Provable Recovery of Boolean Interactions based on Random Forests

Bin Yu

Statistics and EECS, UC Berkeley

BIRS Workshop "Mathematical Statistics and Learning"

Nov. 29, 2021

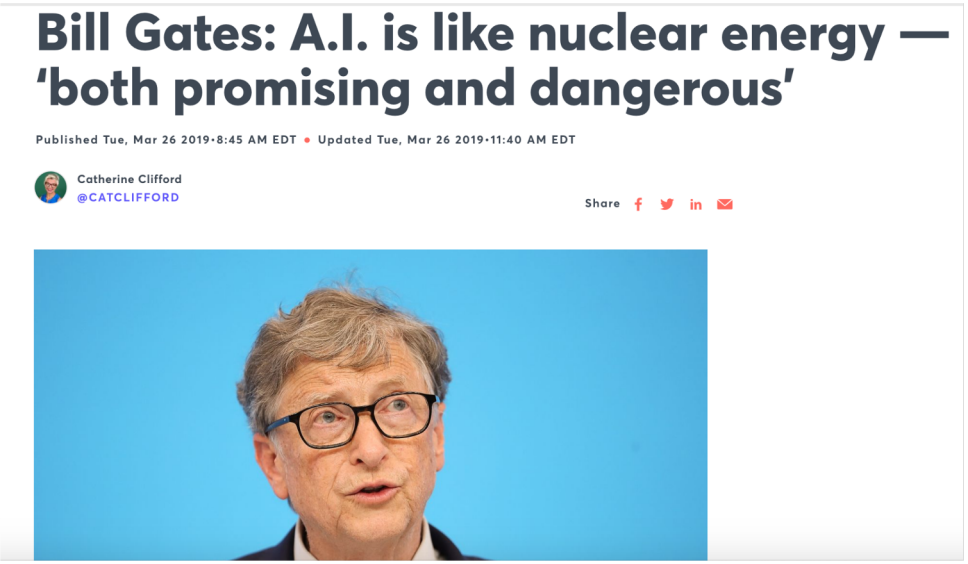
AI is part of modern life

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

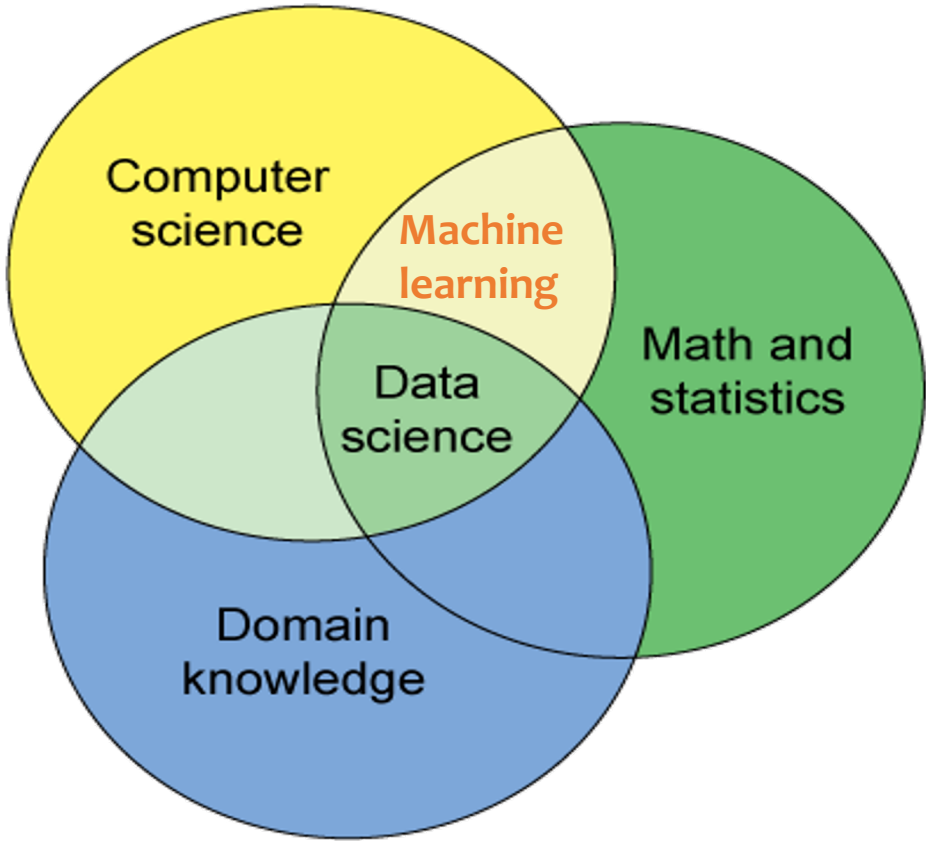
Cosmology

Economics

Political Science

… and beyond

Data science (DS) is a key element of AI

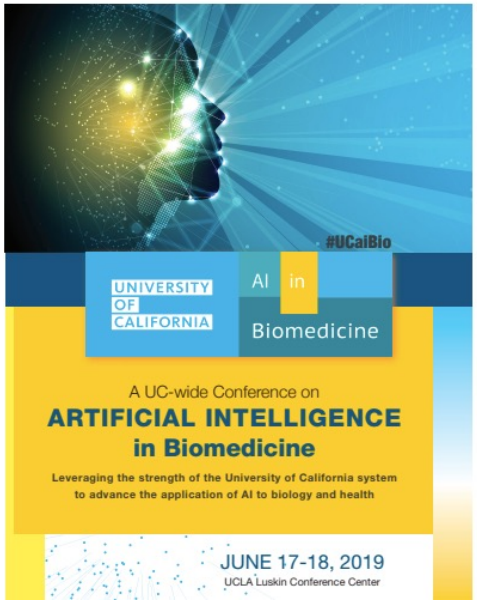


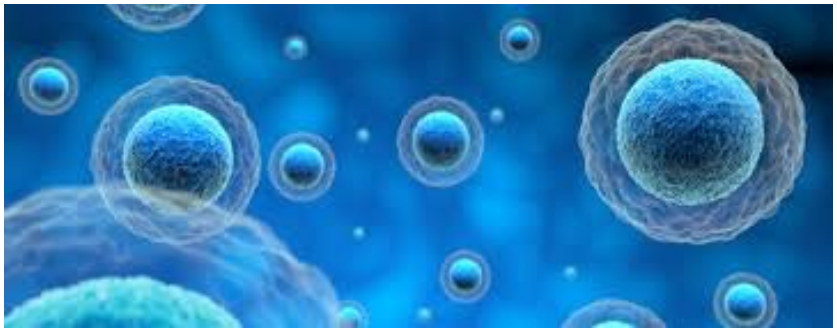
Conway's Venn Diagram

Goal:

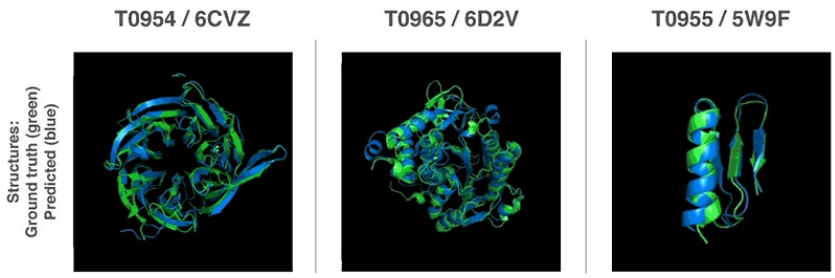
Leverage **algorithms**
to combine **data**
with **domain knowledge**
to make decisions
and generate
new knowledge

Biomedical data problems are pressing





medium.com



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

website of S. Saria at JHU

Trustworthy AI (data science): two complementary approaches

- **Best practices to maximize the promise (preventative)**
- Risk management to reduce the danger (intervention)

