Trustworthy AI through **veridical data science** and **interpretable machine learning**

Bin Yu

Statistics and EECS, UC Berkeley

Virtual Meeting on Statistics and AI

Statistics Committee of the Hungarian Academy of Science

Sept. 16, 2021

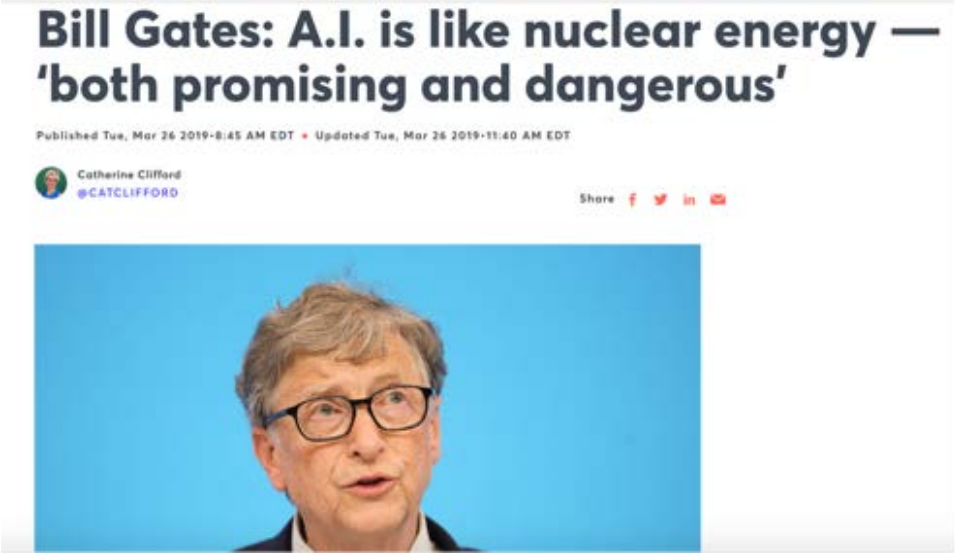
AI is part of modern life

Virtual assistants
(Siri, Alexa, Cortana)

Wearable health devices
(FitBit, Apple watch)

Recommendation systems
(YouTube, Facebook)

Online news



Election campaigns

Self-driving cars

Online gaming

Precision medicine

Biology

Chemistry

Neuroscience

Materials Science

Law

Sociology

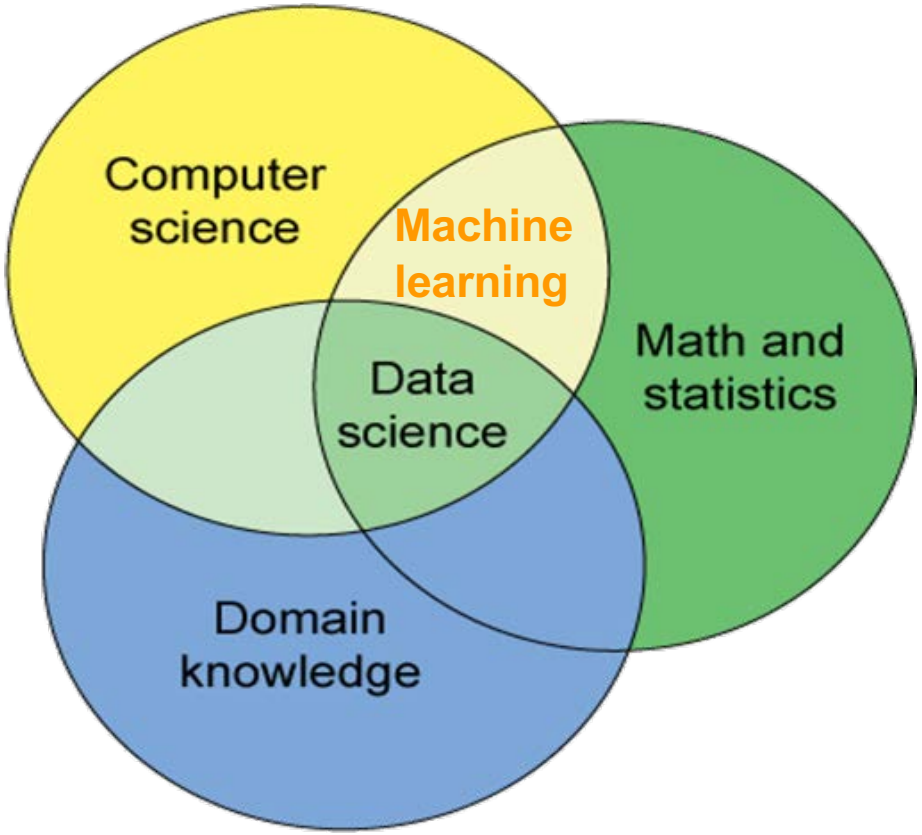
Cosmology

Economics

Political Science

… and beyond

Data science (DS) is a key element of AI



Conway's Venn Diagram

Goal:

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

Trustworthy AI:
two complementary approaches

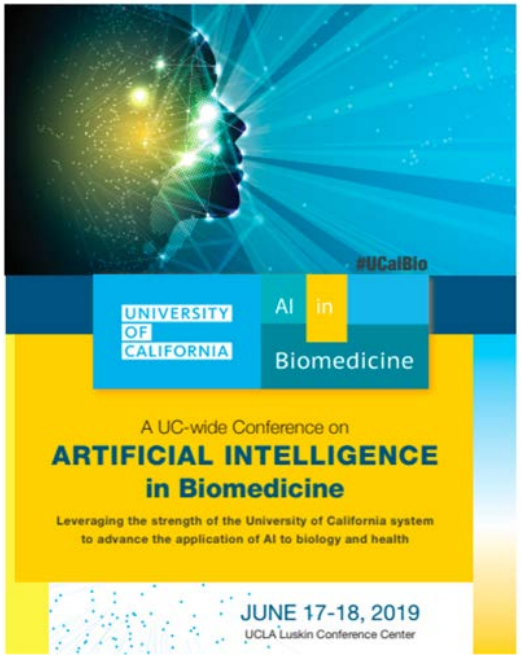
- "Best practices" to **maximize** the **promise** (precaution)
- Damage control to **reduce** the **danger** (intervention)

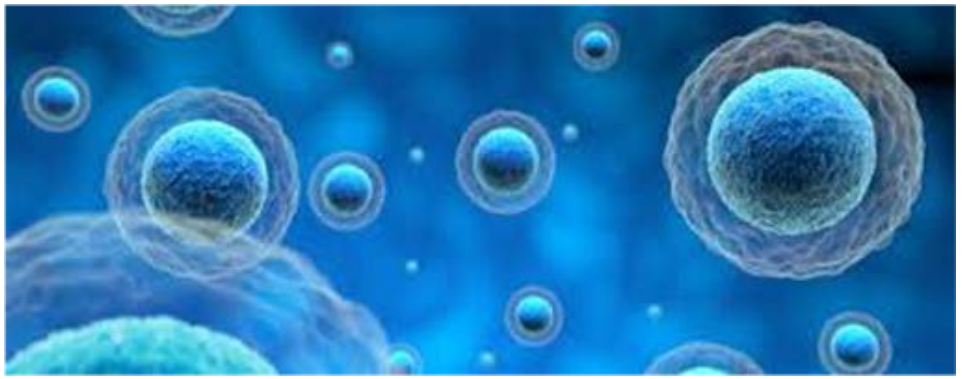
Veridical Data Science (Y. and Kumbier, PNAS, 2020)

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

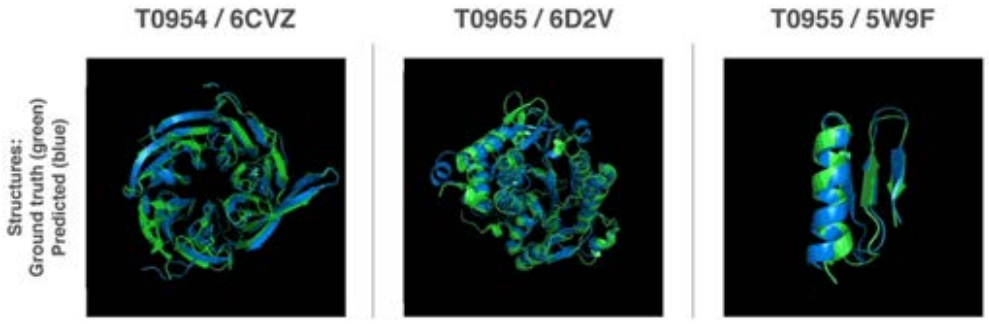
Necessary to realize the promises and mitigates the dangers of AI

Frontiers in biological and medical DS





medium.com



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

Website of S. Saria at JHU

Scientific ML (sML): part of trustworthy AI

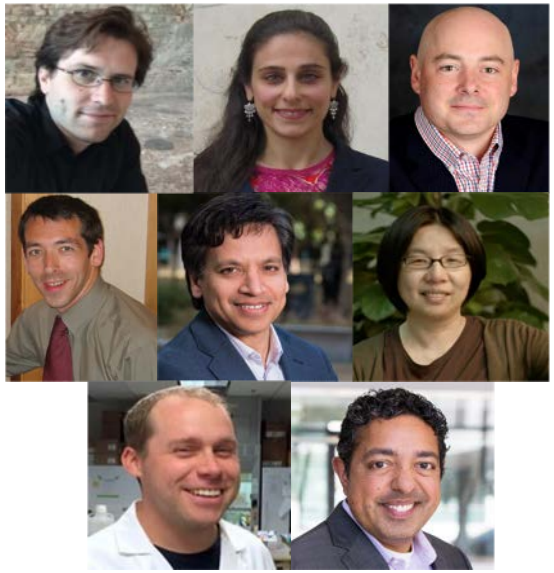
- It uses machine learning/statistics for scientific research to extract, from data, discoveries, theory, and knowledge
- It builds scientific principles/theory in machine learning algorithms
- It iterates between the above two steps
- Results are subject to scientific standards

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

Multi-scale deep learning and single-cell models of cardiovascular health

PIs: Euan Ashley, Rima Arnaout, Ben Brown, Atul Butte, James Priest, Bin Yu

Collaborators: Chris Re, Deepak Srivastava



M. Behr

K. Kumbier

M. Aguirre

A. Cordova-Palomera

Q. Wang

N. Youlton

C. Weldy

W. Hughes

A. Agarwal

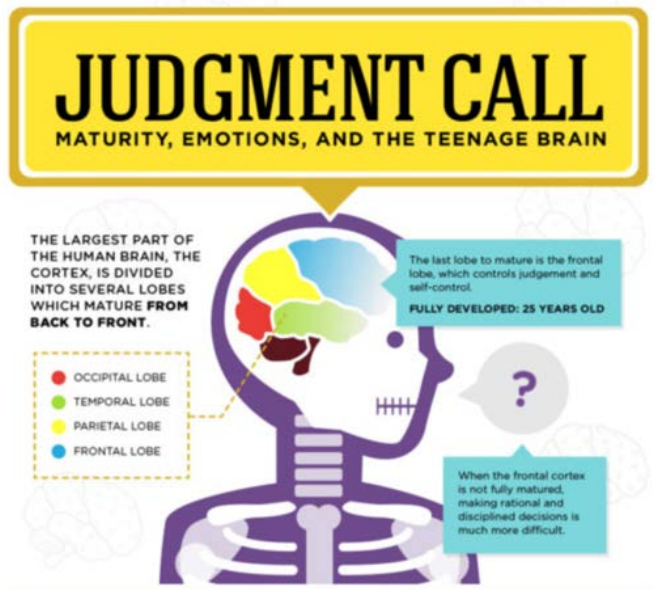
T. Tang

O. Ronen

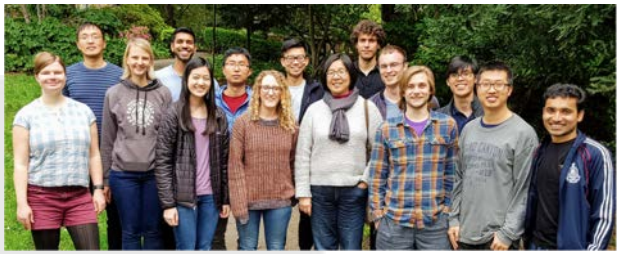
X. Li

Biohub project as a data science problem

- Medical **question**: which genes interact to induce HCM (hypertrophic cardiomyopathy)?
- Which **data** to use? How to **clean**?
- **EDA**: summaries, plots, …
- **Modeling**: Which algorithms to use to find nonlinear interactions?
- **Interpretation & evaluation** of results



