\text{mtry}$ features uniformly at random and optimize CART criterion over subsampled features.
iterative Random Forests (iRFs)
Basu, Kumbier, Brown and Yu (2018)

Core ideas

1. Soft dim reduction using importance index
2. Random interaction trees to find intersections of paths
3. Outer-loop bagging assesses stability

Similar computational and memory costs as RF
Iteratively re-weighted RF stabilize decision paths

Iteratively re-weighted Random Forests

Iter 1
- Input: $\mathcal{D}, w^{(1)} \leftarrow (1/p, \ldots, 1/p)$
- Output: $w^{(2)} \leftarrow \text{Gini Importance}$

Iter 2
- Input: $\mathcal{D}, w^{(2)}$
- Output: $w^{(3)} \leftarrow \text{Gini Importance}$

Iterate through $K - 1$

Feature weights

1 2 3
4 5

importance index

1 2 3
4 5

Re-weighting
Amaratunga et al. (2014)

MDI-oob index: Wed 10:45 AM -- 12:45 PM at East Exhibition Hall B + C #5.
Digression: Interactions in market baskets

Feature-index sets

$Z_1$  
$Z_2$  
$Z_3$  
$Z_4$  

$I_1$  
$I_2$  
$I_3$  
$I_4$
Random Intersection Trees (RIT)
Shah and Meinshausen (2014): fast computation uses sparsity

Randomly sampled class-\( C \) observation

“Survived” interaction
Random Intersection Trees (RIT)
Shah and Meinshausen (2014)

Randomly sampled class-\(C\) observation

“Survived” interaction
Random Intersection Trees (RIT)
Shah and Meinshausen (2014)

Randomly sampled class-\(C\) observation

“Survived” interaction
Generalized RIT for Decision Trees
fast computation uses sparsity

\[ \mathcal{I}_{it} \subseteq \{1, \ldots, p\} \]  \textit{Feature-index set} for leaf node containing observation \( i = 1, \ldots, n \) in tree \( t = 1, \ldots, T \)

\[ Z_{it} \in \{0, 1\} \]  \textit{Prediction} for the leaf node containing observation \( i = 1, \ldots, n \) in tree \( t = 1, \ldots, T \)

\[ S \leftarrow \text{RIT}(\{\mathcal{I}_{it}, Z_{it}\}, C') \]
Stability bagging

Output feature interaction sets with stability scores:

\[ \{S, \text{sta}(S)\} \]

\[ S \subseteq \{1, \ldots, p\} \]

\[ \text{sta}(S) = \frac{1}{B} \sum_{b=1}^{B} 1(S \in S_b) \]

Reference: (Breiman, 1996)
Example: Enhancer activity in *Drosophila*

*Drosophila* blastoderm embryos:
- n=7809 genomic sequences
- p=80 ChIP assays (TF binding, histone modifications)
- Response: enhancer activity

(Bermen et al., 2002; Frise et al. 2010; Fisher et al., 2012; Kvon et al. 2014)
iRF keeps predictive accuracy, and finds stable interactions.
80% of pairwise interactions are validated

| interaction (S) | sta(S) | references |
|-----------------|--------|------------|
| Gt, Zld         | 1      | Harrison et al. (2011); Nien et al. (2011) |
| Twi, Zld        | 1      | Harrison et al. (2011); Nien et al. (2011) |
| Gt, Hb          | 1      | Kraut and Levine (1991a,b); Eldon and Pirrotta (1991) |
| Gt, Kr          | 1      | Kraut and Levine (1991b); Struhl et al. (1992); Capovilla et al. (1992); Schulz and Tautz (1994) |
| Gt, Twi         | 1      | Li et al. (2008) |
| Kr, Twi         | 1      | Li et al. (2008) |
| Kr, Zld         | 0.97   | Harrison et al. (2011); Nien et al. (2011) |
| Gt, Med         | 0.97   | – |
| Bcd, Gt         | 0.93   | Kraut and Levine (1991b); Eldon and Pirrotta (1991) |
| Bcd, Twi        | 0.93   | Li et al. (2008) |
| Hb, Twi         | 0.93   | Zeitlinger et al. (2007) |
| Med, Twi        | 0.93   | Nguyen and Xu (1998) |
| Kr, Med         | 0.9    | – |
| D, Gt           | 0.87   | – |
| Med, Zld        | 0.83   | Harrison et al. (2011) |
| Hb, Zld         | 0.80   | Harrison et al. (2011); Nien et al. (2011) |
| Hb, Kr          | 0.80   | Nüsslein-Volhard and Wieschaus (1980); Jäckle et al. (1986); Hoch et al. (1991) |
| D, Twi          | 0.73   | – |
| Bcd, Kr         | 0.67   | Hoch et al. (1991, 1990) |
| Bcd, Zld        | 0.63   | Harrison et al. (2011); Nien et al. (2011) |
Stable interactions reflect Boolean-type rules

3\text{rd} or 4\text{th} or higher order interactions are suggestions for Crispr experiments
2018: Chan Zuckerberg Biohub Intercampus Award
iRF is a cornerstone