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, control theory, …]

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

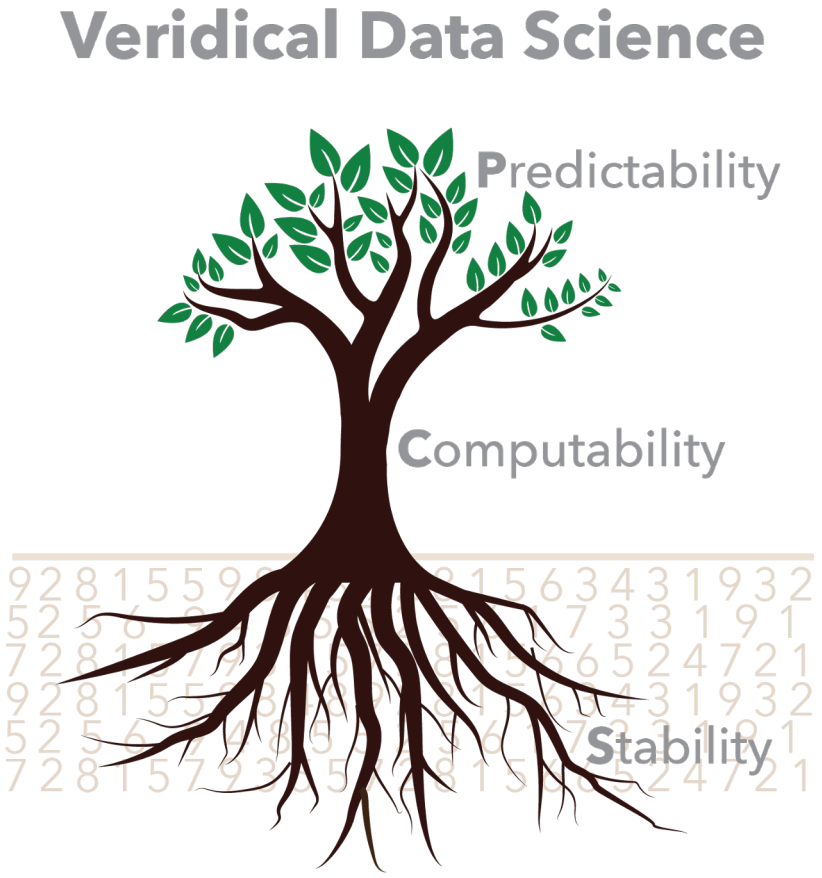


Image credit: R. Barter

2001

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

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

Machine learning

Statistics

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

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

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

2001


Machine Learning
October 2001, Volume 45, Issue 1, pp 5–32 | Cite as
Random Forests
Authors
Authors and affiliations
Leo Breiman
Article
73
Shares
208k
Downloads
27k
Citations

Scientific Machine Learning

- 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 for validation and interpretation
- Algorithms are available through open source software

Rest of the talk

- Motivating case study of PCS and reason for relevant theory

  Iterative random forests (iRF) for predictive and stable Boolean interaction discovery (serving also as non-linear model selection)

- Provable Boolean interaction discovery results

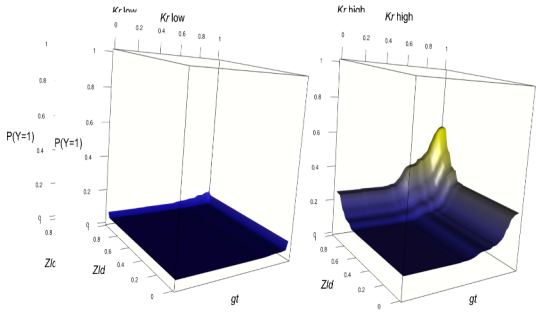
  for a tractable version of iRF, LSSFind,
  under a new relevant generative Local Spiky Sparse (LSS) model

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

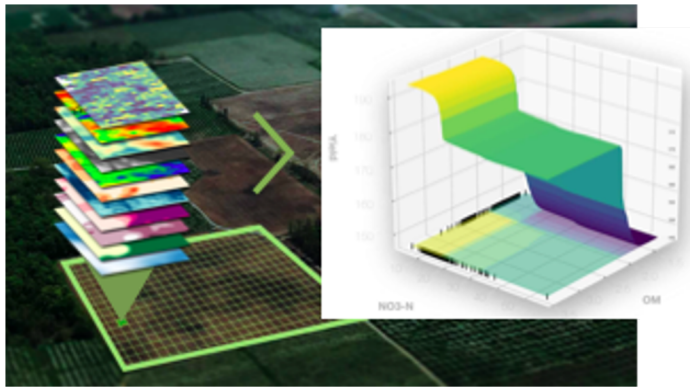


Culmination of 3+ years of work

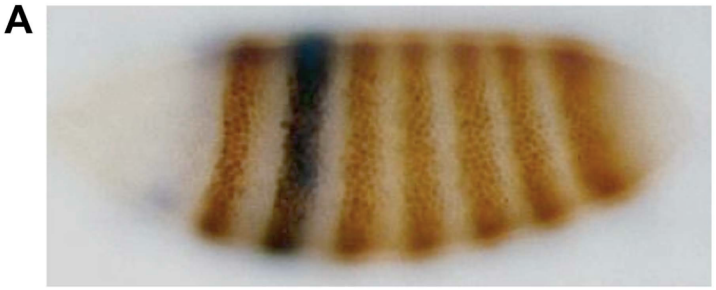
Pattern Recognition *vs.* Pattern Discovery

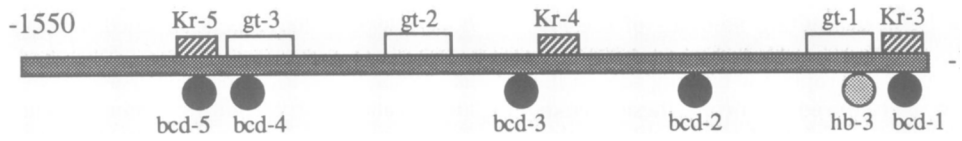
Pattern Recognition:
Finding something for which you already know to look

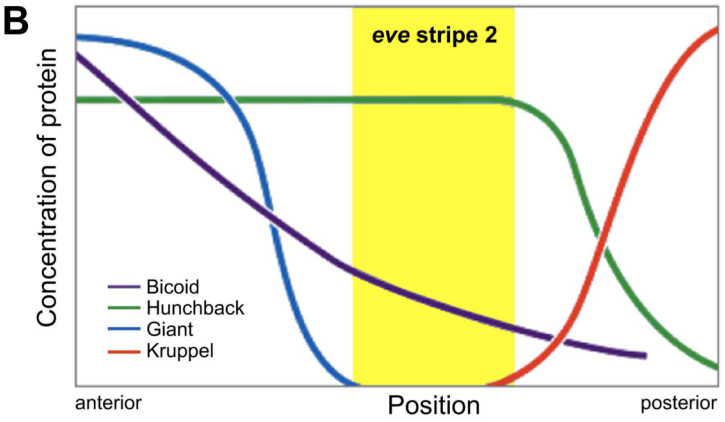
Pattern Discovery:
Identifying structure that hasn't been seen before

