Veridical Data Science

Bin Yu
Statistics and EECS, UC Berkeley

Math/Stats Joint Colloquium, UC Davis
Feb. 13, 2020
**veridical**

/veˈridɪk(ə)l/

*adjective*  
FORMAL

truthful.

- coinciding with reality.

"such memories are not necessarily veridical"
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: ____________
2019

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/

website of S. Saria at JHU
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

Missing: quality control and standardization
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

• DeepTune to characterize neurons

• 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

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

Image credits: R. Barter and toronto4kids.com
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)

Image credits: R. Barter and toronto4kids.com
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

Image credits: groundai.com
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

N. Saito      J. L. Wang   H. Muller

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

Reality

quantitative and qualitative narratives

Models

Stability formulation

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 Deep Learning interpretable
by adding stability over 18 models
The DeepTune framework for modeling and characterizing neurons in visual cortex area V4

Abbasi-Asl, Chen, Bloniarz, Oliver, Willmore, Gallant, and Y. (submitted, 2018)
https://www.biorxiv.org/content/early/2018/11/09/465534

Culmination of 3+ years of work

Reza Abbasi-Asl
Yuansi Chen
Adam Bloniarz

In collaboration with

Mike Oliver
Ben Willmore
Jack Gallant
Interface between Neuroscience and Deep Learning

- Human visual cortex V4 is a **difficult** and **elusive** area.

- Deep convolutional neural networks

![Brain Diagram](http://cs231n.github.io/assets/nn1/neural_net2.jpeg)

![Neural Network Diagram](http://cs231n.github.io/assets/nn1/neural_net2.jpeg)
V1 decoded by Hubel and Wiesel (1959)

V1: orientation and location selectivity, and excitatory and inhibitory regions.

Nobel Prize in 1981
**V4** has been probed by **synthetic polar and hyperbolic gratings and complex shape stimulus**

Gallant et al. 1993, 1996

David et al (2006)

- Cartesian
- Polar
- Hyperbolic
V4 has been probed by synthetic convex and concave boundary stimuli.

Pasupathy and Connor 1999, 2002

The stimuli were created by systematically combining convex and concave boundary elements.
Our data collection: 71 V4 neurons
(from the Gallant Lab at UC Berkeley)

Well-isolated visual neurons

Neuronal behavior is probed using sequences of natural images
Related works


Parallel developments in the DiCarlo Lab at MIT:

Here we replicate their predictive results and aim at interpretation and understanding.
Questions to answer

1. How do we characterize V4 neurons?

   If we can characterize a neuron, we then know how to generate data-driven hypotheses.

2. How much do Convolutional Neural Networks (CNNs) resemble brain function?
DeepTune in a nutshell

Transfer predictive learning based on CNN+reg to derive 18 state-of-art predictive models for our V4 neurons (prediction)

Stable interpretation via DeepTune images over 18 models suggest what V4 neurons do (stability)

As a result, we provide some support for resemblance of CNNs to primate brain, and generate image stimuli for closed-loop experiments
Transfer learning...

Step 1
Training CNN

Images from ImageNet dataset

CNN used in Training for classification task

Labels of Images

Step 2
Feature Extraction And Fitting

Limited Images used in experiment

Early layers of trained CNN

Linear regression fitting via Ridge or Lasso

V4 neural activity

Prediction performance across different layers of CNN (AlexNet): N2 works well for V4
DeepTune image generation: Neuron 1

DeepTune Image(s):
Maximizing a (regularized) fitted model
Stable curve patterns across structurally compressed models

DeeTune image from full network

DeepTune images from compressed networks

Abbasi-Als and Y. (2017)
Top *curve* images from training set based on a model for neuron 1
Top curve images from test data set without models for Neuron 1
Stable predicted neuron activity from three deep nets +Lasso for a particular neuron
Dealing with multiple predictive models

CNN + regression: 18 models

**Interpretation** via stability over multiple provides testable (prescriptive) characterizations of V4 neurons

We combat “**model-hacking**” via “**stability principle**”
Neuron 1 seems a curve neuron and DeepTune images provide intervention stimuli.