One of the 6 awards

Project leaders:
Rima Arnout and Atul Butte (UCSF)
James Priest and Euan Ashley (Stanford)
Ben Brown and Bin Yu (UC Berkeley)

Collaborators:
Chris Re (Stanford), Deepak Srivastava (UCSF)

Image credits: Rima Arnout.
Interpreting iRF results generates biological hypotheses
Other examples of interpretation need

- FDA wants interpretation of DL algorithms for radiology
- Stimuli to characterize a neuron
- Phrases making a sentence negative
(Faithful) interpretation builds trust

EU's General Data Protection Regulation (GDPR) (2016) gives a “right” to explanation, and demands ML/Stats algorithms to be **human interpretable**

Image credit: [https://christophm.github.io/interpretable-ml-book/](https://christophm.github.io/interpretable-ml-book/)
Some related work

- Lipton (2017)
- Doshi-Velez and Kim (2017)
- Molnar (2019) book
“We define interpretable machine learning as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model. Here, we view knowledge as being relevant if it provides insight for a particular audience into a chosen problem. These insights are often used to guide communication, actions, and discovery.”
iML through the PDR desiderata

- **P**- Predictive accuracy for reality check average (global) and point-wise (local)

- **D**- Descriptive accuracy: the degree to which an interpretation method objectively captures the relationships learned by machine learning models (both post-hoc and model-based methods can increase D)

- **R**- Relevancy: interpretation method is “relevant” if it provides insight for a particular audience into a chosen domain problem

Relevancy often plays a key role in determining the tradeoff between predictive and descriptive accuracy
iML-PDR in one figure

(P) Predictive accuracy

(D) Descriptive accuracy

(R) relevancy

R is key in the trade-off of P and D
Model-based interpretability

• Sparsity (e.g. small sparse logistic regression for lung cancer prediction)

• Simulatability (e.g. small decision tree for lung cancer prediction)

• Modularity (e.g. generalized additive models, layers in DL)

• Domain-based feature engineering (e.g. credit score)

• Model-based feature engineering (e.g. clustering and dimensionality reduction like PCA)
Post-hoc interpretability

• Data set level (global) interpretation (feature and interaction importance, statistical significance score, visualization)

• Prediction-level (local) interpretation (feature importance and alternatives)

Murdoch et al (2019) contains many examples from our own work and others’ work to illustrate PDR.
Agglomerative Contextual Decomposition (ACD)

(1) How can we get feature-interaction importance for a DNN model prediction in general? (ICLR 2018)

(2) How can we visualize these feature-interactions in an understandable way? (ICLR, 2019)

(3) How can we use the importance scores and prior info to debias algorithms? (submitted, 2019)
Previous work (post-hoc interpretation)

- gradient-based methods
  - LIME  
    Ribeiro et al. (2016)
  - Integrated Gradients (IG)  
    Sundarajan et al. (2017)

- contribution-based
  - Occlusion / saliency maps  
    Dabkowski & Gal (2017)
  - SHAP  
    Lundberg & Lee (2017)
CD: Contextual Decomposition
(Murdoch, Liu and Y. (2018). ICLR)

• Given a LSTM with weights, CD gives a prediction-level score for each part of the input to “explain” the prediction

\[ \text{LSTM}(w_1, \ldots, w_T) = \text{SoftMax}(\gamma_T + \alpha_T) \]

• \( \gamma_T \) corresponds to contributions solely from the phrase, \( \alpha_T \) other factors
Agglomerative Contextual Decomposition (ACD)

CD is generalized to DNNs.
ACD is a hierarchical clustering algorithm with visualization, where the joining metric is CD score

*Singh, *Murdoch, Y. (2019). ICLR

CD/ACD code: github.com/csinvac/acd
a great ensemble cast can’t lift this heartfelt enterprise out of the familiar.
prediction: puck

skates are important

puck is important
Human experiments

Telling a good model from a “bad” one using only interpretations

Whether Interpretation instills trust or not
Improving models by regularizing ACD explanations

Rieger, Singh, Murdoch, Y. (2019). In submission

github.com/laura-rieger/deep-explanation-penalization
Using CD to identify fundamental cosmological parameters of the universe

In Progress

W. Ha, C. Singh, F. Sapienza
F. Lanussen, V. Boehm

Yu group

@ Berkeley Center for Cosmological Physics
Cosmological parameters such as $\Omega_M$, determine evolution of universe

$\Omega_M \rightarrow \text{Map of mass in the universe}$

Adaptation of NASA WMAP Science Team Image
CNN predicts well, but what does it learn?