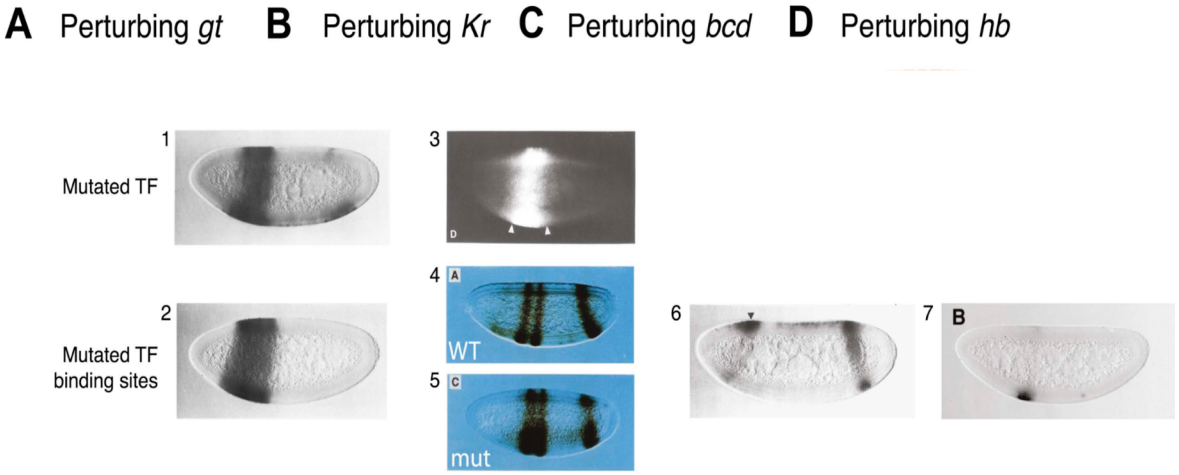


Order-4 interaction regulate *eve* stripe 2





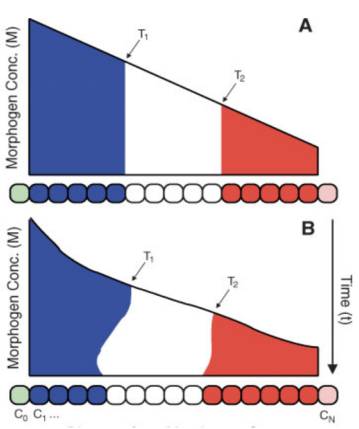




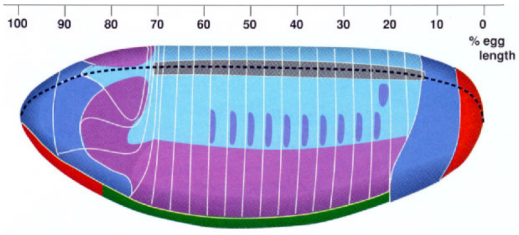
Goto et al. (1989), Harding et al. (1989), Small et al. (1992),
Isley et al. (2013), Levine et al. (2013)

Capturing the form of genomic interactions

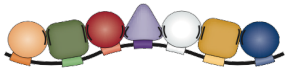
- Interactions are high-order and combinatorial in nature
- Interactions can vary across space and time as biomolecules carry out different roles in varied contexts
- **Interactions** exhibit **thresholding behavior**, requiring sufficient levels of constitutive elements before activating

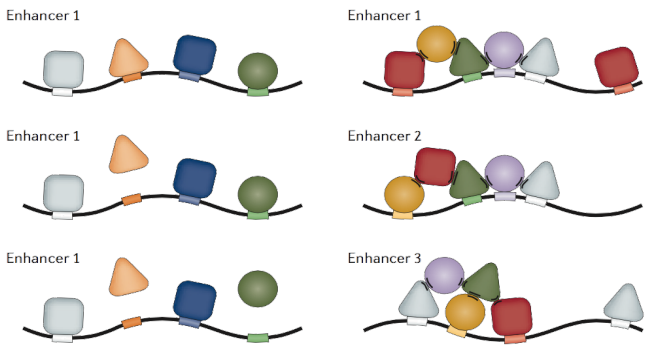


(Wolpert, 1969;
Jaeger and Reinitz, 2006)



(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,

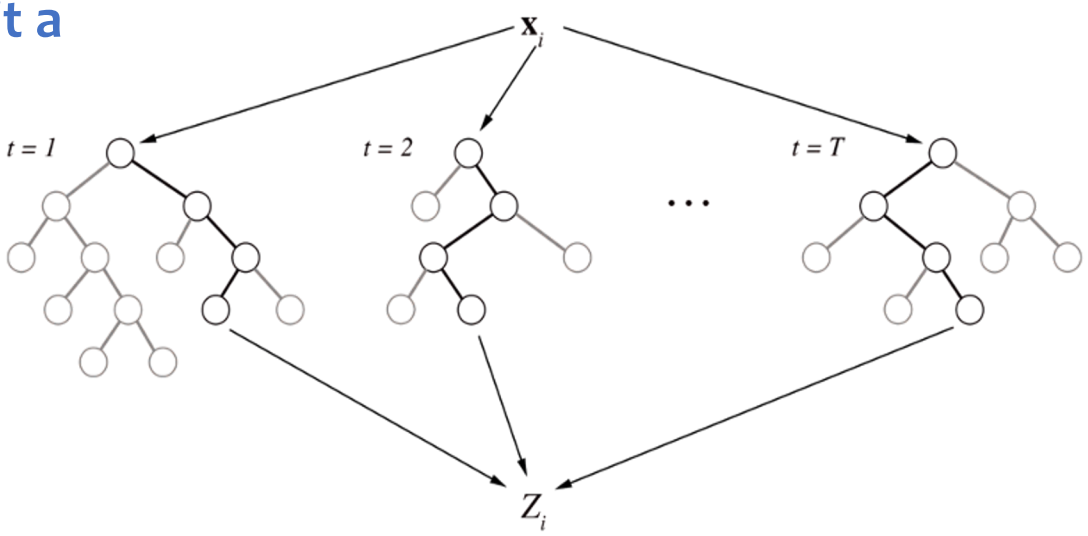
$$S \subseteq \{1, \ldots, p\}, |S| = s$$

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.



Previous works using RF for interaction discovery

- Key idea: co-occuring features on the same path imply interaction
  Wan et al (2009), Yoshida and Koike (2011), …

- Problem: these features are unstable

iterative Random Forests (iRF)

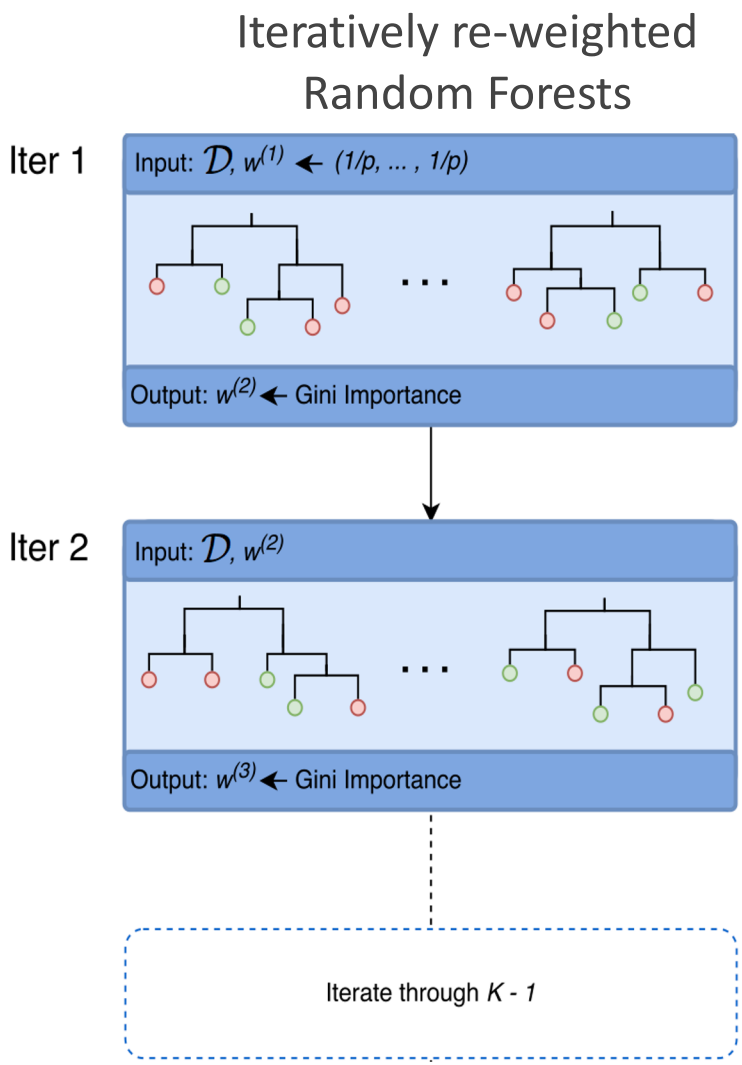
Basu, Kumbier, Brown and Yu (2018)

