Veridical Data Science

Bin Yu

Statistics and EECS, UC Berkeley

Breiman Lecture, NeurIPS

Vancouver, Dec. 10, 2019

ve·rid·i·cal

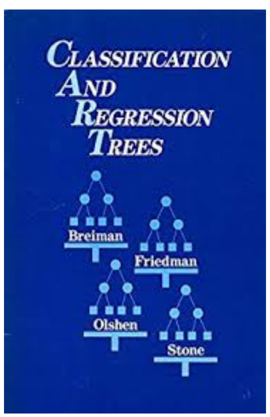
/vəˈridək(ə)l/

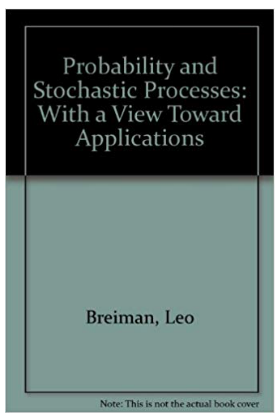
adjective FORMAL

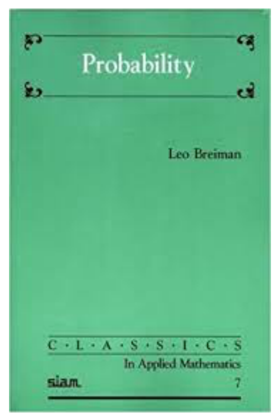
truthful.

- coinciding with reality.
  "such memories are not necessarily veridical"

Leo Breiman (1928-2005):
a data scientist and a modern day polymath







2001

Image credit: en.Wikipedia.org

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:

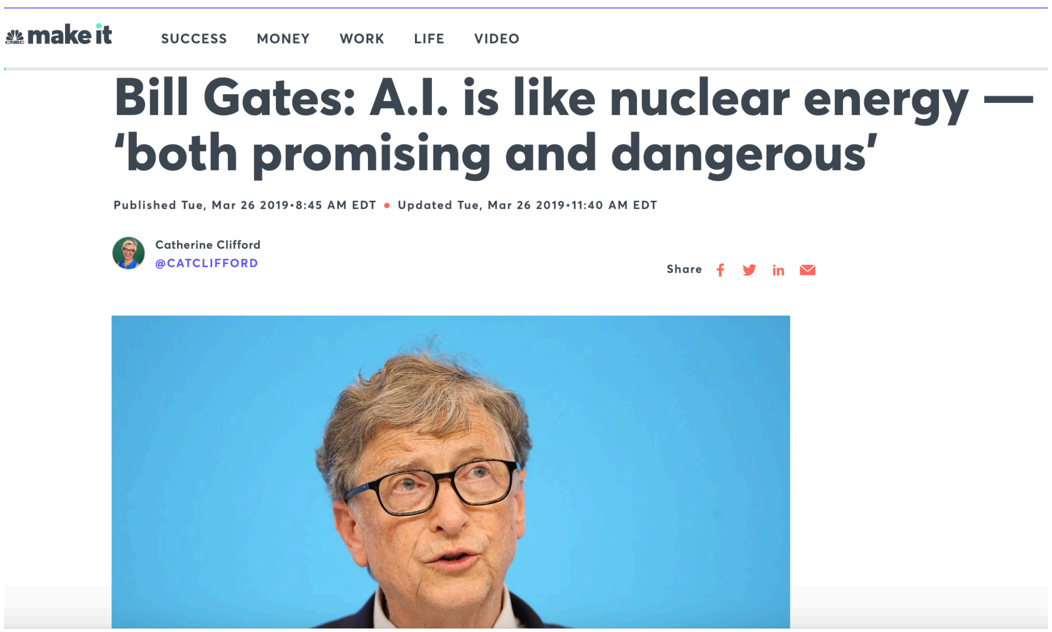
Machine Learning
October 2001, Volume 45, Issue 1, pp 5–32 | Cite as