Data Science Life Cycle (DSLC): A holistic view



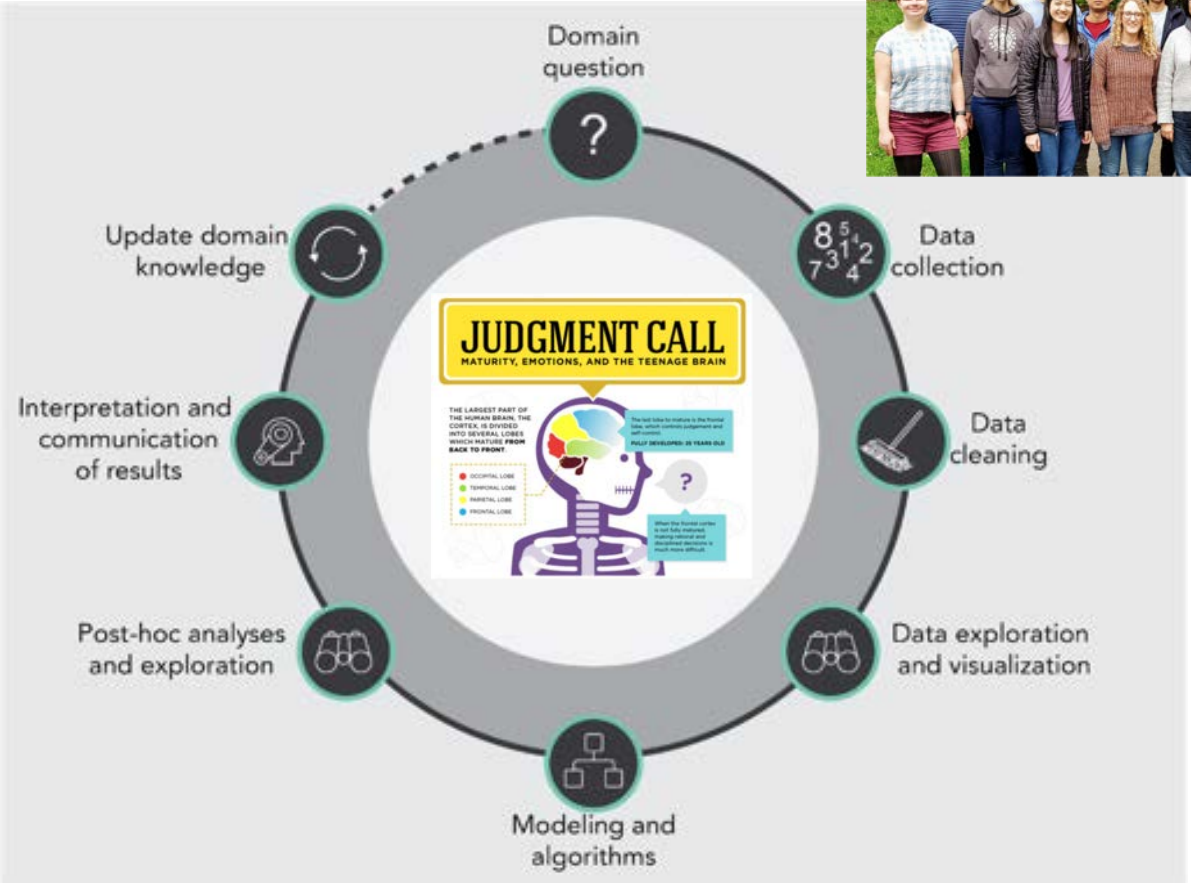


Image credits: R. Barter and toronto4kids.com

To maximize promise,

Scientific conclusions must **reflect reality** and be **stable** to human judgment calls throughout the integrated data science life cycle (DSLC).

Thus, data science requires **quality control** and **standardization** inspired by empirical practice.

2001

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

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 $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:

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)

Machine learning **Statistics**

Cross-validation

Deep Learning, AlphaGo, AlphaFold, self-driving cars, …

Linear model, Logistic regression, PCA, p-value, t-test, …

Predictability, Computability, Stability (**PCS**) framework

one culture

for veridical data science

Rest of the talk

PCS conceptual framework:

veridical data science with reliable and reproducible results

Interpretable Machine Learning

What is interpretable ML or iML?

iterative random forests (iRF) (based on PCS)

Adaptive wavelet distillation (AWD) of deep learning networks

PCS-based software:

Veridical Flow and simChef

2001

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

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 $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:

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

$$\text{response variables} = f(\text{predictor variables, random noise, parameters})$$

Machine learning

Statistics

Deep Learning, AlphaGo, AlphaFold, self-driving cars, …

Linear model, Logistic regression, PCA, p-value, t-test, …

Image credit: https://www.lib.berkeley.edu/news_events/bridge/sfobay.html

PCS framework for veridical data science: one culture

Y. and Kumbier (PNAS, 2020)

Three principles of data science:

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

(**C**)omputability [ML]

(**S**)tability [Stats]

Unifies, streamlines, and expands ideas and best practices in **both** ML and Stats

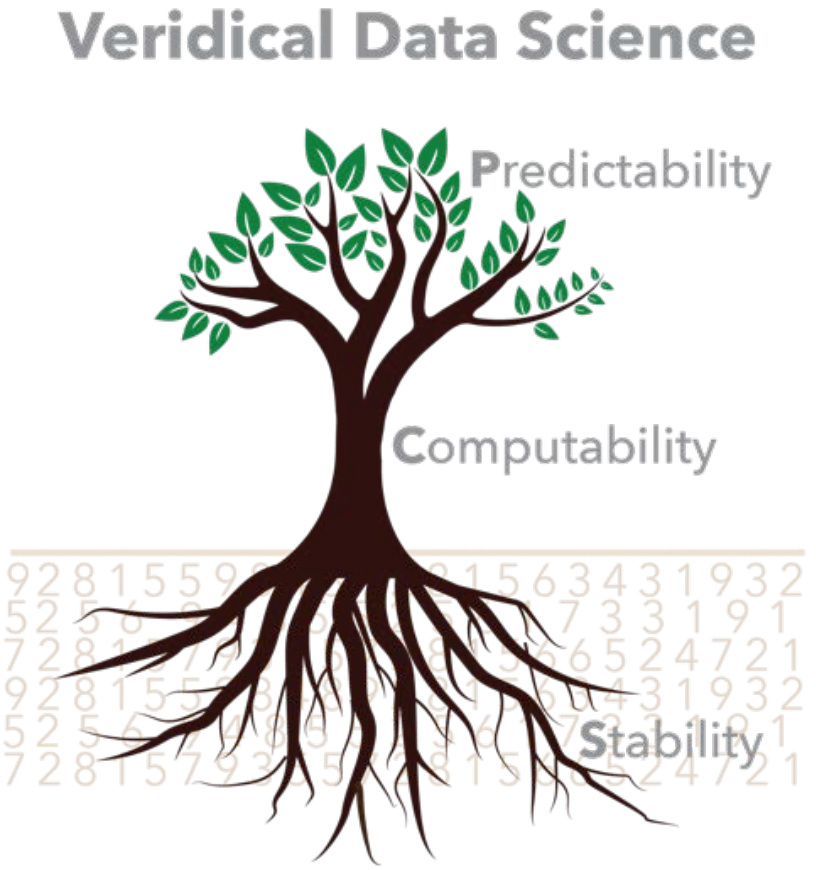


Image credit: R. Barter

The stability principle

Reproducibility *is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in* ***stability*** *of statistical results relative to* ***reasonable perturbations*** *to data and to the model used.*

- Yu (2013) [Stability]

PCS connects science with engineering

- **Predictability** and **stability** embed the scientific principles of *prediction* and *replication*