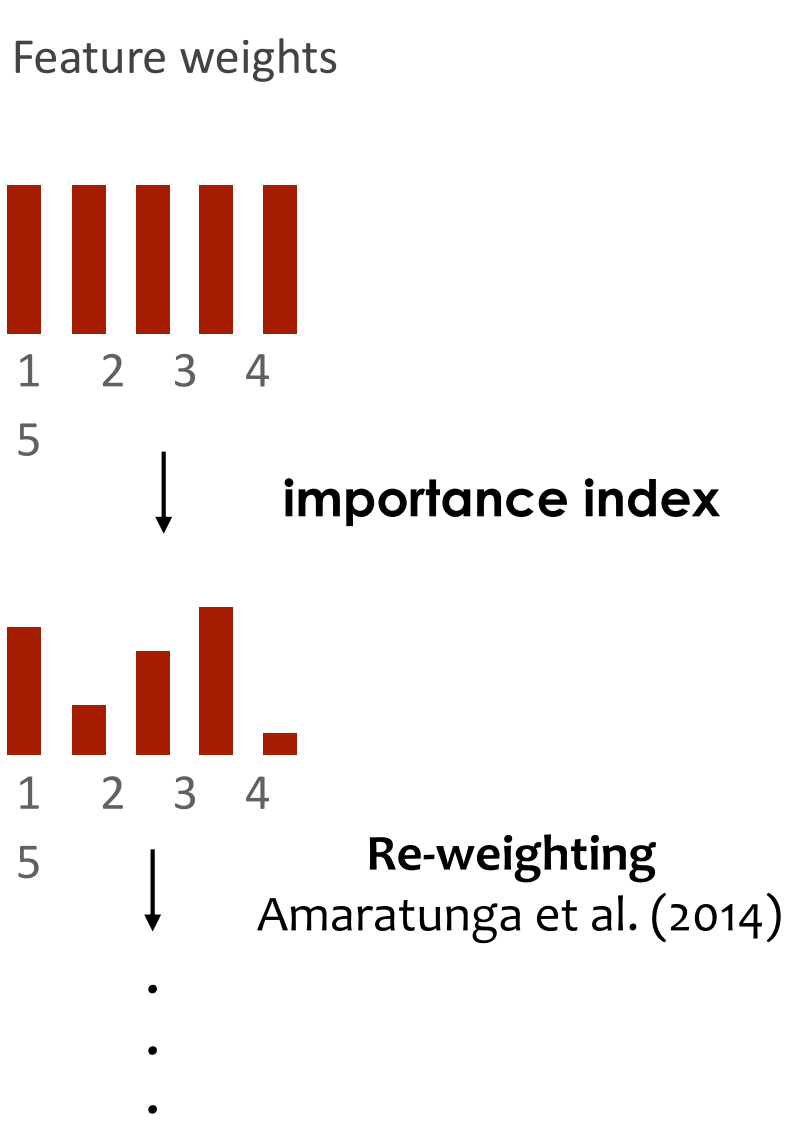
Core idea: add stability to RF

1. Soft dim reduction using importance index
2. Random interaction trees (RIT) 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





Generalized RIT for Decision Trees

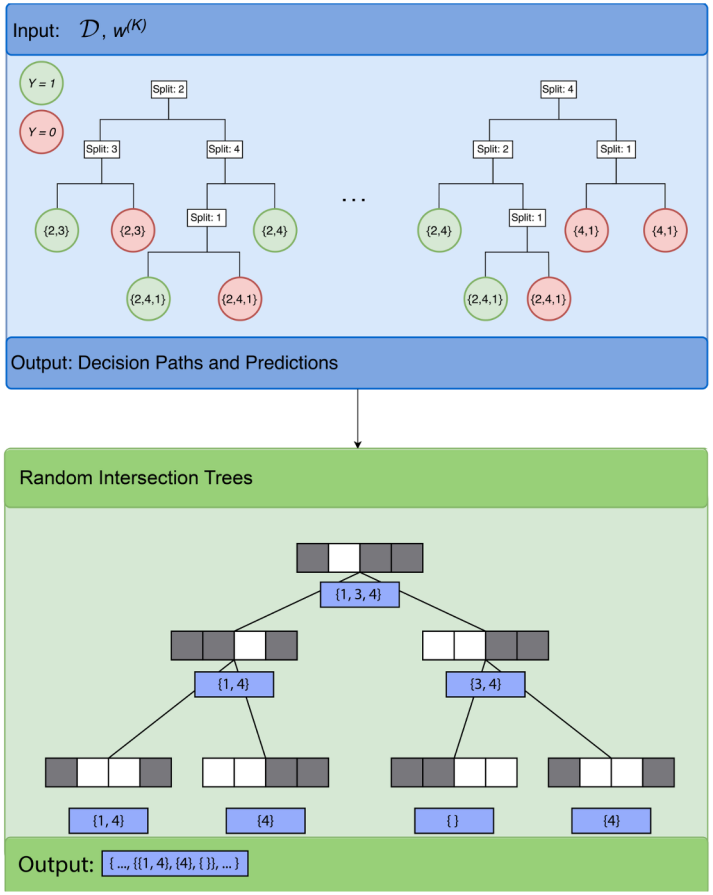
fast computation uses sparsity

(Random Intersection Trees (RIT), Shah and Meinshausen, 2014)

$\mathcal{I}_{i_t} \subseteq \{1, \ldots, p\}$ *Feature-index set* for leaf node containing observation $i = 1, \ldots, n$ in tree $t = 1, \ldots, T$

$Z_{i_t} \in \{0, 1\}$ *Prediction* for the leaf node containing observation $i = 1, \ldots, n$ in tree $t = 1, \ldots, T$