Random Forests

Authors | Authors and affiliations

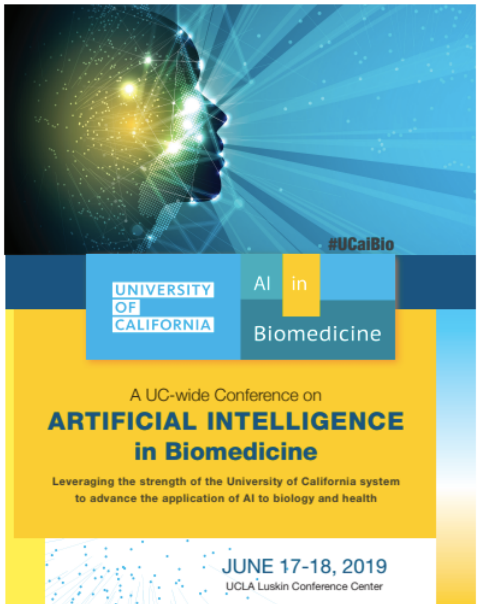
Leo Breiman

2019 AI is part of modern life

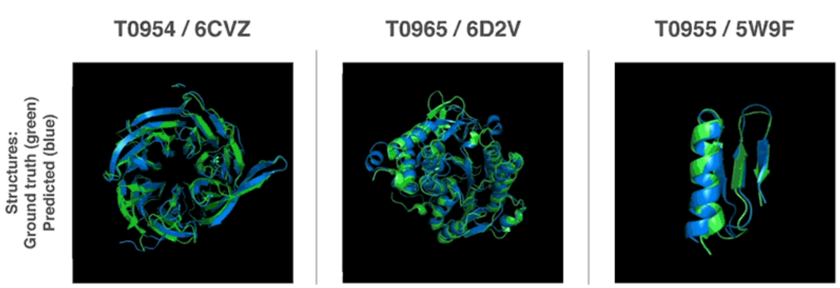


Alexa, Siri, …
Wearable health devices
Streaming videos, on-line gaming, …
On-line news
Self-driving cars
Election campaigns
Precision medicine
Biology
Neuroscience
Cosmology
Material science
Chemistry
Law
Political science
Economics
Sociology
…