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

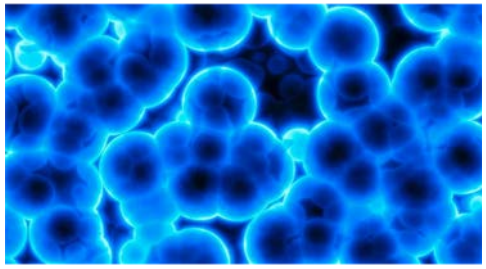


Image credits: nstat.org, hub.jhu.edu, vox.com, Andras Libal

PCS in a nutshell

Predictability for reality check

Stability analysis over human judgment calls

Computability is implicit in P and S

Data Science Life Cycle

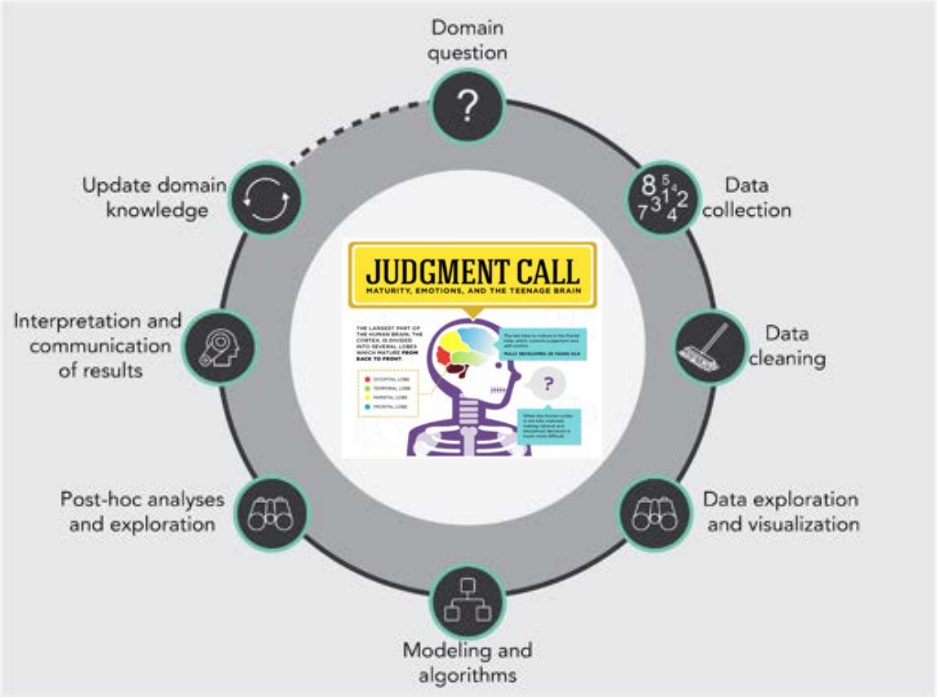


Image credits: R. Barter and toronto4kids.com

PCS in a nutshell

Predictability for reality check

Stability analysis over human judgment calls

Computability is implicit in P and S

Stability analysis "shakes" every part of DSLC to check robustness

"Stability analysis" defined **broadly** and **flexibly**

Data Science Life Cycle

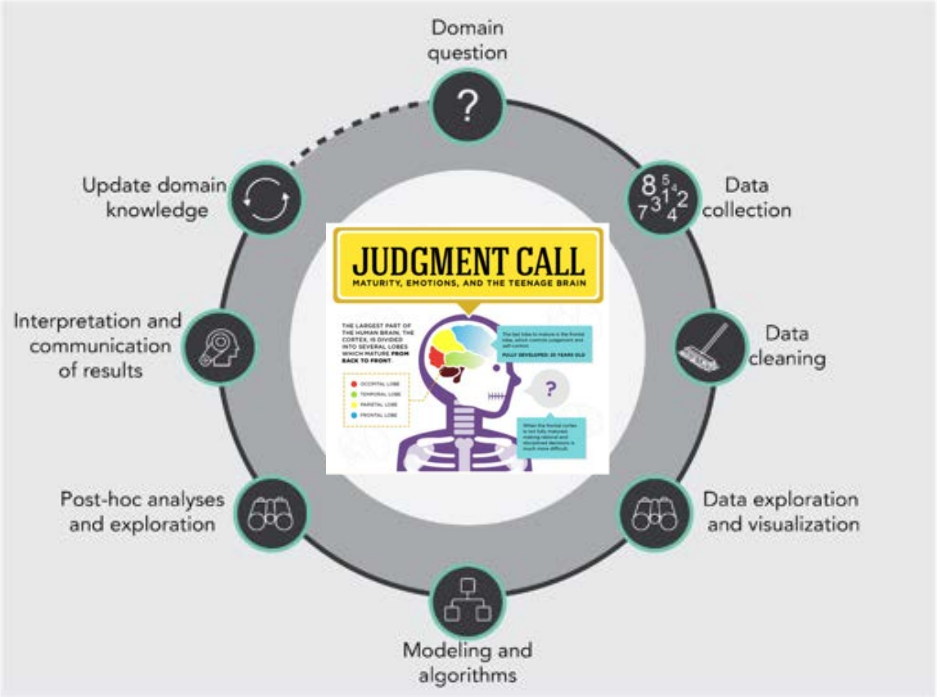
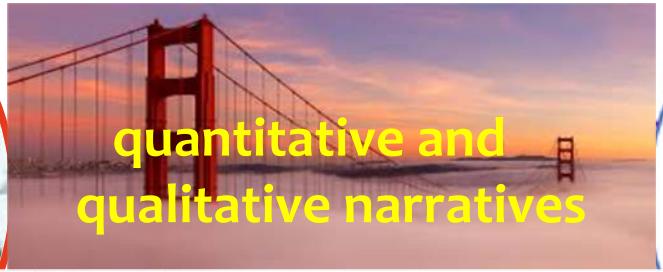
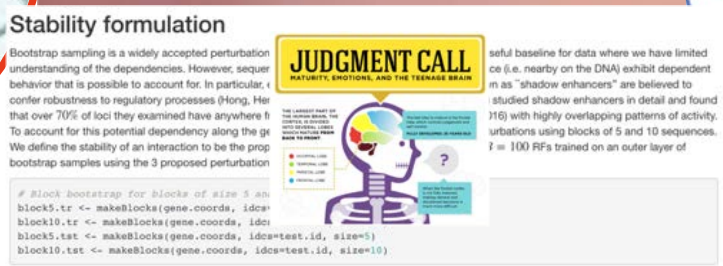


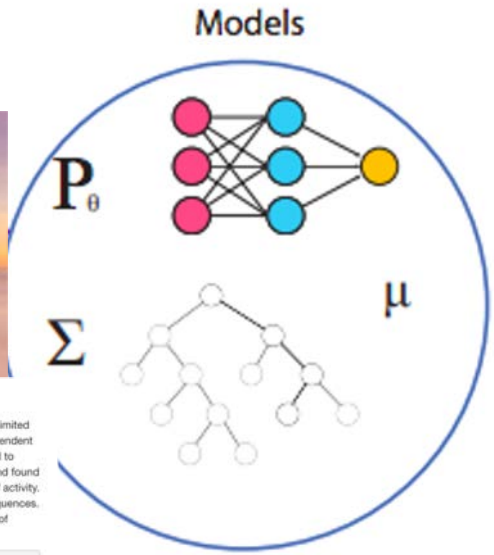
Image credits: R. Barter and toronto4kids.com

PCS documentation [on GitHub (JupyterNotebook)] bridges reality and models

Reality







Mental Construct

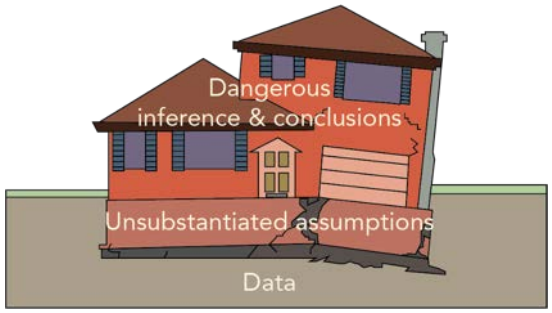


Image credits: Rebecca Barter

What is PCS to a doctor?

"The PCS framework builds a **working relationship** between data and the clinical world."

"PCS is a 'look under the hood' to ensure that the **conclusions found** are what **the data genuinely suggest**. In all, PCS is a holistic approach to helping the clinician **understand, interpret, and build the science** we need to help our patients."

Dr. Aaron Kornblith, ER, UCSF

main medical collaborator on PCS stress-testing of PECARN CDR

Data preprocessing (shoe-leather work) causes data perturbations

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

Growth in a Time of Debt

By CARMEN M. REINHART AND KENNETH S. ROGOFF*

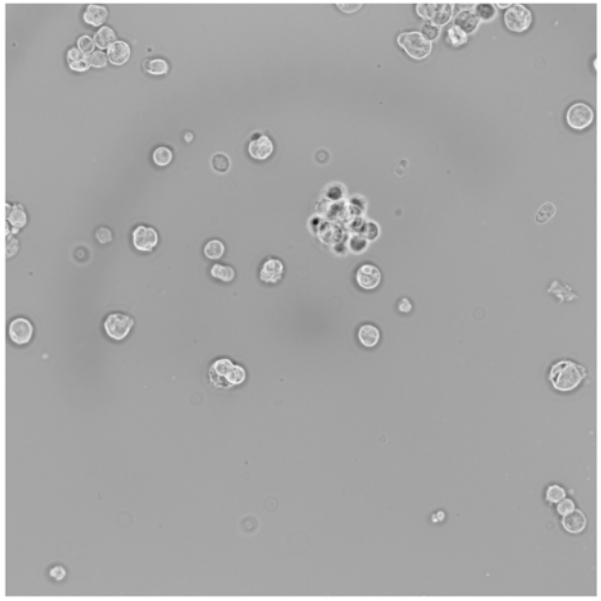
RR paper was covered widely in popular media, often as "high debt/GDP ratio is bad for growth," to support austerity policies in UK and Europe.

Herdon, Ash and Pollin (2014) found that RR had exclusive data selection (cleaning), coding errors, and unconventional weighting.

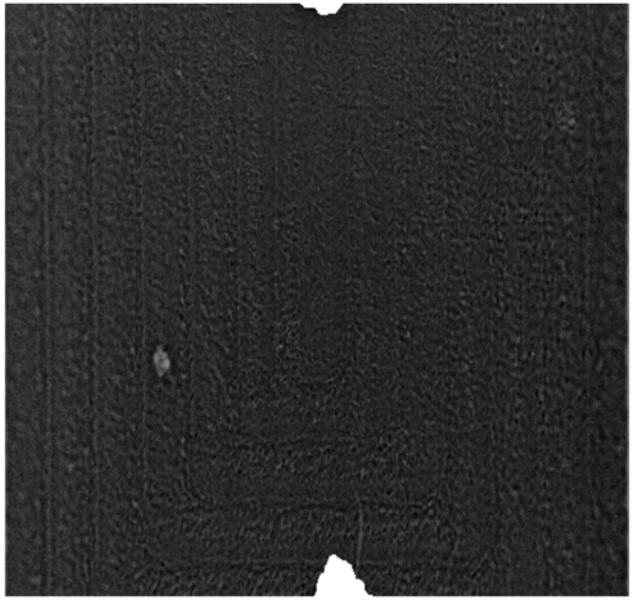
When corrected, RR's conclusion fails to hold.

Data choice
causes data perturbations

High resolution



Low resolution



HCM cell size study after knockout experiments by Ashley Lab at Stanford

Data partitioning

causes data perturbations

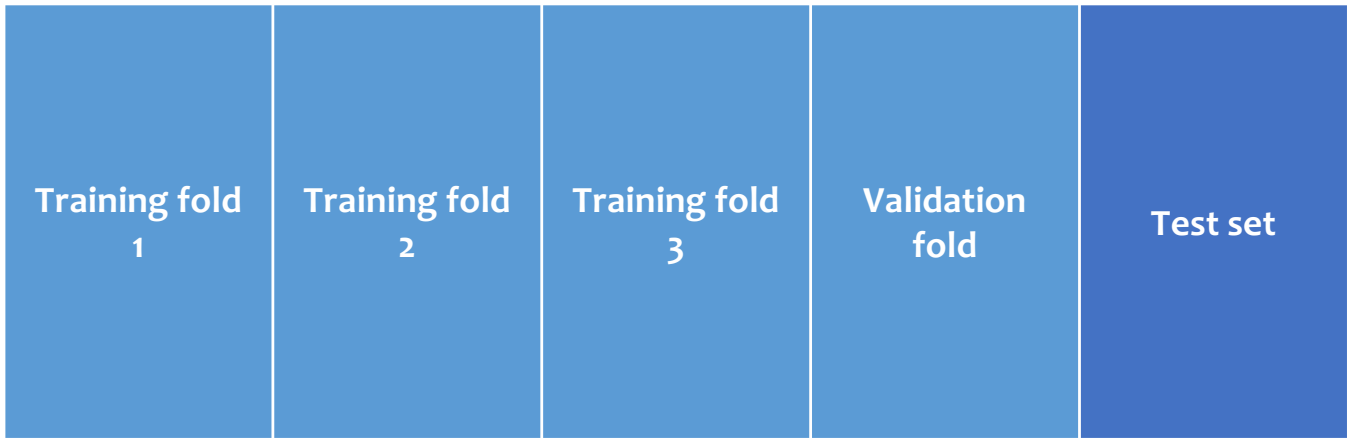
different people create data splits
in different ways

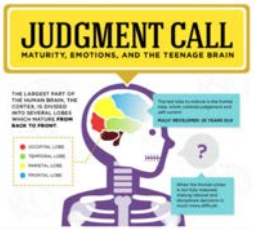
Data perturbations to assess P

Data splitting (stratified by treatment & outcome)