Need to go beyond just identifying important pixels...
CD can measure the importance of different frequencies in the image to the model’s prediction.
Goals of (faithful) interpretation

• Save on data collection
• understand which features drive the predictions
• give trust to using deep learning
• distill the DL model into a simple model (e.g. *generative and mechanistic*)

Success of these goals serves as validation

“Data science process: one culture”
Summary

Veridical data science (trustworthy AI) through

- **PCS** framework (workflow and documentation on github) advocating best practices for a responsible, reliable, reproducible and transparent DSLC to reach trustworthy data conclusions

- **PCS** inference incorporating data and model (researcher) perturbations

- **PDR** interpretation framework guides selection and evaluation of interpretation methods

- Case studies: iRF (siRF), ACD (*DeepTune omitted)

- Domain knowledge is important and **PCS** generates testable hypotheses towards causality

Hope PCS and PDR are useful for your projects
PCS next steps

- **PCS-compliant** projects
- Unpacking PCS for emergency medicine and social science
- Theory on PCS and fast algorithms to implement perturbations
- PCS computing platform
- PCS-guided DS book in prep
PDR next steps

- Cosmology projects (CNN-ACD, and iRF)
- Cancer drug discovery project (PCS-compliant)
- Epistasis discovery project (PCS-compliant)
- Simons Inst workshop at Berkeley on June 29 – July 2, 2020 “Interpretable Machine Learning in Natural and Social Sciences”

(co-organizers: Hima Lakkaraju, Zack Lipton, David Madigan, and BY. , part of Simons summer cluster with Shai Ben-David and Ruth Urner)
People make “veridical” happen

Problem to solve
Question to answer
Domain knowledge

Critical thinking
Algorithms
Inference

People

Humanly understandable conclusions

Relevant theory

Thanks to my group
Opportunities and challenges

Within DS/ML/AI community, we need

• transdisciplinary, trans-methodological people with communication skills

• position and vision papers

• attention to energy consumption impact on climate change
Opportunities and challenges

Outfacing for DS/ML/AI community, we need

- A few COMMON, robust and reliable “products”

- Certification and labels for open-source and SAFE software

- **Rigorous evaluation process of new algorithms** (modularity is a virtue) (e.g. taking things apart like in red-tagging in software development)
For veridical data science, academic/industry/government leadership and funding agencies need to incentivize

• Quality research and **trustworthy publication**, not paper counting

• “Team-brain” to solve complex transdisciplinary problems

• Fair collaborative environment so that the best arguments win
Our papers

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

   (old title: Three principles of data science: predictability, computability and stability (PCS))

2. Definitions, methods and applications in interpretable machine learning
   J. Murdoch, C. Singh, K. Kumber, R. Abbasi-Asl, and B. Yu (2019), PNAS
Upcoming book on data science by MIT Press
Coming at the end of 2021 with a free on-line interactive version in the spring

Veridical Data Science: A Book
Bin Yu¹² and Rebecca Barter¹
¹Department of Statistics, UC Berkeley
²Department of Electrical Engineering and Computer Science, UC Berkeley

What skills does the book teach?
Veridical Data Science (VDS) will teach the critical thinking, analytic, human-interaction and communication skills required to effectively formulate problems and find reliable and trustworthy solutions. VDS explains concepts using visuals and plain English, rather than math and code.
The primary skills taught are:

Critical thinking
Readers will learn to:
Formulate answerable questions using the data available
Scrutinize all analytic decisions and results
Document all analytic decisions
Appropriate common techniques to unfamiliar situations
Deal with real, messy data

Technical skills
Data cleaning
- Algorithmic
  - Dimension reduction
  - Clustering
  - Least Squares & ML
  - Regularization
- Stability-based inference
  - Inference
  - Causal Inference
  - Perturbation Intervals
  - Trustworthiness Statements

Data merging
- Exploratory Data Analysis
- Data merging

Communication
- Exploratory Visual Summaries
  - Preparing explanatory visual and numeric summaries for explaining data and findings to an external audience
- Written reports
  - Preparing written analytic reports for case studies based on real, messy data

Core guiding principles for the book

The DS Lifecycle

The Data Science Lifecycle is an iterative process that takes the analyst from problem formulation, data cleaning, exploration, algorithmic analysis, and finally to obtaining a verifiable solution that can be used for future decision-making.

Blending together concepts from statistics, computer science and domain knowledge, the data science life cycle is an iterative process that involves human analysts learning from data and refining their project-specific questions and analytic approach as they learn.

Three realms

Readers will learn to view every data problem through the lens of connecting the three realms:
1. the question being asked and the data collected (and the reality the data represents)
2. the algorithms used to represent the data
3. future data on which these algorithms will be used to guide decision-making.
Guiding the reader to connect the three realms is a means of guiding the reader through the data science lifecycle.

Intended Reader/Audience

Anyone who wants to learn the intuition and critical thinking skills to become a data scientist or work with data scientists.

Neither a mathematical nor a coding background is required.

VDS could form the basis of a semester- or multi-semester-long introductory data science university course, either as an upper-division undergraduate or early graduate-level course.

Interested? Get in touch!

Bin Yu
Email: binyu@stat.berkeley.edu
Website: https://www.stat.berkeley.edu/~binyu/Site/Welcome.html

Rebecca Barter
Email: rebeccabarter@berkeley.edu
Website: www.rebeccabarter.com
Twitter: @rbarter
Thank you!

Visit Bin Yu’s website for more info
https://binyu.stat.berkeley.edu/