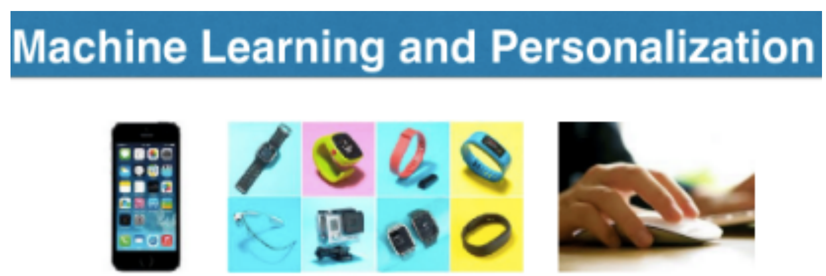
Biomedical data problems are pressing



medium.com



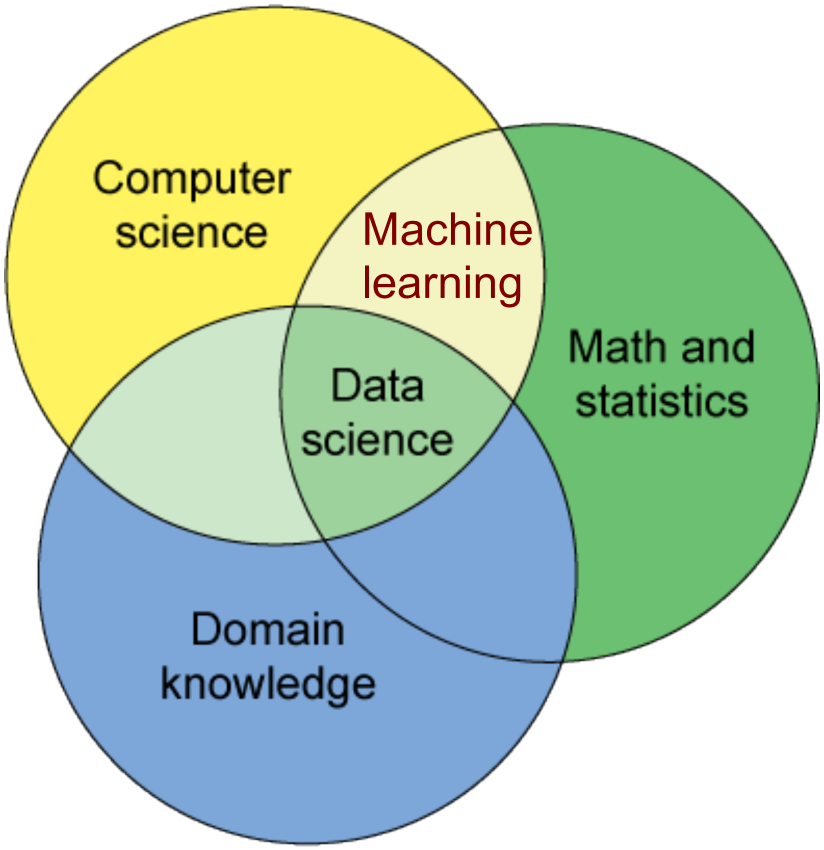
https://deepmind.com/blog/alphafold/



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

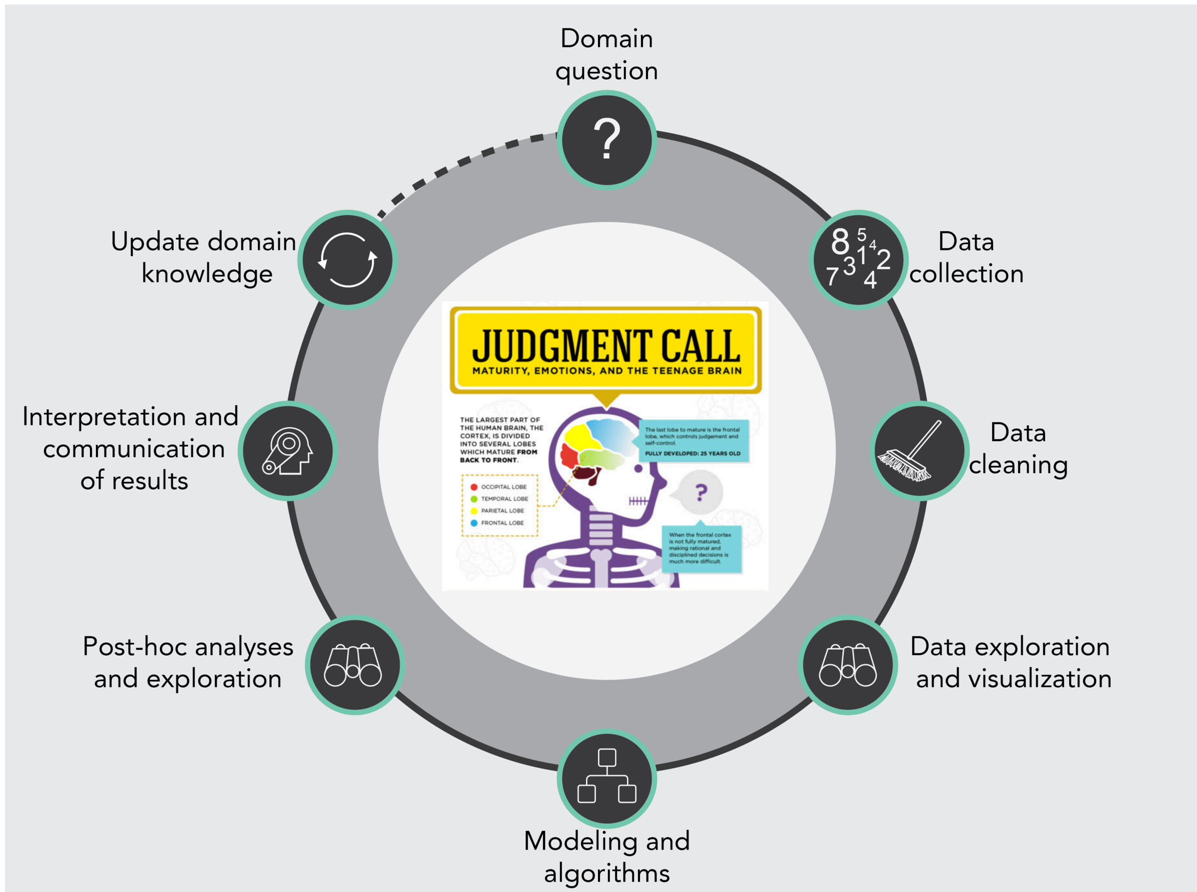


Image credits: R. Barter and toronto4kids.com

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

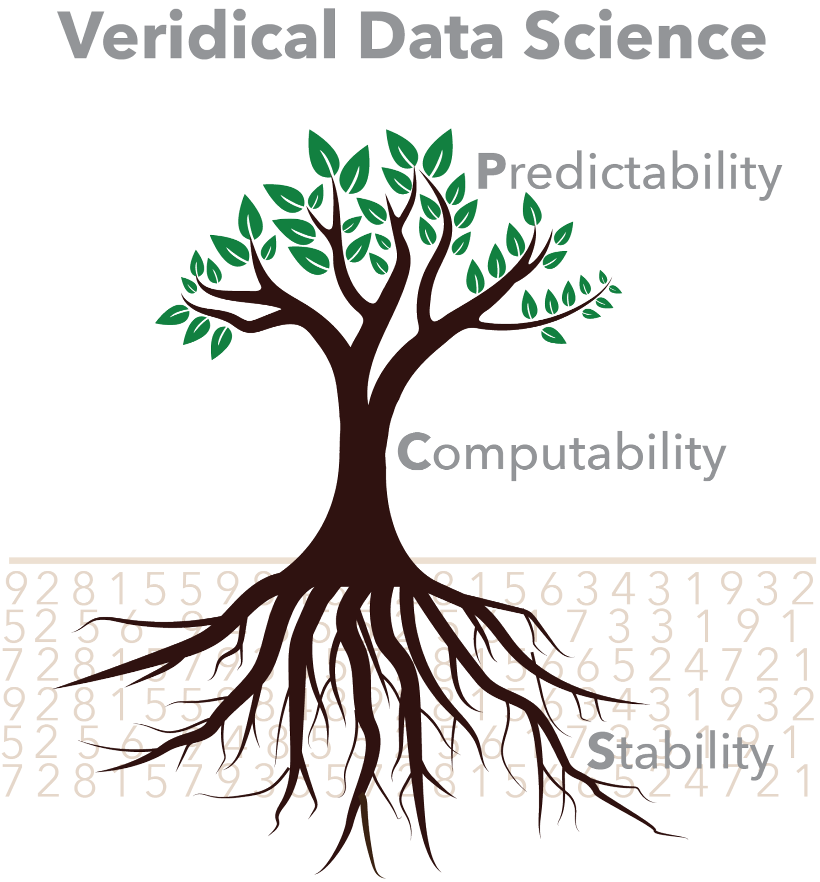
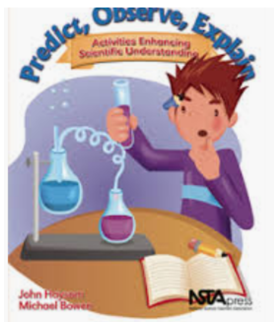
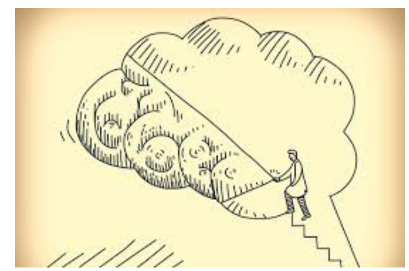