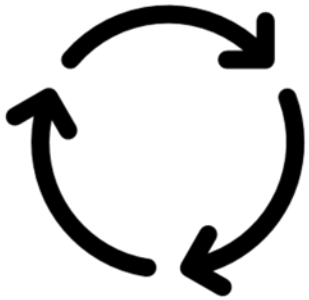
Training folds (cross-validation)

Works well if data units are symmetric





Surrogates for future data **(time split)**



12 settings (or perturbations):
permute val. fold 4 times * re-split 3 times

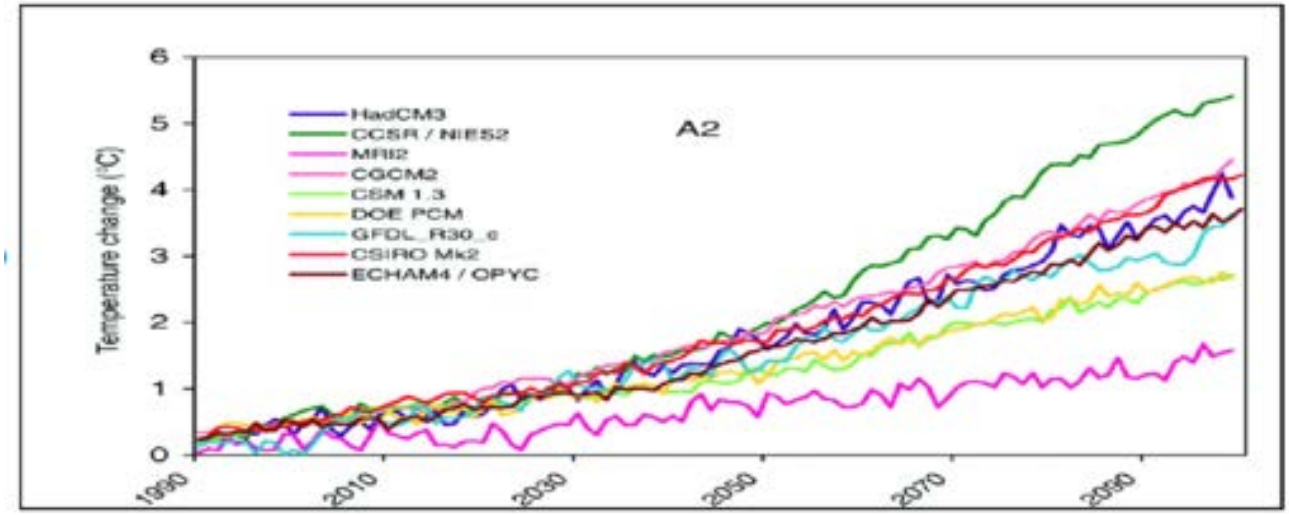
Dwivedi et al (2020) *Stable discovery of interpretable subgroups via calibration in causal studies.* *International Statistical Review*

Choice of method/algorithm causes model perturbations:

different people prefer different methods/algorithms

Some recommendations from PCS

- Make sure "model" means the same thing for a multi-disciplinary team
- Keep multiple versions of annotated or cleaned data
- Keep multiple models after some prediction/reality check, such as in climate modeling



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

Global mean-temp change

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 the data evidence generation **to evaluate the strength** of evidence.

Traditionally, the **p-value** has been used as evidence for decisions, but its use has been so **problematic** that psychology journals have **banned** it.

PCS Inference (Yu and Kumbier, 2020)

Predictability

Use prediction error for model checking

Stability

Assessed across data and model perturbations
(with data perturbation broadly interpreted)

Computability

Implicitly required by P and S

Includes data-inspired simulation

Does not assume a probabilistic generative model -- similar to bootstrap-based inference when the probabilistic model approximates reality well

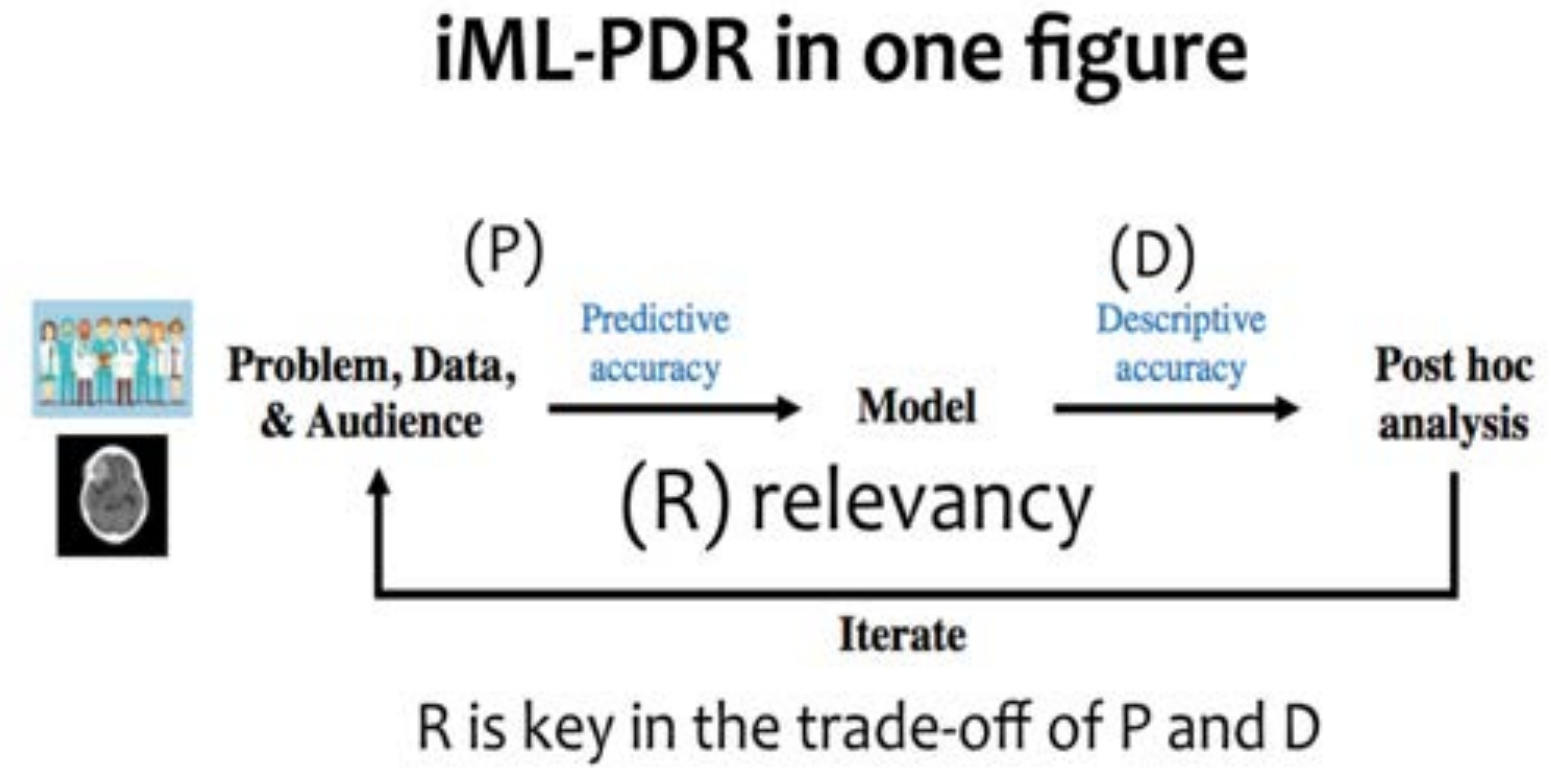
Interpretable ML (iML): necessary for sML and trustworthy AI

(Murdoch, Singh, Kumbier, Abbasi-Asl, and Y., PNAS, 2019)

"**Definitions, Methods and Applications in Interpretable Machine Learning**"

"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 PDR: **P**redictive accuracy, **D**escriptive accuracy and **R**elevancy



Two vignettes: from predictive models to interpretations

Q: what genes drive a heart disease HCM?

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

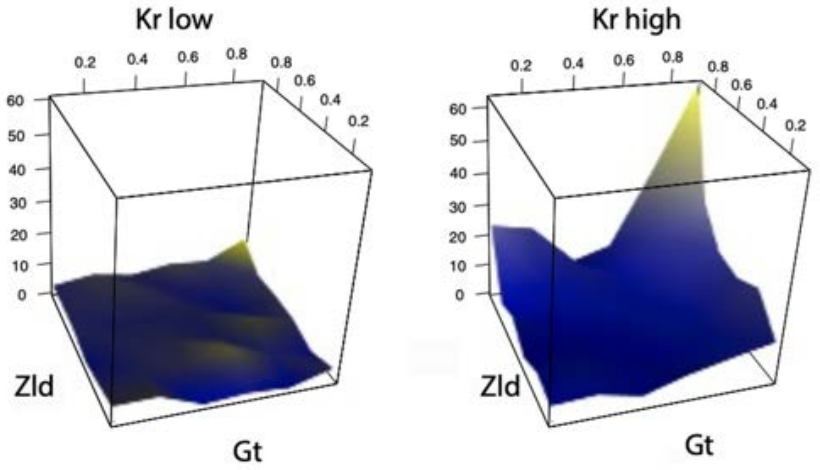
Sumanta Basu[a,b,c,1], Karl Kumbier[d,1], James B. Brown[c,d,e,f,2], and Bin Yu[c,d,g,2]

Co-authors

S. Basu

K. Kumbier

B. Brown

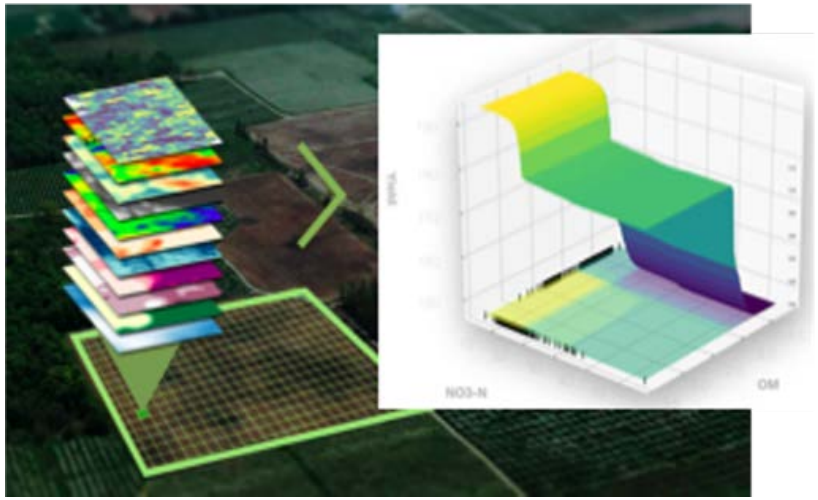


Culmination of 3+ years of work

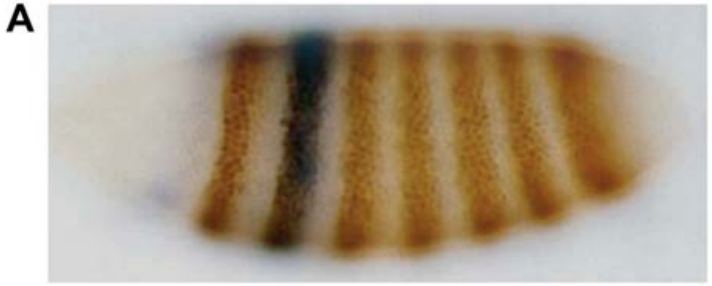
Pattern Recognition *vs.* Pattern Discovery