$$\mathcal{S} \leftarrow \mathrm{RIT}(\{\mathcal{I}_{i_t}, Z_{i_t}\}, C)$$



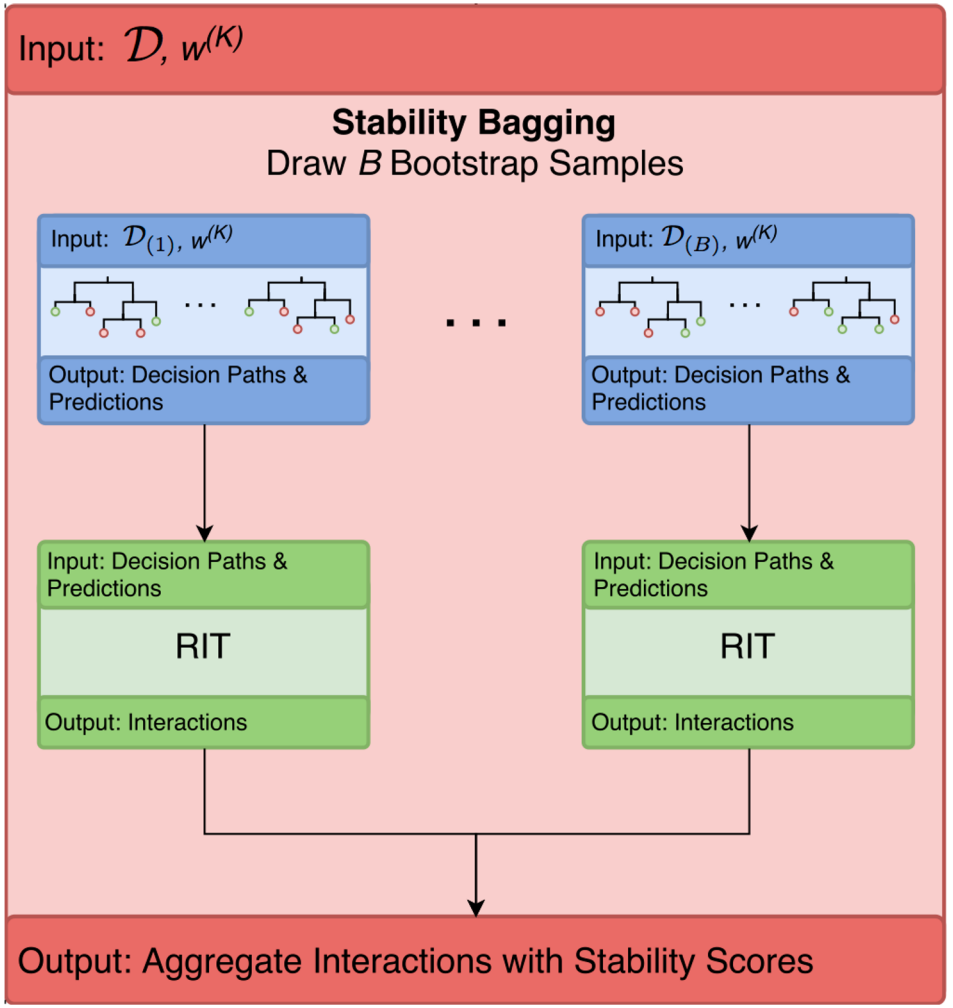
Stability bagging

Output feature interaction sets with stability scores:

$$\{S, sta(S)\}$$

$$S \subseteq \{1, \ldots, p\}$$

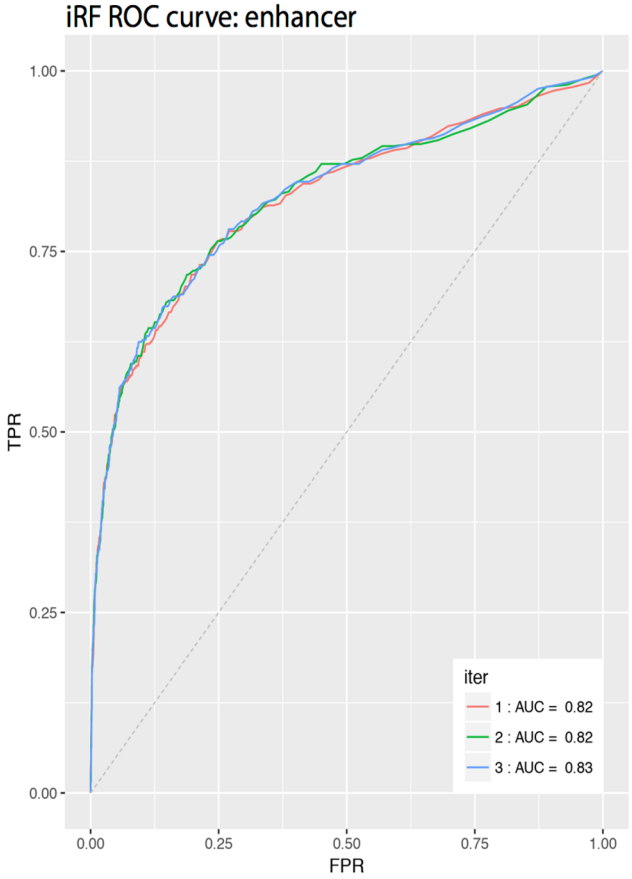
$$sta(S) = \frac{1}{B} \cdot \sum_{b=1}^{B} 1(S \in \mathcal{S}_b)$$

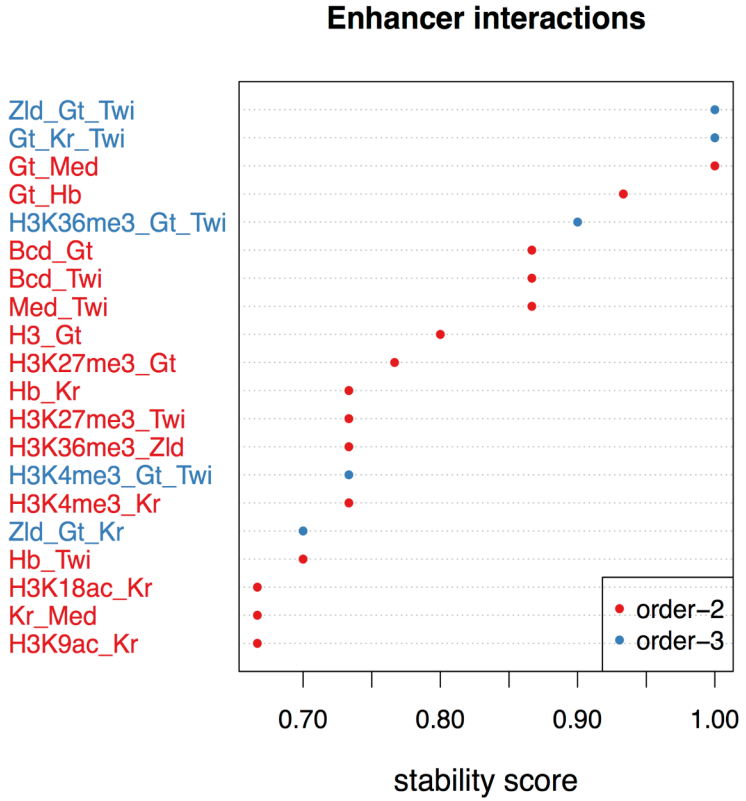


iRF uses PCS = RF (P) + RIT (C) + Stability (S)

iRF keeps predictive accuracy, and finds stable interactions

for a Drosophila enhancer prediction problem





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

signed iterative Random Forests (siRFs)

Kumbier, Basu, Brown and Yu (2021)