Image credit: R. Barter

PCS connects science with engineering

- Predictability and stability embed two scientific principles: prediction and replication



- Computability is a necessity and includes data-inspired simulations

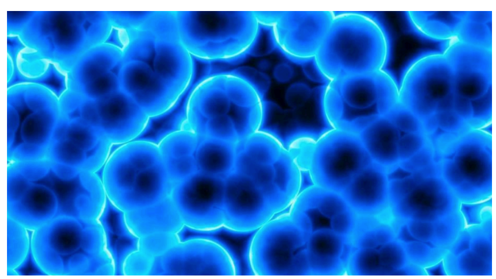


Image credits: nstat.org, hub.jhu.edu, vox.com, Andras Libal

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**.

Image credit: designnews.com

Stability tests DSLC by "shaking" every part

DSLC

Shakes come from human decisions

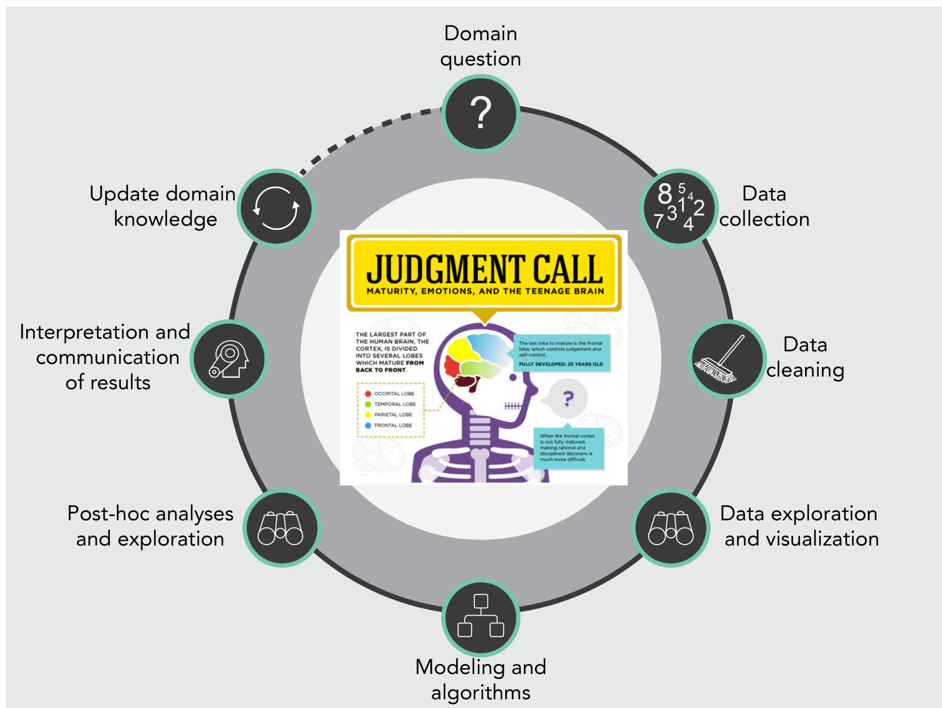
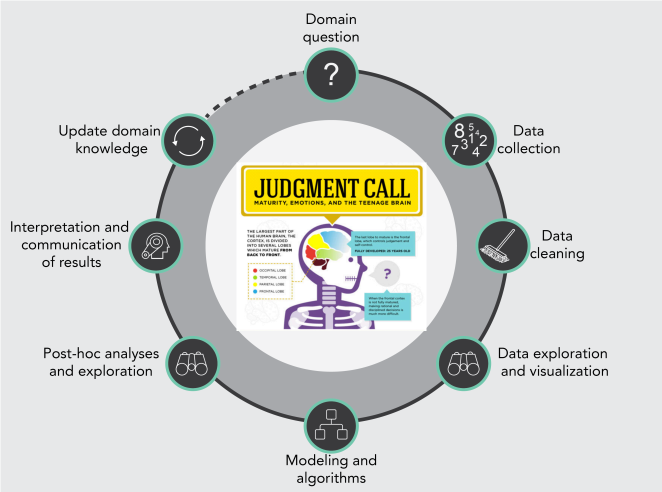