18 DeepTune images from 18 predictive models

Ridge

<table>
<thead>
<tr>
<th>Layer 2</th>
<th>Layer 3</th>
<th>Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
</tr>
<tr>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
</tr>
</tbody>
</table>

LASSO

<table>
<thead>
<tr>
<th>Layer 2</th>
<th>Layer 3</th>
<th>Layer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
</tr>
<tr>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
</tr>
</tbody>
</table>
Consensus DeepTune

- **Single model DeepTune:** Use gradient ascent to find stimuli that maximize one of the CNN+Regression model output

- **Consensus DeepTune:** The models have to agree with each other to create a DeepTune pattern. (Stability)

\[
|\nabla f(x)| = \text{element-wise min}_{i=1 \ldots \#\text{models}} |\nabla f_i(x)|
\]
Consensus DeepTune from 10 initializations

Neuron 1
Hierarchical clustering of "good" neurons through DeepTune Images on CNN feature space
Neuron 1: regularity of spacing between curves
an artifact of convolution filter size

18 DeepTune images from 18 predictive models

Ridge
Layer 2 Layer 3 Layer 4
AlexNet
VGG
GoogleNet

LASSO
Layer 2 Layer 3 Layer 4
Layer 2 Layer 3 Layer 4
DeepTune images or parts are “verifiable” in closed-loop experiments

- cropped DeepTune images as stimulus images
- randomly cropped and combined images
- cropped images with varied sizes

Already done in
Interpreting DeepTune results generates neuroscience hypotheses
Other examples of interpretation need

- FDA wants interpretation of DL algorithms for radiology
- Iterative random forests (iRF) for non-linear interactions
- 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/
Some related work

- Lipton (2017)
- Doshi-Velez and Kim (2017)
“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)

- Simulatibility (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  

• 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

The movie was not good

CD

Negative
Agglomerative Contextual Decomposition (ACD)

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


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

In submission

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

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

Yu group

In Progress
Cosmological parameters such as $\Omega_M$, determine evolution of universe
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
People make “veridical” happen

Problem to solve
Question to answer
Domain knowledge

Critical thinking
Algorithms
Inference

Humanly understandable conclusions

People

Relevant theory

Thanks to my group
Opportunities and challenges

Within Stats/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 Stats/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
   (old title: Three principles of data science: predictability, computability and stability (PCS))
   (Yu and K. Kumbier, 2019)
   https://arxiv.org/abs/1901.08152

2. Definitions, methods and applications in interpretable machine learning
Upcoming book on data science

Coming (2021?)

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

What skills do we teach?

Data Science In Action (DSIA) will teach the critical thinking, analytic, and communication skills required to effectively formulate problems and find reliable and trustworthy solutions.

DSIA teaches the reader skills that are adaptable to any data-based problem. The primary skills taught are:

- Reliability and trustworthiness (Principles)
- Communication skills required to effectively formulate problems and find reliable and trustworthy solutions.

Core guiding principles

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
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 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.

DSIA 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.

The PCS framework provides concrete techniques for finding evidence for the connections between the three realms.

Predictability: if the patterns found in the original data also appear in withheld or new data, they are said to be predictable. If an analysis or algorithm finds predictable patterns, then these patterns are likely to be capturing real phenomena.

Computability: algorithmic and data efficiency and scalability is essential to ensuring that the results and solutions (e.g. a predictive algorithm) can be applied to new data.

Stability: minimum requirement for reproducibility. If results change in the presence of minor modifications of the data (e.g. via perturbations) or human analytic decisions, then there might not be a strong connection between the analysis/ algorithms and the reality that underlies the data.
Berkeley’s DS Intellectual and Organizational Vision

Summary of the 2016 Report by the Faculty Advisory Board of the Data Science Planning Initiative
Prepared: 19 August 2016
Cathryn Carson, FAB Chair

Contents
A. Rationale for action: Why Berkeley, why now
B. Recommendations
   1. Organizational form: Core and connections
   2. Faculty FTE: Campus-wide surge and strategic foci
   3. Fundraising pillar and revenue generation
C. Situational challenges and next steps
D. The Faculty Advisory Board

CS/Stat Faculty
co-creating and co-teaching
data8.org and ds100.org

DS Interim Dean: D. Culler

New DS Major, Fall 2018

Associate Provost J. Chayes
Div of Computing, DS and Society

Data8 Spring19 – 1500 students

Data100 Spring19: 1,100 students