Core ideas

1. Soft dim reduction using importance index

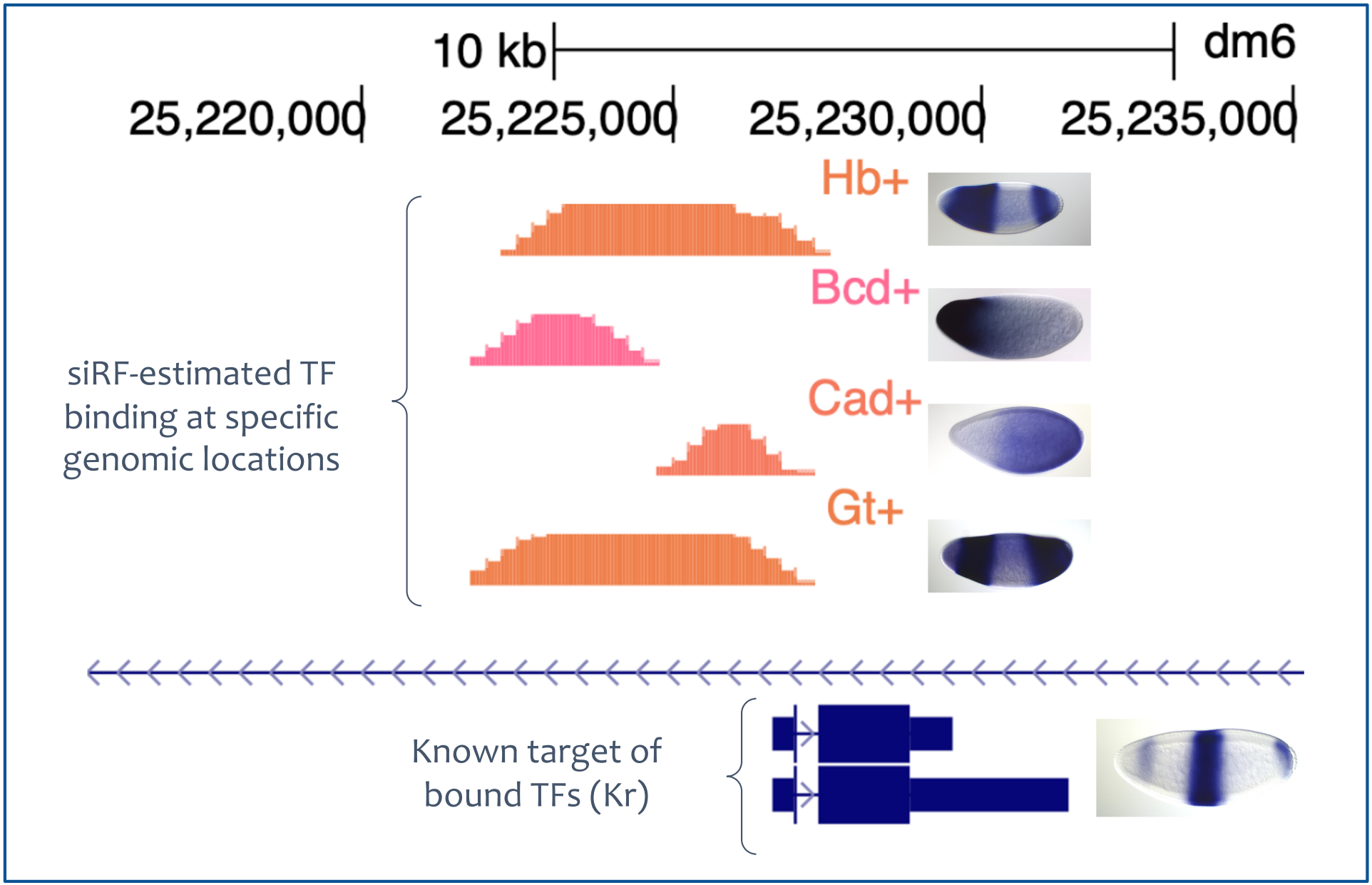
2. Random interaction trees to find intersections of signed paths

   e.g. split $X1 < 0.3$ will be coded as (1, -1)
   split $X2 > 0.5$ will be coded as (2, +1)

3. Outer-loop bagging assesses stability

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

(Kumbier et al, 2021)

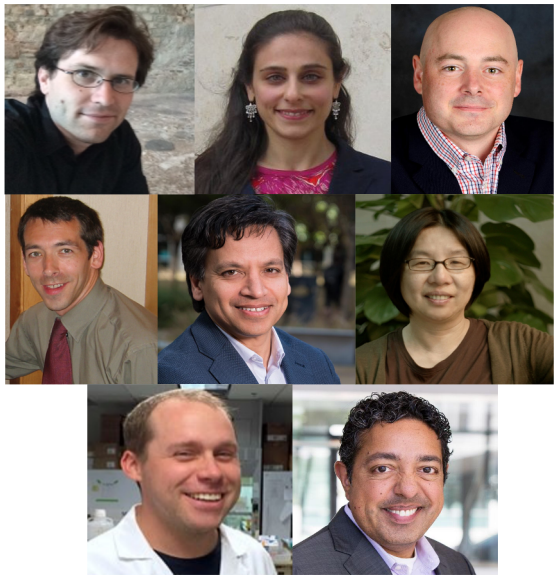


Berkeley UNIVERSITY OF CALIFORNIA UCSF Stanford CHAN ZUCKERBERG BIOHUB

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





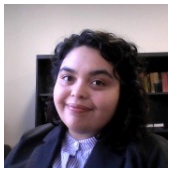



M. Behr K. Kumbier M. Aguirre A. Cordova-Palomera Q. Wang N. Youlton









C. Weldy W. Hughes A. Agarwal T. Tang O. Ronen X. Li A. Kenney

Biohub project for cardiovascular health

Cardiovascular phenotypes from MRI: (n = 30,000 UKBiobank subjects)

1. Continuous phenotype Left Ventricle Mass (LVM) -- proxy for a well-known heart disease called Hypertrophic Cardiomyopathy (HCM)

2. Much less data available (rare variants)

3. Genetic association more complex:

   predictability and stability are both low

Yu Group at Berkeley has found 4 predictive and stable gene-pairs discovered singed iterative random forests (siRF) that might drive LVM.

Ashley Lab at Stanford medical school have carried out siRNA transfection experiments with promising preliminary results.

Another form of evaluation

Theory to understand interaction discovery using tree ensembles from RF under relevant generative models

Previous work on RF theory

- Regression function estimation

consistency, rates of nonparametric convergence (under smoothness conditions), asymptotic normality

Breiman (2004), Biau (2012), Mentch and Hooker (2015), Scornet et al (2015), Duroux and Scornet (2016), Wager and Athey (2018)…

- Feature importance measures: dealing with noisy features, permutation test based measures

Loupoe et al (2013), Li et al (2019), Loecher (2020), Zhou and Hooker (2020),…
Ishwaran (2007), Strobl et al (2008), Janitza et al (2016), Nembrini et al (2019), Debeer and Strobl (2020), …

Towards evaluating iRF theoretically for interaction discovery consistency

- New LSS model: linear combination of Boolean interactions
- Theoretical tractable version of iRF: LSSFind based on Depth-Weighted Prevalence (DWP) computed from an RF tree ensemble
- Interaction discovery consistency of LSSFind under regularity conditions
- Simulation studies

M. Behr Y. Wang X. Li

arXiv.org > math > arXiv:2102.11800

Search..

Help | A

Mathematics > Statistics Theory

[Submitted on 23 Feb 2021 (v1), last revised 1 Mar 20

Provable Boolean Interaction Recovery from Tree Ensemble obtained via Random Forests

Merle Behr, Yu Wang, Xiao Li, Bin Yu

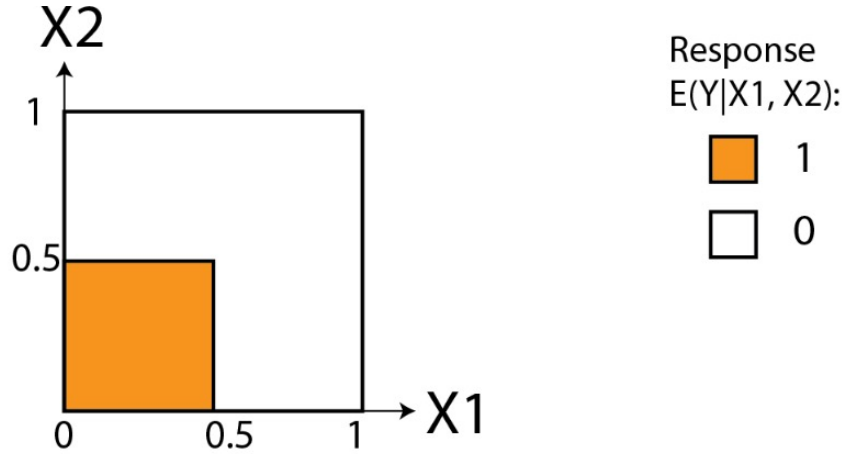
New Benchmark Model for Studying RF or siRF:

Local Spiky Sparse (LSS) model

$$E(Y|X) = \beta_0 + \sum_{k=1}^{K} \beta_k \prod_{j \in S_k} \mathbf{1}(X_j \leq \gamma_j)$$

Possible interpretation: K non-overlapping pathways
Model not identifiable with overlapping interactions

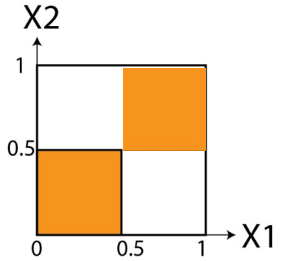
$$E(Y|X_1, X_2) = \mathbf{1}(X_1 \leq 0.5) \cdot \mathbf{1}(X_2 \leq 0.5).$$



Example of a non-identifiable LSS model

One representation of the regression function

$$\mathbf{1}(X_1 < 0.5, X_2 < 0.5) + \mathbf{1}(X_1 > 0.5, X_2 > 0.5)$$

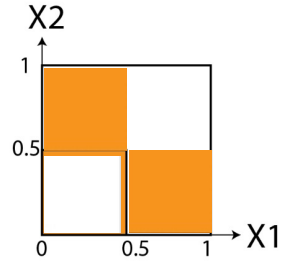


Signed interactions $\{(1,-1),(2,-1)\}$ and $\{(1,+1),(2,+1)\}$

Equivalent representation

$$1 - \mathbf{1}(X_1 < 0.5, X_2 > 0.5) - \mathbf{1}(X_1 > 0.5, X_2 < 0.5)$$

1 -



Signed interactions $\{(1,-1),(2,+1)\}$ and $\{(1,+1),(2,-1)\}$

Non-sign interactions are identifiable

What do we want in a tractable version of iRF?

- Re-weighting is hard to analyze, but its goal is to find stable interactions on the paths from an RF tree ensemble
- Random interaction trees (RIT) is hard to analyze – omit it

Idea:

Hard-thresholding impurity index on a path to bring in stability and no RIT

Stable path set: $\hat{\mathcal{F}}_\epsilon(\mathcal{P}, T, \mathcal{D})$

- Defining stability through thresholding impurity index ϵ
- Given a tree T and a path P, this set gives the indices and corresponding signs of first-appearing stable features with impurity index larger than a threshold ϵ

$$\hat{\mathcal{F}}_\epsilon(\mathcal{P}, T, \mathcal{D}) := \{(k_t, b_t) \mid \ t \text{ is an inner node of } \mathcal{P} \text{ with } \Delta_I^n(t) > \epsilon \text{ and feature } k_t \text{ appears first time on } \mathcal{P}\}$$

Depth-weighted prevalence (DWP) for a given RF tree ensemble: conditioning on data

- Prevalence is a form of stability
- Signed feature set $S^{\pm} \subset [p] \times \{-1, +1\}$

  Example: {(1, -1), (1, +1), (3, +1)}

- For a given $S^{\pm}$,

  DWP is the depth-weighted probability of the stable set $\hat{\mathcal{F}}_{\epsilon}(\mathcal{P}, T, \mathcal{D})$

$$\mathrm{DWP}_{\epsilon}(S^{\pm}) = P_{(\mathcal{P},T)}(S^{\pm} \subset \hat{\mathcal{F}}_{\epsilon} \mid \mathcal{D})$$

Randomness in T comes from RF tree construction;
Given a tree, randomness in P uses 2^{-d} probability for any path with depth d.

Main results: Defining properties of DWP (stability) under LSS

1. *(General upper bound)* $\mathrm{DWP}_\epsilon(S^\pm) \leq 2^{-|S^\pm|}$

Major assumptions:

- LSS with 1-subgaussian additive noise, non-overlap basic Boolean sets, independent features; mtry is of order p,….
- Lower bounds on LSS parameters to stay away from zero with a gap

2. For true (union) interactions, the upper bound is achieved with equality

3. For wrong interactions, DWP has a gap from the upper bound

- *(Interaction lower bound) when* $S^\pm$ *is a union signed interaction as in Definition 1, we have*

$$\mathrm{DWP}_\epsilon(S^\pm) \geq 2^{-|S^\pm|} - b(\epsilon) - r_n(\mathcal{D}, \epsilon);$$

- *(Non-interaction upper bound) w hen* $S^\pm$ *is not a union signed interaction, then,*

$$\mathrm{DWP}_\epsilon(S^\pm) \leq 2^{-|S^\pm|}\left(1 - \frac{C_m^s}{2}\right) + r_n(\mathcal{D}, \epsilon),$$

with

$$r_n(\mathcal{D}, \epsilon) \xrightarrow{p} 0 \quad \text{as } n \to \infty,$$

LSSFind utilizes the defining properties

LSSFind returns signed feature sets that achieve the bound.

Algorithm 1: LSSFind(m_{try},ϵ, η, $s_{\max}$)

Input: Dataset $\mathcal{D}$, RF hyperparameter m_{try}, impurity threshold $\epsilon > 0$, prevalence threshold $\eta > 0$, and maximum interaction size $s_{\max} \in \mathbb{N}$.

Output: A collection of sets of signed features.

Train an RF using dataset $\mathcal{D}$ with parameter m_{try}.;

return $\{S^{\pm} \subset [p] \times \{-1, +1\}$ such that $|S^{\pm}| \leq s_{\max}$ and $2^{|S^{\pm}|} \cdot \text{DWP}_{\epsilon}(S^{\pm}) \geq 1 - \eta\}$.

It is really intriguing that LSSFind achieves "model selection" consistency without estimating any model parameters in LSS.

Note that iRF doesn't estimate them either.

$$E(Y|X) = \beta_0 + \sum_{k=1}^{K} \beta_k \prod_{j \in S_k} \mathbf{1}(X_j \leq \gamma_j)$$

Proof ideas

General upper bound is a counting problem (it holds for any tree)

For the lower bound and achievability on signal interactions in LSS

- Prove the results for the population case: very delicate even with smart notations

Feature space is divided into regions by a tree. Based on true interactions in LSS, impurity index decrease behaves differently in these different regions. Stable paths with non-zero impurity decreases correspond to true interactions.

- Use uniform convergence results for VC classes for finite sample case

Insights from theoretical analysis

WLOG, assume features are uniform on [0,1]

- Higher-order interactions are difficult because the number of samples fall into an order L interaction region: O(2^{-L})
- Highest possible stability in DWP: O(2^{-L}) for L-order interactions
- Hard thresholding on impurity index in LSSFind does not well allow “weak” interactions to show, indicating advantage of iRF that uses a data-driven soft-threashoding
- mtry should not be too large or too small, backing up choice of mtry in RF

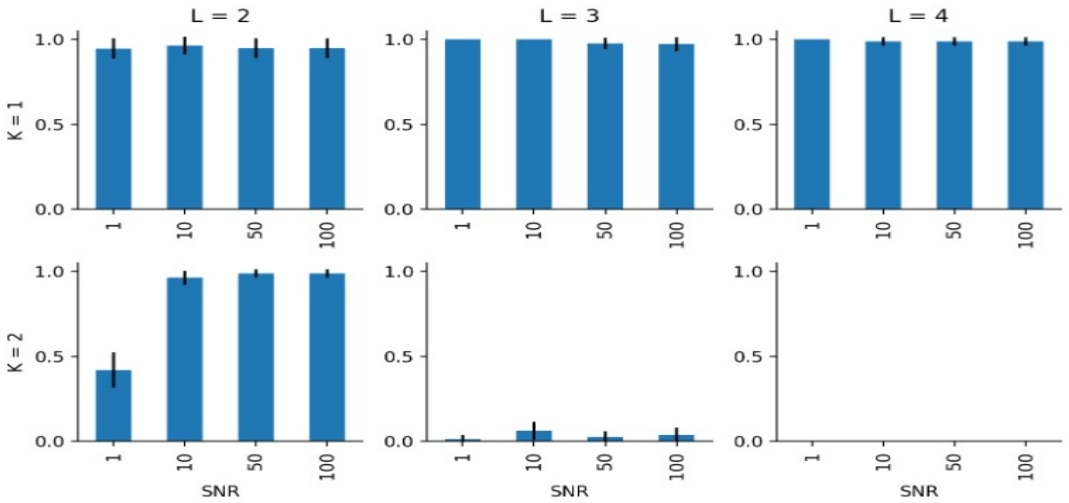
Simulation studies: exact discovery (Jaccard distance)