Image credits: R. Barter and toronto4kids.com

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**

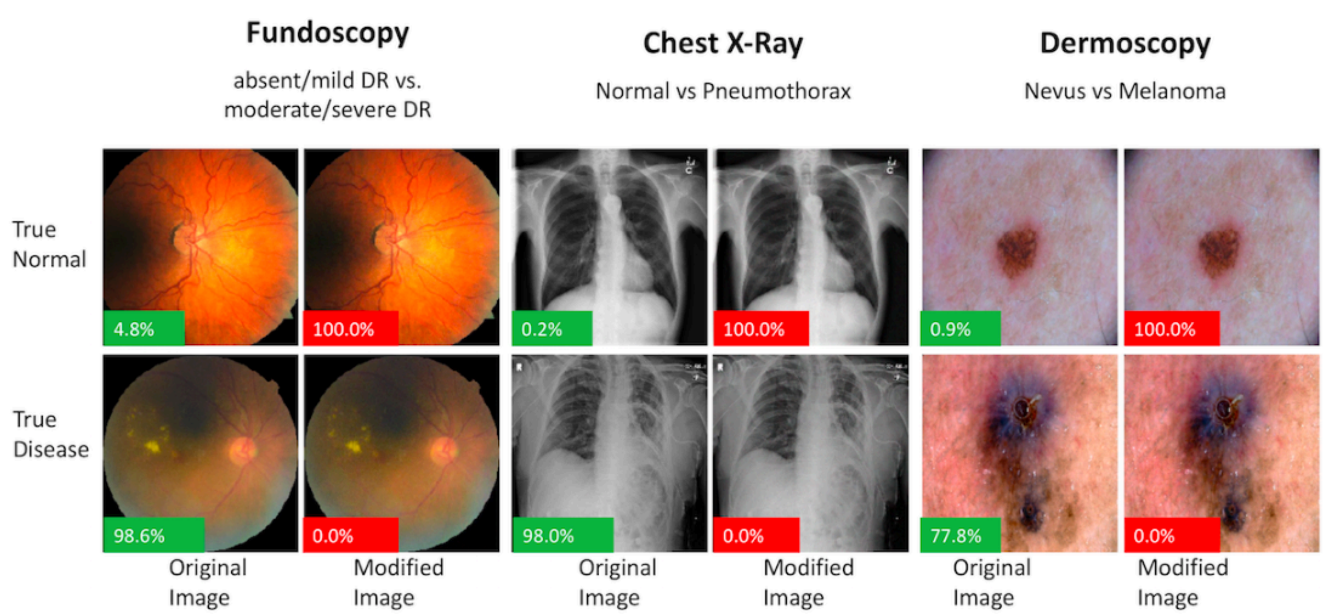


Image credits:groundai.com

Data perturbations (new)

- **Data pre-processing (cleaning) matters**

THE NEW YORKER

THE REINHART AND ROGOFF CONTROVERSY: A SUMMING UP

By John Cassidy April 26, 2013

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

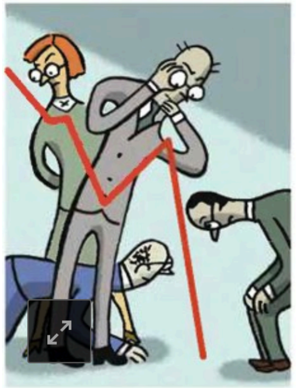
By Carmen M. Reinhart and Kenneth S. Rogoff*

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

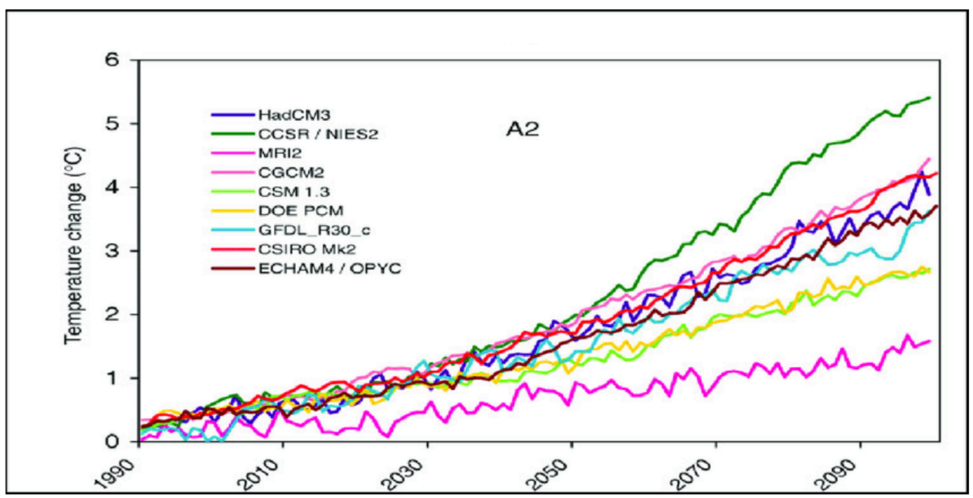

H. Larochelle


A. Beygelzimer


F. d'Alché-Buc


E. Fox

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)