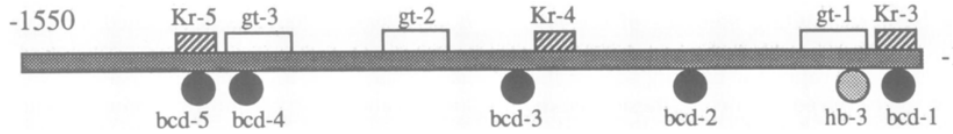
Pattern Recognition:
Finding something for which you already know to look

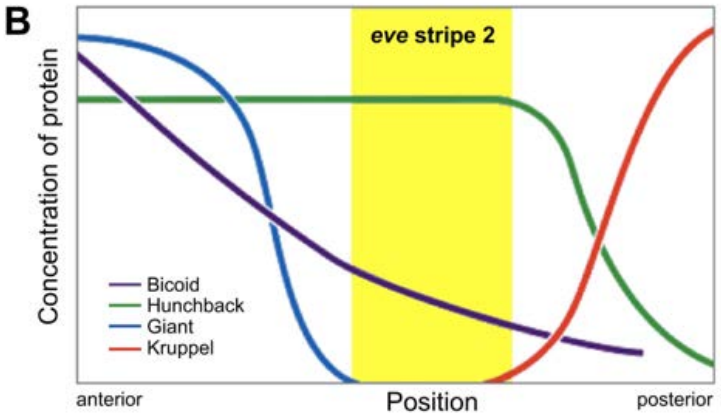
Pattern Discovery:
Identifying structure that hasn't been seen before

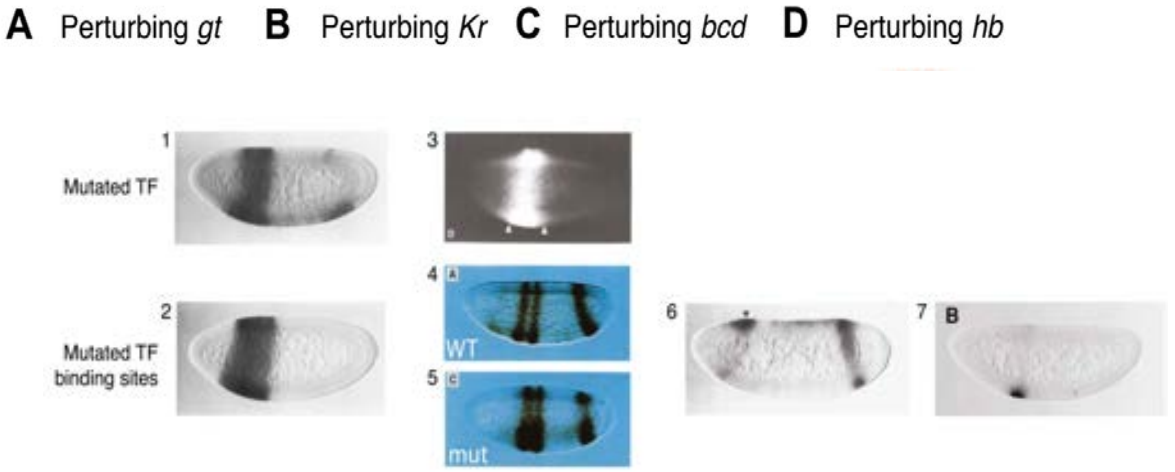


Order-4 interaction regulates *eve* stripe 2





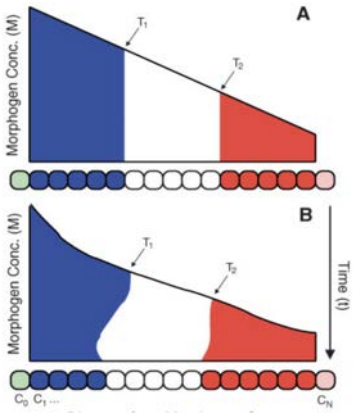




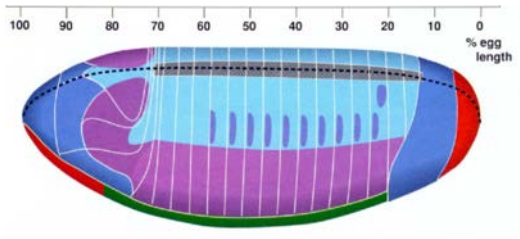
Goto et al. (1989), Harding et al. (1989), Small et al. (1992), Isley et al. (2013), Levine et al. (2013)

Capturing the form of genomic interactions

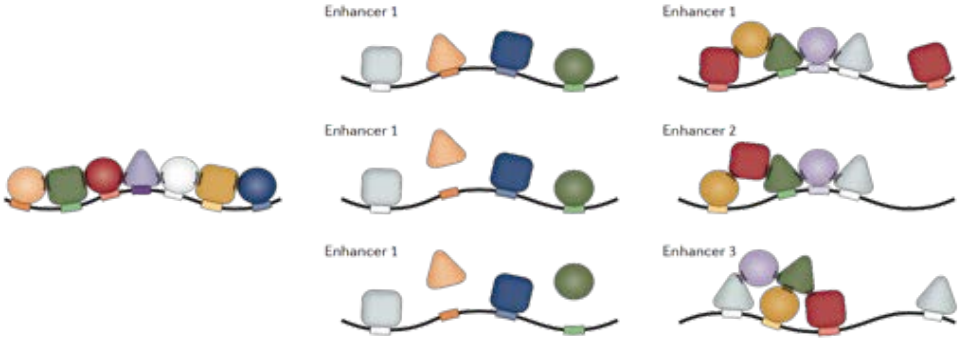
- Interactions are **high-order** and **combinatorial** in nature
- Interactions can **vary across space and time** as biomolecules carry out different roles in varied contexts
- Interactions exhibit **thresholding behavior**, requiring sufficient levels of constitutive elements before activating



(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

$$R(\mathbf{x}) = \underbrace{\prod_{i \in S} 1\{x_i > t_i\}}$$

Order-s interaction,

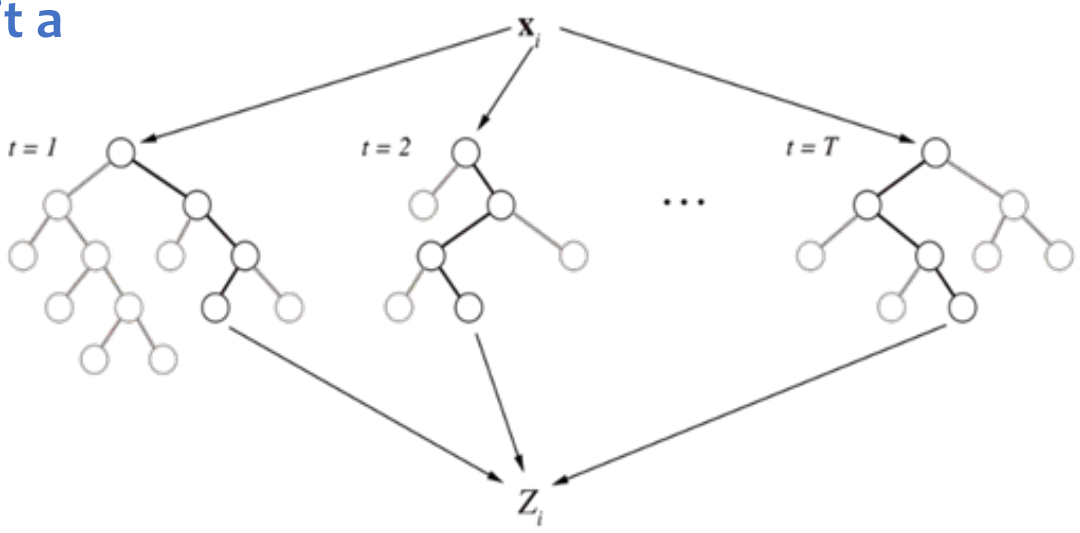
$$S \subseteq \{1, \ldots, p\}, |S| = s$$

Random Forests (RF) (Breiman, 2001)

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

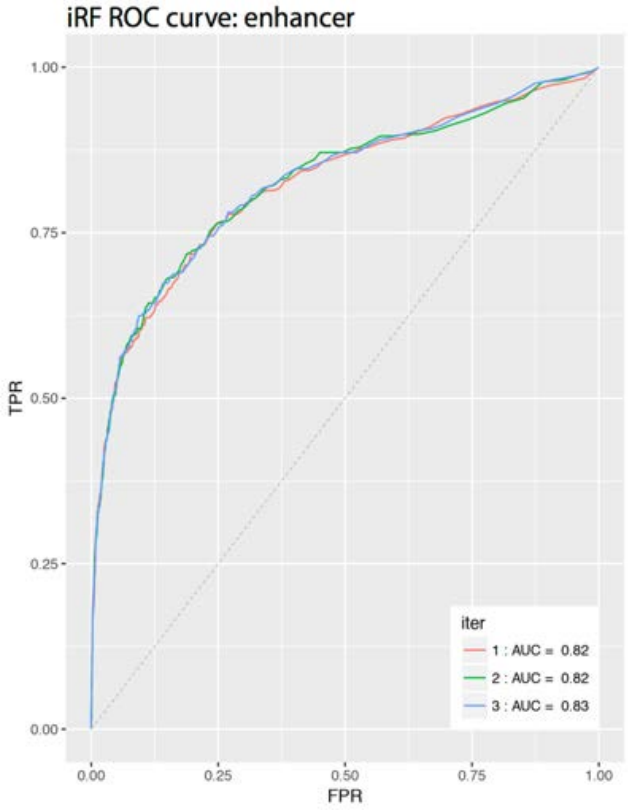
1. Grow CART trees to purity.

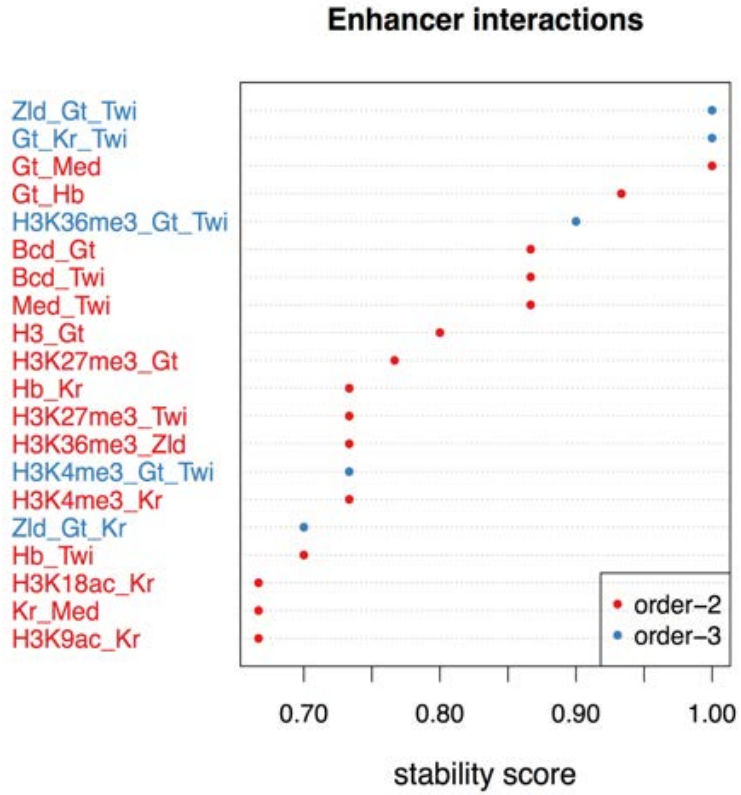
1. When selecting splitting feature, choose a subset of `mtry` features uniformly at random and optimize CART criterion over subsampled features.