(p=20, n=1,000; K= # terms in LSS, L= order of interaction)

- Correct LSS model

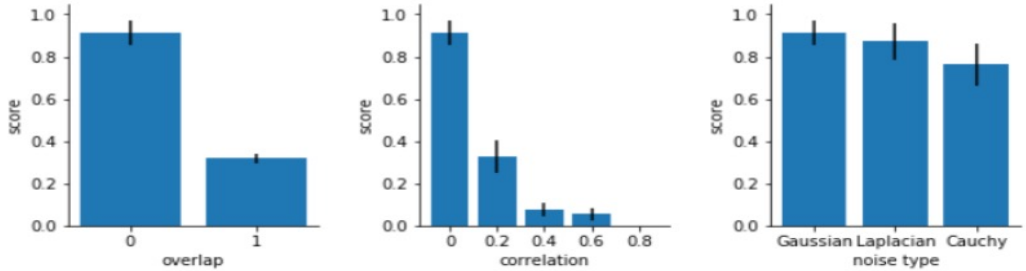
1-term in LSS

2-terms in LSS



More difficult with more terms and higher-order interactions

- Misspecified (L=2, K=2)



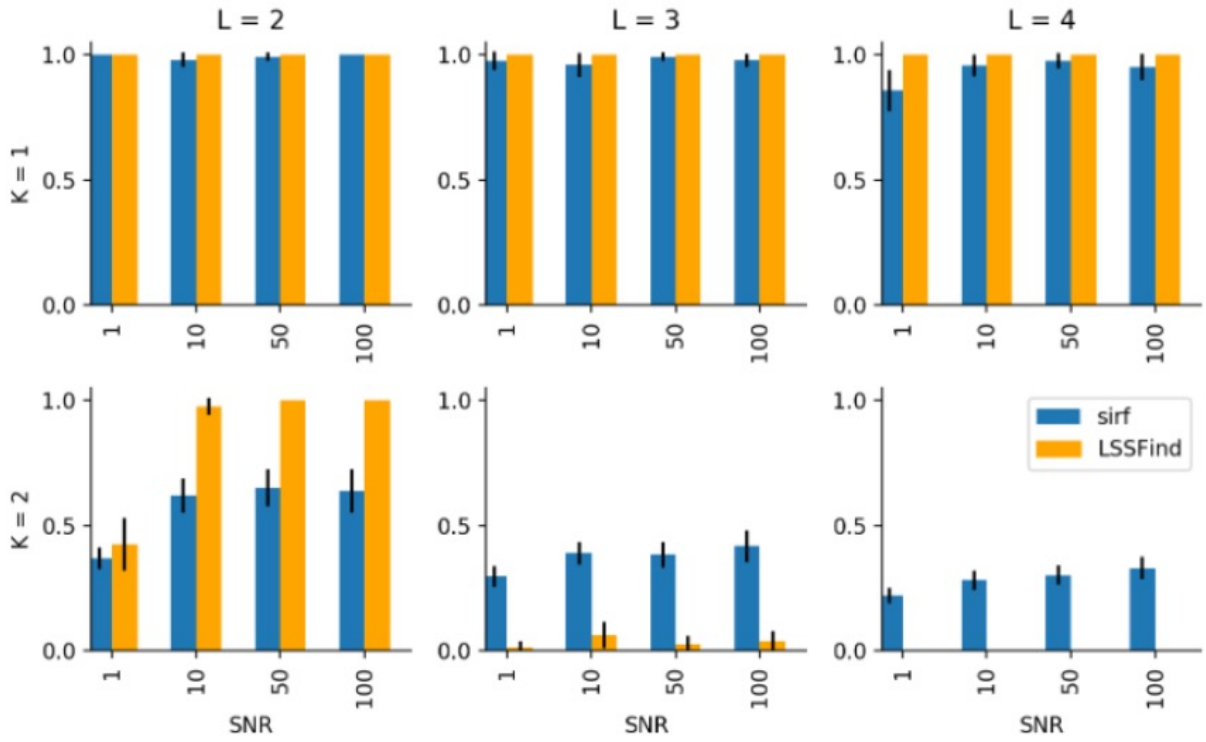
Overlapping Boolean terms and dependent features are problematic

Not so much noise distribution

Comparing LSSFind and siRF

With a practical metric (not the impractical exact discovery) that looks at the whole collection of individual features (not for each term)

siRF or (signed) iRF is very competitive, esp. for more terms and higher-order interactions



Summary

- iRF (siRF) is a practical algorithm to find Boolean interactions with empirical successes in some genomics problems – case study of PCS
- Relevant theoretical analysis is an integral part of evaluating siRF
- A new relevant Boolean model: LSS
- LSSFind based on DWP is shown to be consistent under LSS
- Simulation studies verify theoretical results and show siRF is more robust and better for higher-order interactions than LSSFind
- The simulation set-up most relevant to practice is for mis-specified models and relaxed metric.

Thank you!

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

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

3. M. Behr, Y. Wang. X. Li, B. Yu (2021). **Provable Boolean Interaction Recovery from Tree Ensemble obtained via Random Forests.** https://arxiv.org/abs/2102.11800

PCS Software Projects

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!

Book by Yu and Barter with MIT Press
Free on-line interactive copy (plan: 2022 spring)

Veridical Data Science: A Book

Bin Yu[1,2] and Rebecca Barter[1]
[1]Department of Statistics, UC Berkeley
[2]Department of Electrical Engineering and Computer Science, UC Berkeley


Berkeley
UNIVERSITY OF CALIFORNIA

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 processing
Data cleaning
Exploratory Data Analysis
Data merging

Algorithmic
Dimension reduction
Clustering
Least Squares & ML
Regularization

Stability-based inference
Inference
Causal Inference
Perturbation Intervals
Trustworthiness Statements

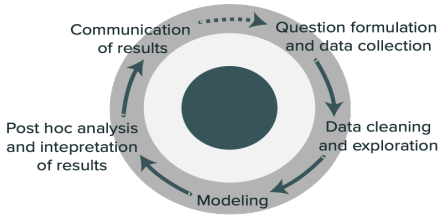
Communication

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

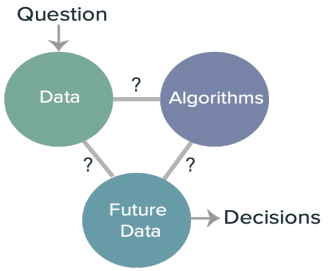
The DS Lifecycle



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.

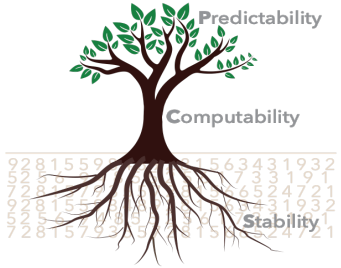
Blending together concepts from statistics, computer science and domain knowledge, the data science life cycle is an iterative process that involves human analysts learning from data and refining their project-specific questions and analytic approach as they learn.

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.

PCS framework



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 efficiently 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.

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: @rlbarter