Global mean-temp change

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

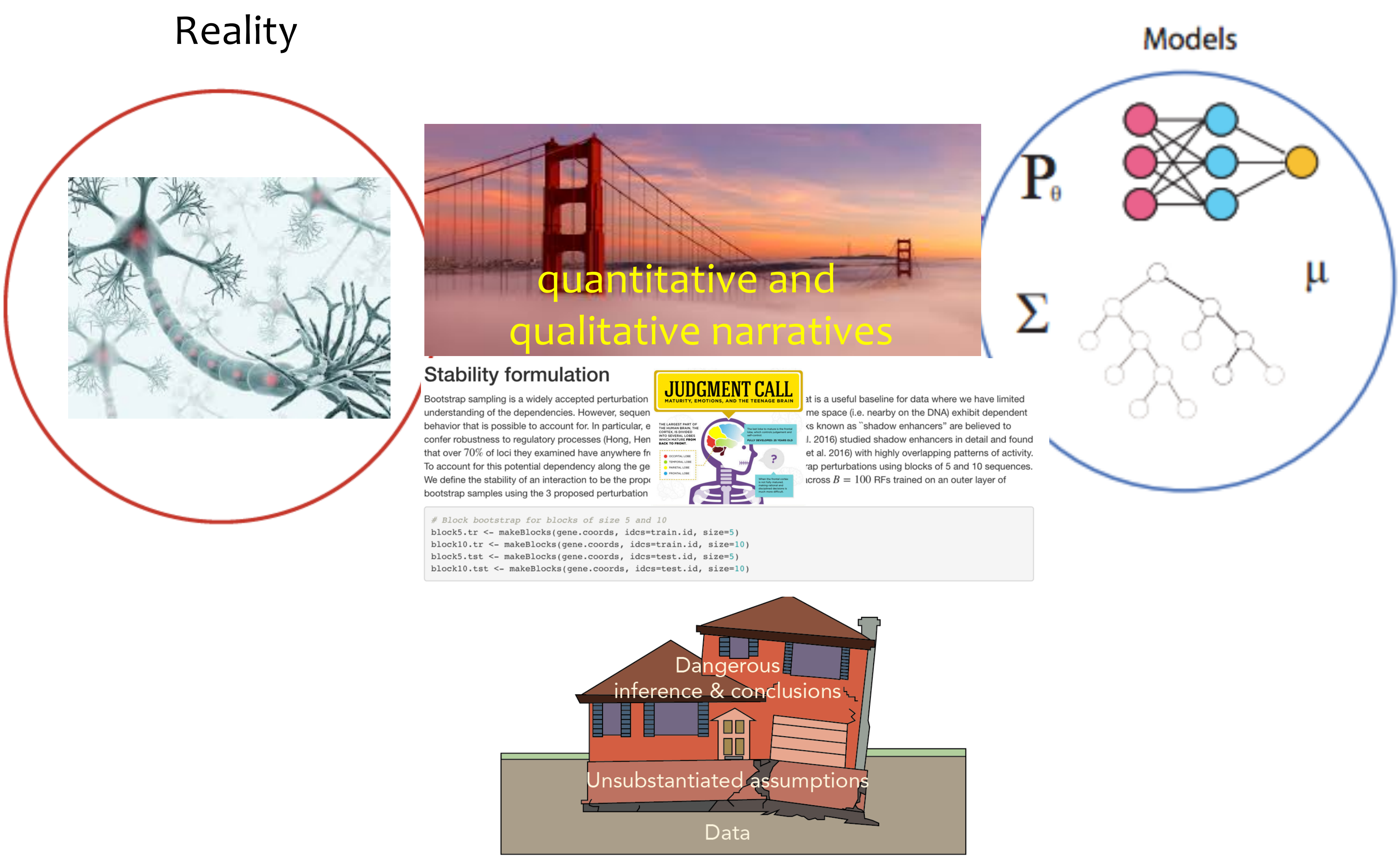


Image credit: Rebecca Barter

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

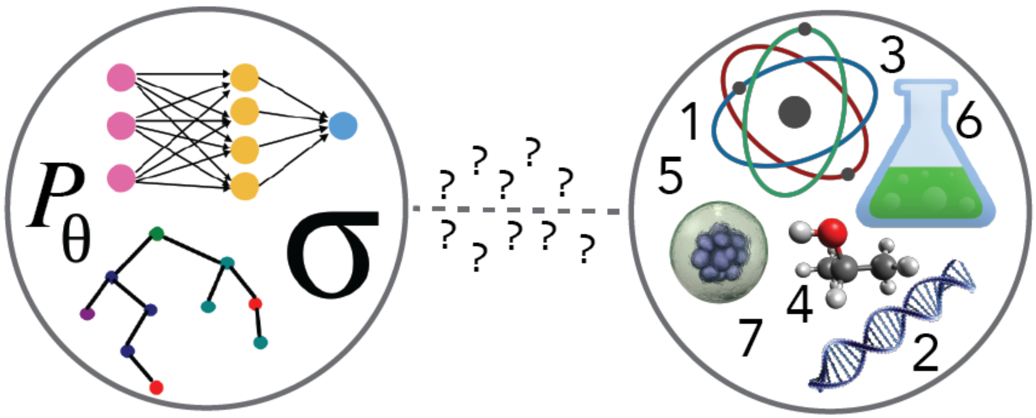
Expanding statistical inference under PCS

- 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

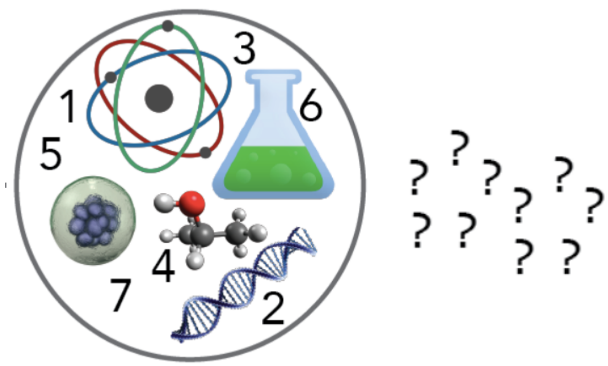
“It is not p-value’s fault”