Iterative random forests (iRF) finds predictive and stable interactions

for a Drosophila enhancer prediction problem



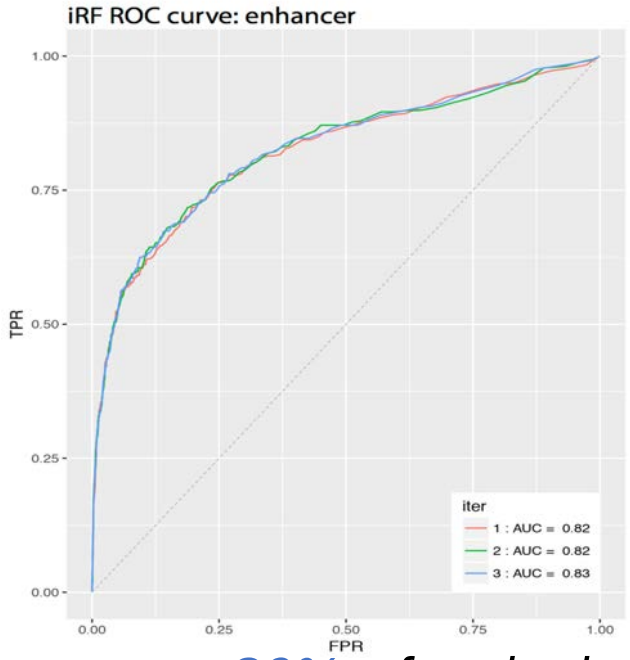


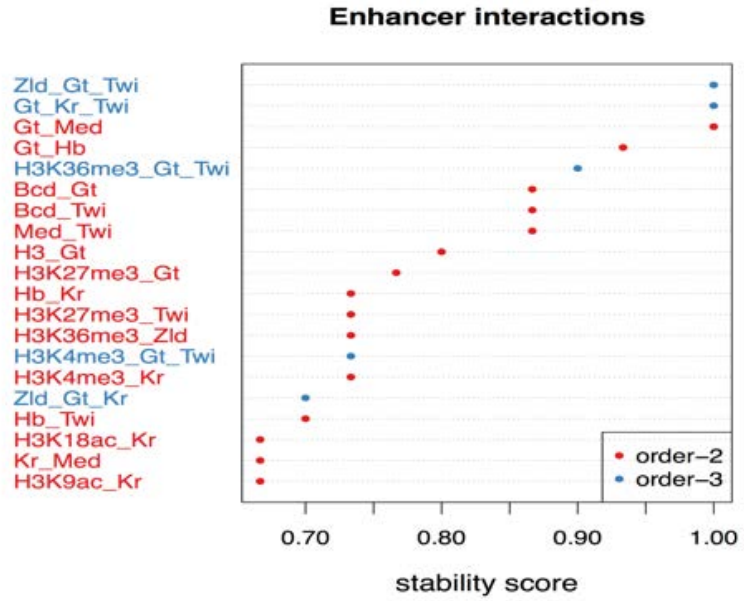
80% of pairwise gene interactions are validated by past biological experiments in the literature

Three-way interactions are recommendations

Iterative random forests (iRF) keep predictive accuracy, and finding stable interactions

for a fruitfly enhancer prediction problem from TFs



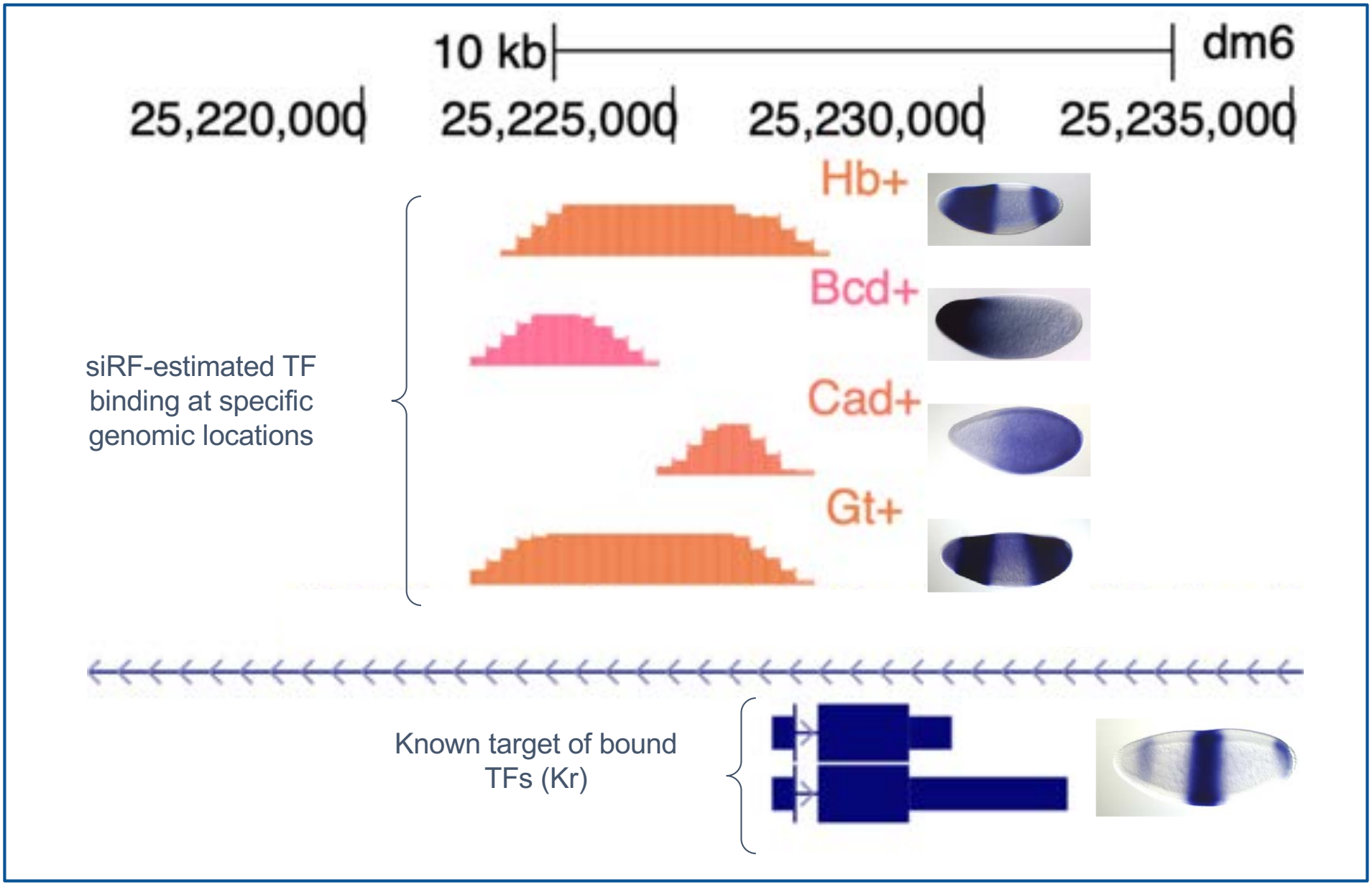


80% of pairwise interactions are validated by past biological experiments in the literature

Three-way interactions are recommendations

siRF-estimated fruitfly TF binding will be made available as UCSC genome browser track

(Kumbier et al, 2021)



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

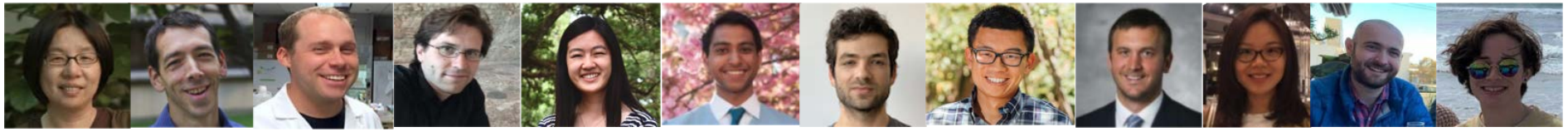
Biohub project for cardiovascular health

Cardiovascular phenotype LVM from cardiac MRI: (n = 30,000 in UKBiobank)

- Left Ventricle Mass (LVM): proxy for a well-known heart disease called Hypertrophic Cardiomyopathy (HCM) (rare variant with 1 in 500)
- Genetic association complex: predictability and stability both low

Yu Group at Berkeley found 4 **predictive and stable gene pairs** that might drive LVM, discovered using their method of iterative random forests (iRF).

Ashley Lab at Stanford Medical School are carrying out **siRNA transfection experiments** with promising preliminary results.



Adaptive wavelet distillation (AWD) from neural networks through interpretations

Co-authors




W. Ha, **C. Singh,** F. Lanusse, S. Upadhyayula

https://arxiv.org/abs/2107.09145

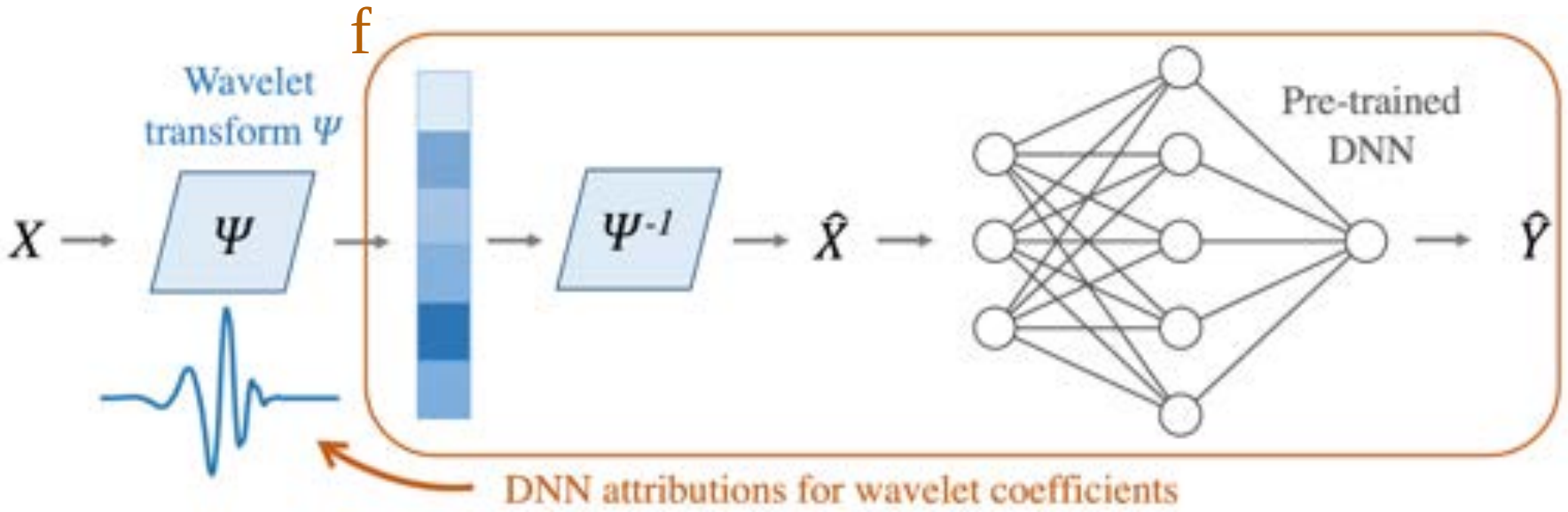
CODE: https://github.com/Yu-Group/adaptive-wavelet-distillation