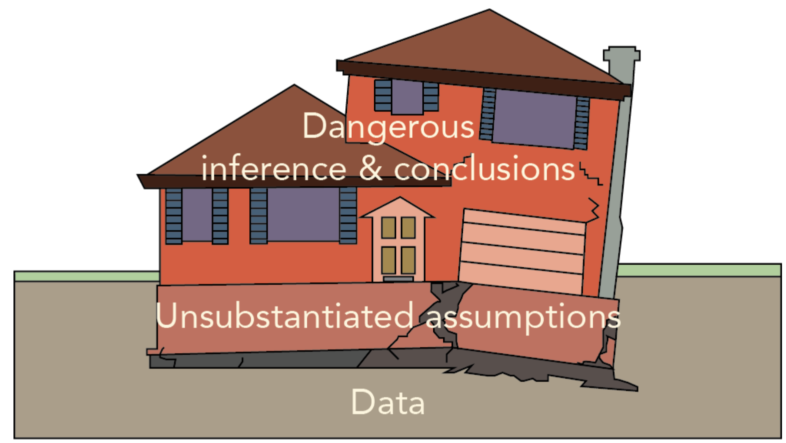
“The p-value is a very valuable tool, but when possible it should be complemented – not replaced - by confidence Intervals and effect size estimates” – Yoav Benjamini

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

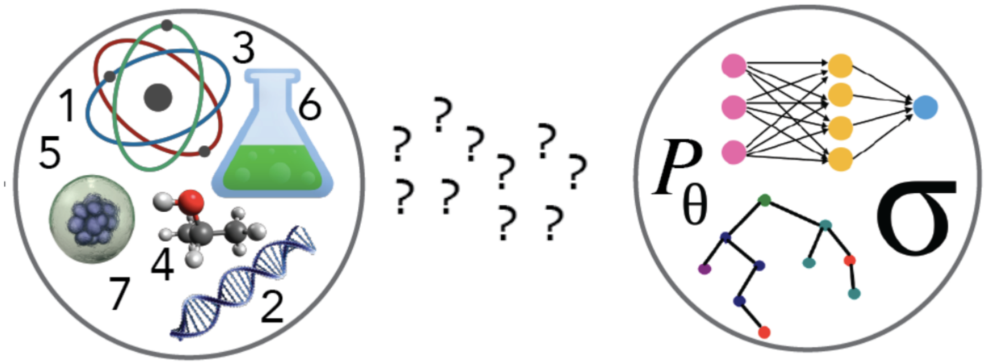


P_θ



- 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)

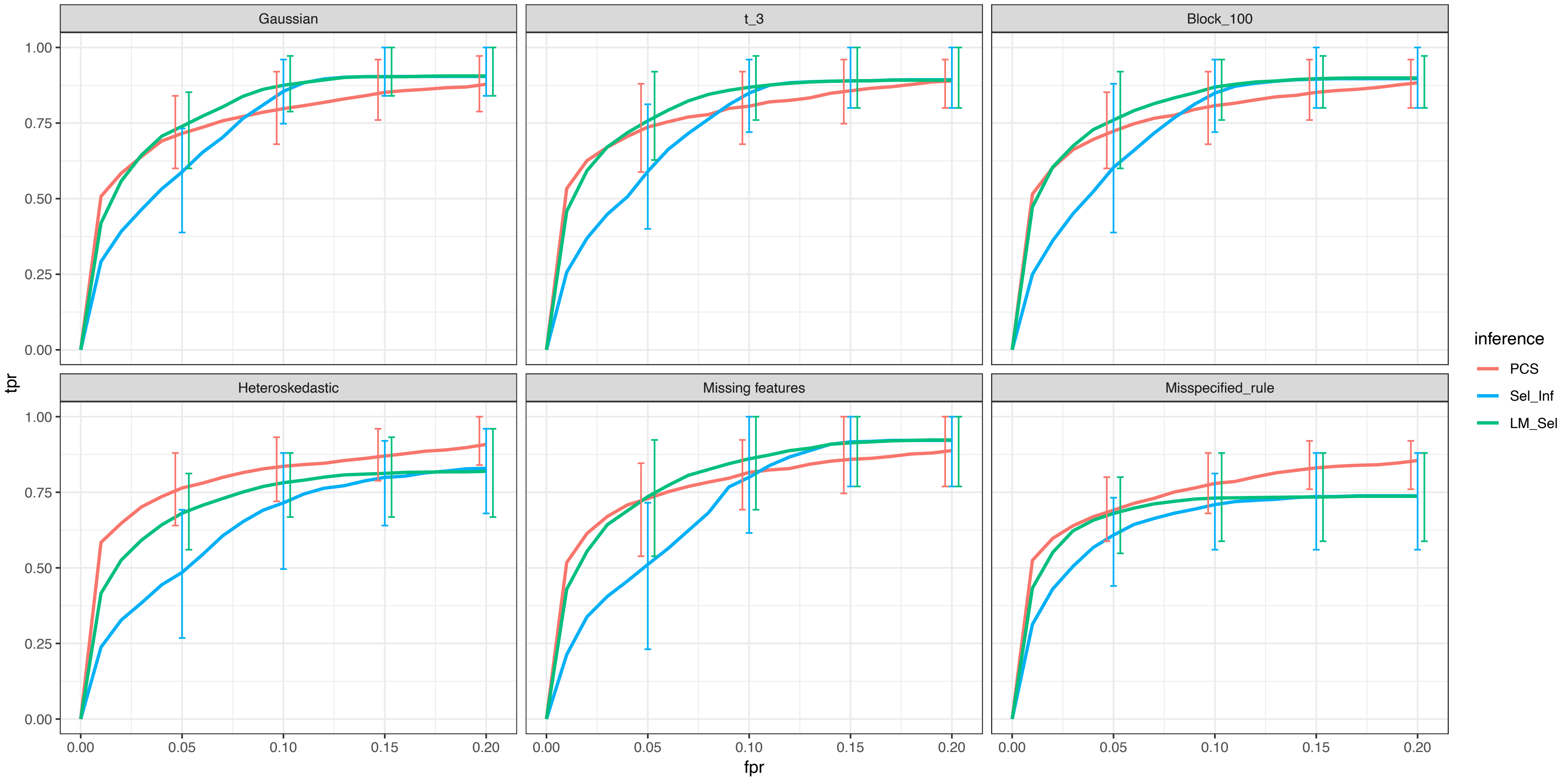
3. **Target value perturbation distribution:** Evaluate the target of interest across "appropriate" data and model perturbations