AWD goal

Given a domain where a Deep Neural Network predicts well, distill it into a simple **learned** wavelet transform

Improves interpretability, compression, and efficiency

Adaptive Wavelet Distillation (AWD)

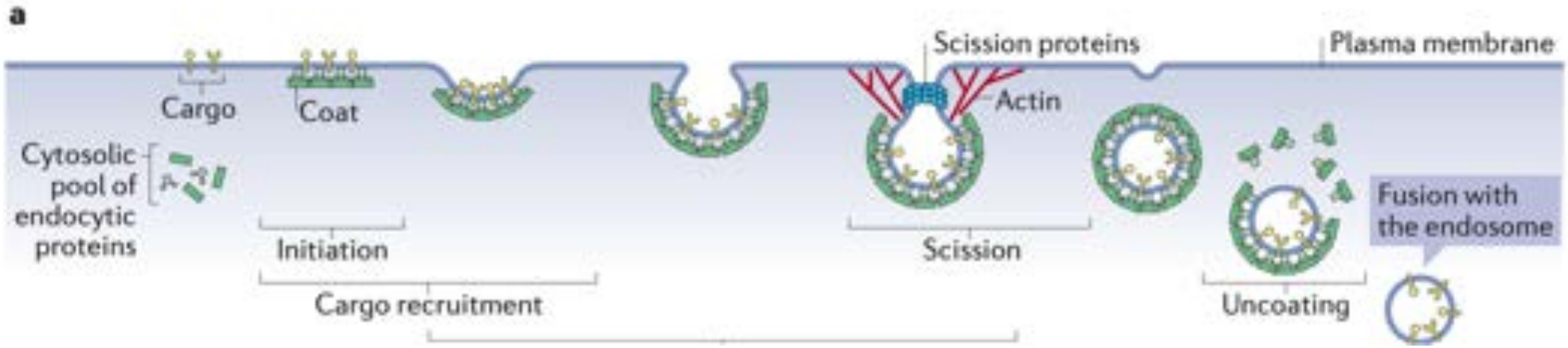


$$\mathcal{L}(h,g) = \underbrace{\frac{1}{m}\sum_i \|x_i - \hat{x}_i\|_2^2}_{\text{Reconstruction loss}} + \underbrace{\frac{1}{m}\sum_i W(h,g,x_i;\lambda)}_{\text{Wavelet loss}} + \underbrace{\gamma \sum_i \|\mathrm{TRIM}_f(\Psi x_i)\|_1}_{\text{Interpretation loss}},$$

AWD finds h and g to minimize L(h, g) through (stochastic) gradient descent

Validation of AWD methodology in a biology problem

Clathrin-mediated endocytosis (CME)

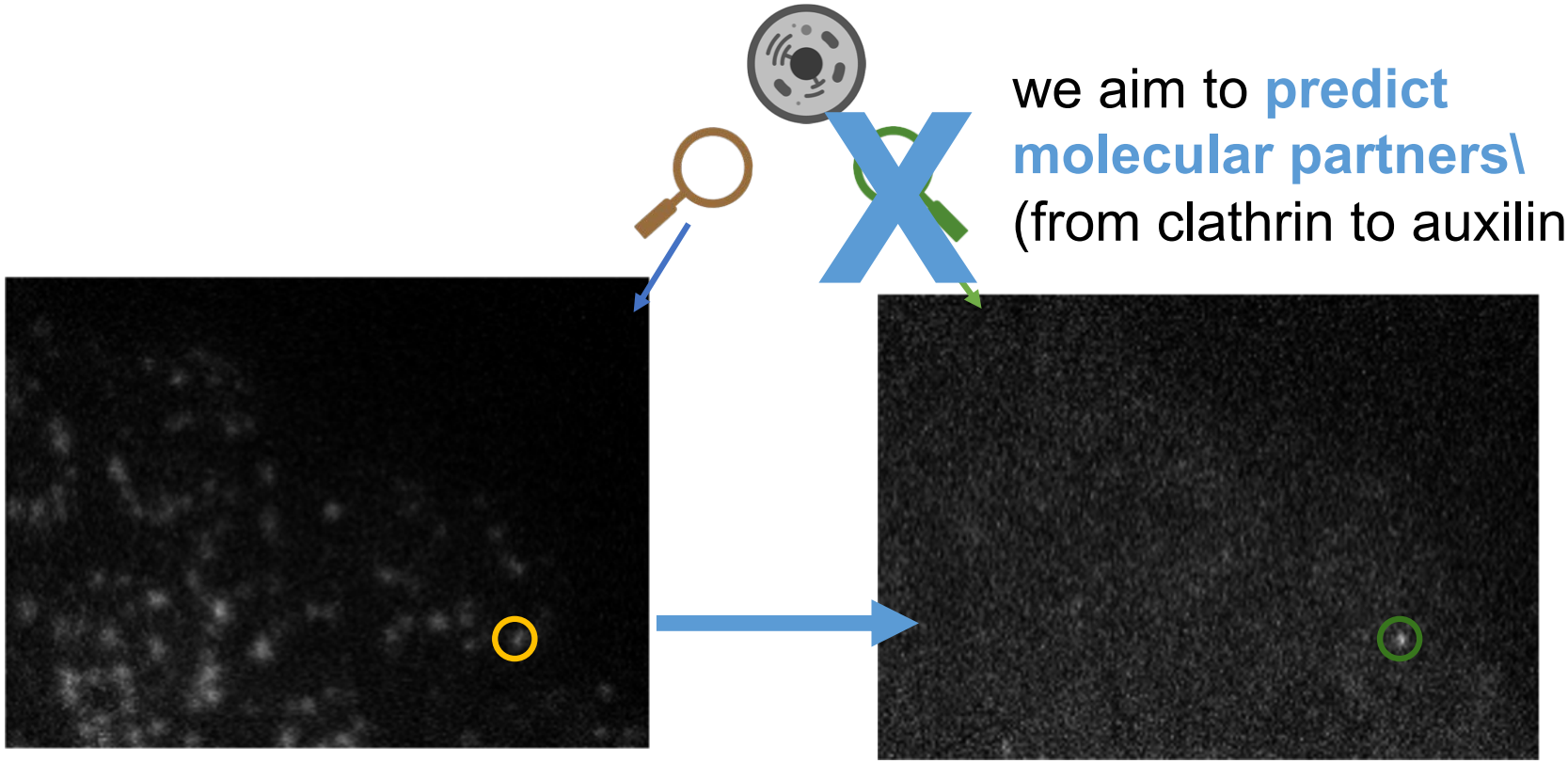


CME ``is a key process in vesicular trafficking that transports a wide range of cargo molecules from the cell surface to the interior.''

-- Kaksonen and Roux (2018) *Nature Reviews.*
https://www.nature.com/articles/nrm.2017.132

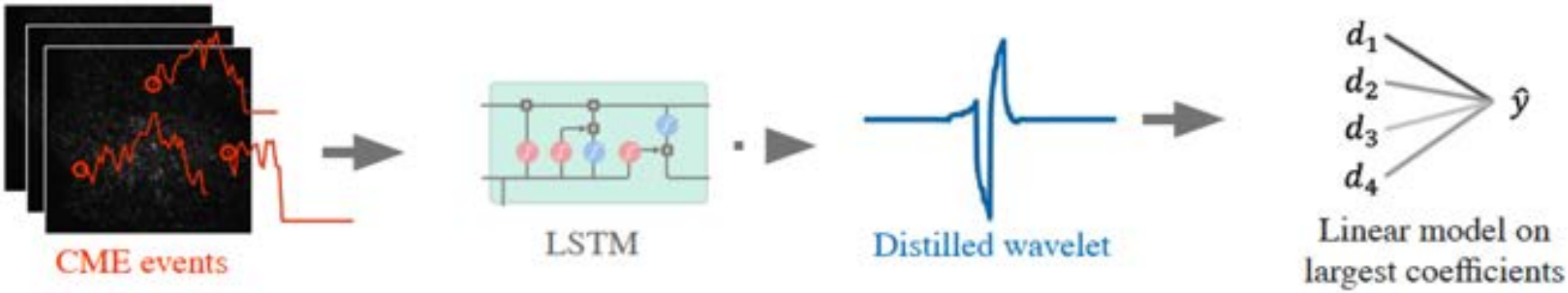
Tracking molecular partners is a central problem in cell biology

...but is experimentally difficult



we aim to **predict molecular partners**
(from clathrin to auxilin)

AWD performs best: prediction, compression, speed-up



For each scale, we select 6 largest coefficients or 30 coefficients

R^2 score

- proportion of variance in Y that is explained by the model

AWD	DB5 Wavelet	LSTM	AWD w/o interp. loss
0.263	0.197	0.237	0.231

A large build up in clathrin fluorescence followed by sharp drop is a highly predictive signature of a successful CME event

PCS Softwares: easy to use

Design Principles:

Transparent (**P**)
Realistic (**P**)

Intuitive (**C**)
Modular (**C**)
Efficient (**C**)

Reproducible (**S**)

Veridical Flow

PCS-style data analysis made easy!

simChef

PCS-style simulations made easy!

Veridical Flow (on-going):
a PCS wrapper for ML in Python

Make prediction screening and stability analysis **simple**, **reproducible**, and **computationally efficient**

Create intuitive capabilities to **inspect**, **manipulate**, and **visualize** the ML pipeline


A. Agarwal


J. Duncan


R. Kapoor


C. Singh

github.com/Yu-Group/veridical-flow

Veridical Flow: a PCS wrapper for ML in Python

Agnostic to underlying ML framework

Closely integrated with capabilities of Ray & MLflow (both from CS at Berkeley)

- Distributed computation 

- Tracking results across perturbations
- Visualization of performance metrics
- Packaging trained models for reproducibility and real-world deployment

github.com/Yu-Group/veridical-flow

simChef (on-going):
an R package for PrinCipled Simulations

- Provide **powerful tools** for **evaluating methods** efficiently across a **variety of scenarios** (perturbations)

- Encourage **adherence to principles** of strong, reliable, trustworthy data science


J. Duncan


C. F. Elliott


T. Tang

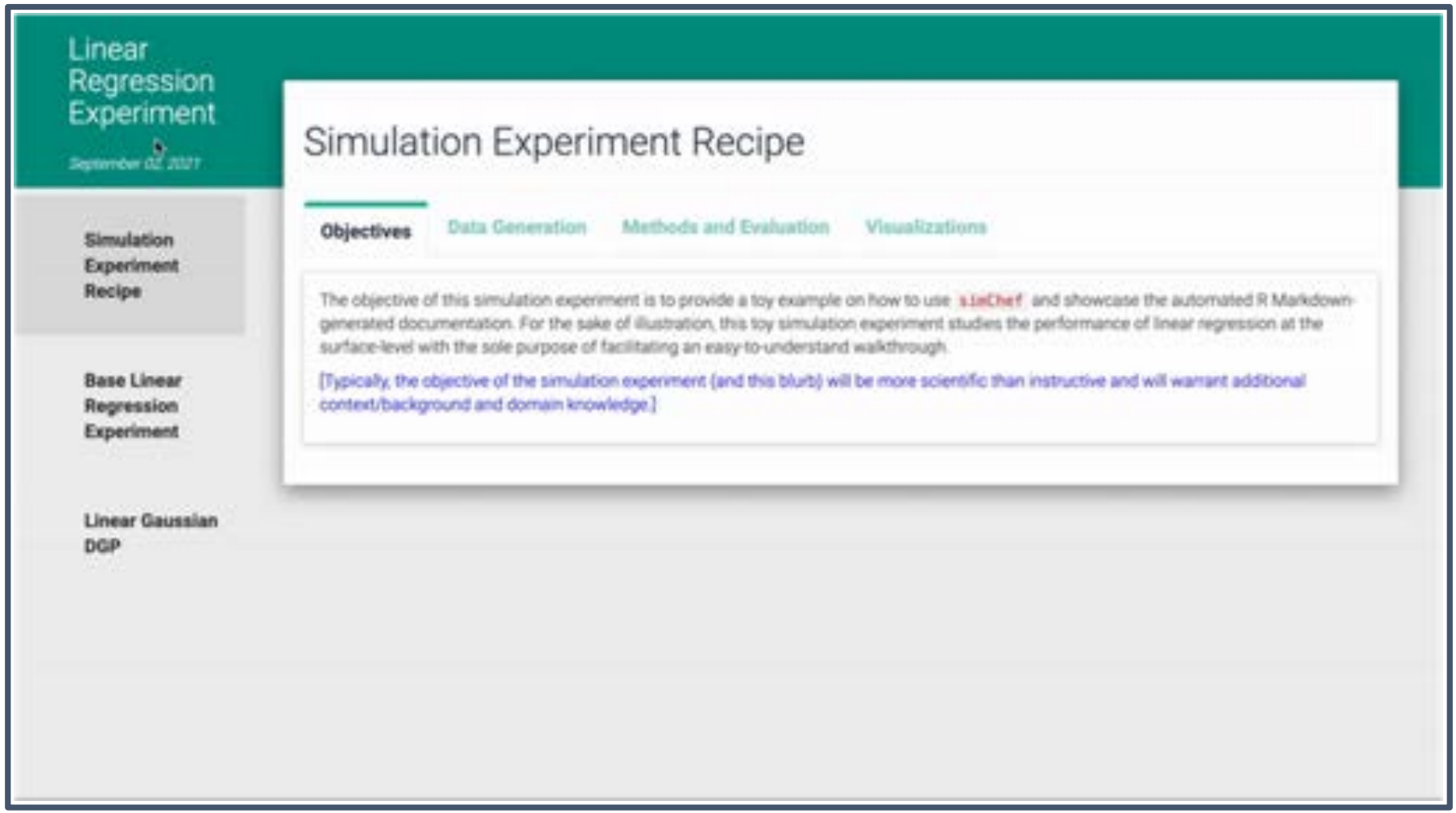

M. Behr


K. Kumbier

github.com/Yu-Group/simChef

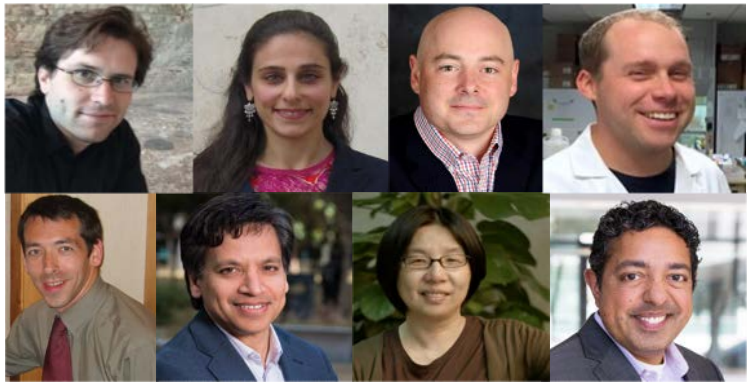
simChef Interactive PCS Documentation

- Automated generation of **transparent** and **reproducible documentation** in R Markdown for **rapid** and **veridical communication** of scientific results.



github.com/Yu-Group/simChef

PCS inference case study: epiTree for epistasis discovery



Post-docs & Students:


M. Behr


K. Kumbier


M. Aguirre


A. Cordova-Palomera

Learning epistatic polygenic phenotypes with Boolean interactions
https://www.biorxiv.org/content/10.1101/2020.11.24.396846v1

epiTree uses iRF for interaction selection and new PCS p-value for epistasis testing

Summary

- **Veridical data science** is necessary for trustworthy AI
- **PCS framework** (with **documentation on GitHub**) for Veridical data science (8 successful PCS case studies including iRF)
- **Scientific ML** is part of trustworthy AI: it needs **interpretable ML**
- iML based on **PDR**, decision trees from **iRF**, **AWD** from DNNs
- **Domain knowledge** is important, and building on this, **PCS** generates testable results for **external validation** (e.g., experiments and follow-up studies)
- Importance of reliable and easy-to-use **software**
- **Trustworthy AI needs a fair reward system and incentives**

Software from the Yu Group

stat.berkeley.edu/~yugroup/code.html

- iRF, siRF
- epiTree
- Superheat
- Veridical Flow **(on-going)**
- simChef: PCS Simulation package in R **(on-going)**

Also, newly accepted by the *Journal of Open Source Software* (with 10k+ downloads, and 240+ github stars):

imodels (Python) for concise, transparent, and accurate predictive modeling: provides a simple interface for fitting & using state-of-the-art interpretable models compatible with scikit-learn.

Hope PCS is useful for your projects

Bin Yu website: https://binyu.stat.berkeley.edu/
Papers: https://binyu.stat.berkeley.edu/papers

1. B. Yu and K. Kumbier (2020), **"Veridical data science",** PNAS. --- PCS framework

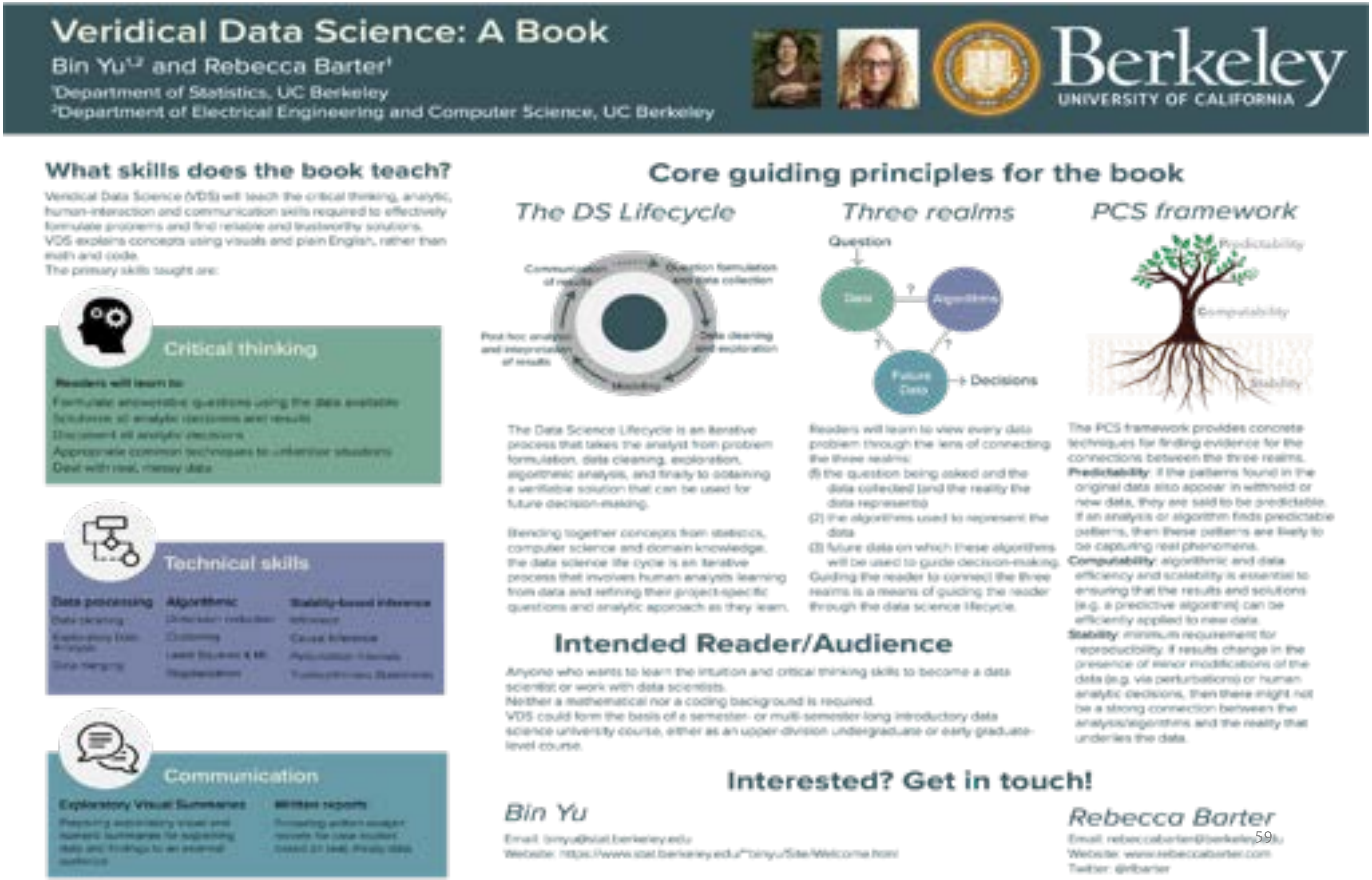
2. S. Basu, K. Kumbier, B. Brown and B. Yu (2018). **"Iterative random forests to discover predictive and stable high-order interactions",** PNAS

3. K. Kumbier, S. Basu, J. Brown, S. Celniker, B. Yu **(2018) Refining interaction search through signed iterative Random Forests (signed iRF or siRF)** **https://arxiv.org/abs/1810.07287**

4. **M. Behr, K. Kumbier**, M. Aguirre, R. Arnaout, E. Ashley, A. Butte, R. Arnout, B. Brown**, J. Priest, B. Yu** (2020). **"Learning epistatic polygenic phenotypes with Boolean interactions" https://www.biorxiv.org/content/10.1101/2020.11.24.396846v1**

5. Singh, Ha and Yu (2021). Interpreting and improving deep-learning models with reality checks (iML review paper including AWD) https://arxiv.org/abs/2108.06847

Book using PCS for DSLC by Yu and Barter with MIT Press
free online interactive copy (plan: 2021 fall)



Berkeley Computing, Data Science, and Society **(CDSS led by Jennifer Chayes)**

About ▾ Academics ▾ Research ▾ News ▾ Events ▾ Support Us

A data science program for everyone

Data8 on EdX

Berkeley UNIVERSITY OF CALIFORNIA

Professional Certificate in
Foundations of Data Science

I'm interested

Data Science Major
Data8 (1000+ students)
Data100 (1000+)
Data102 (200+)
co-created and co-taught by Stats and EECS faculty

What you will learn

- How to interpret and communicate data and results using a vast array of real-world examples from different domains
- How to make predictions using machine learning and statistical methods
- Computational thinking and skills, including the Python programming language for analyzing and visualizing data
- How to think critically about data and draw robust conclusions based on incomplete information

Expert instruction
3 skill-building courses

Self-paced
Progress at your own speed

4 months
4 - 6 hours per week

$537.30 ~~$597~~ USD
For the full program experience