4. **Perturbation interval reporting:** Summarize the target value perturbation distribution.

Feature importance study: PCS performs well

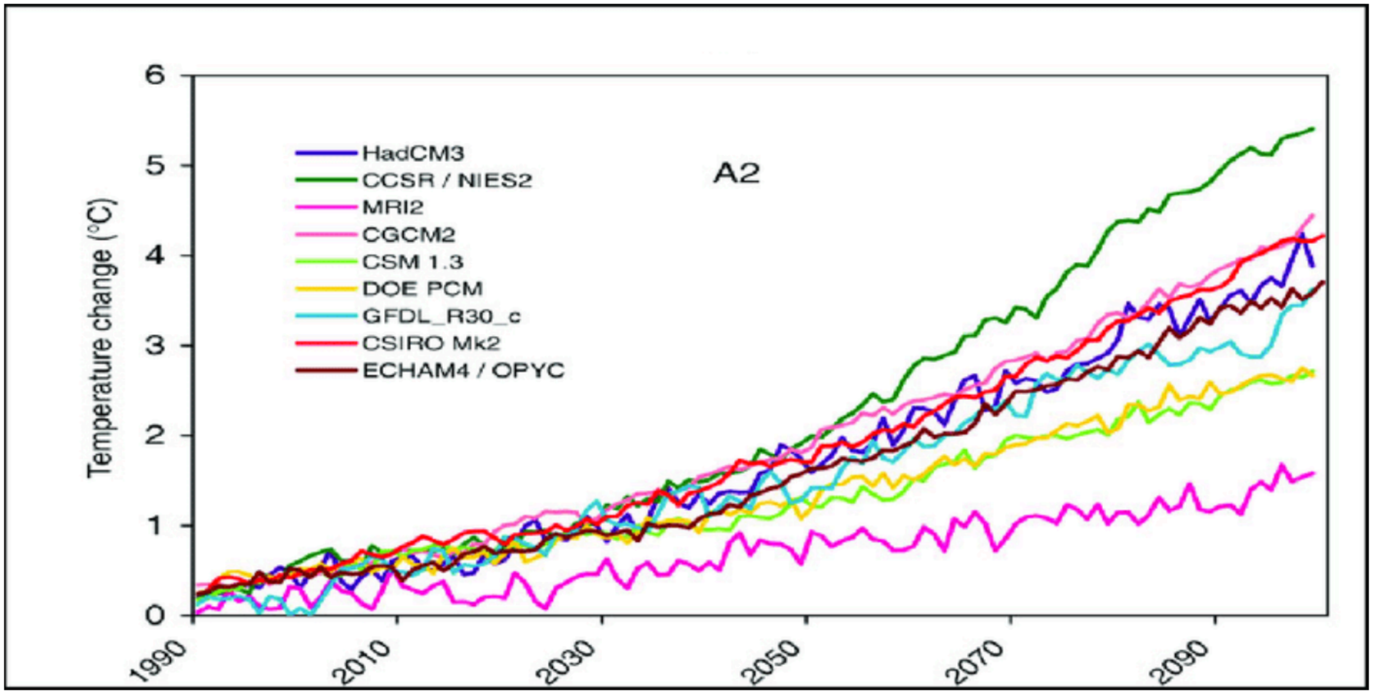
simulation results for lasso feature selection in linear model *n=1000, p=630*

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



Global mean-temp change

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

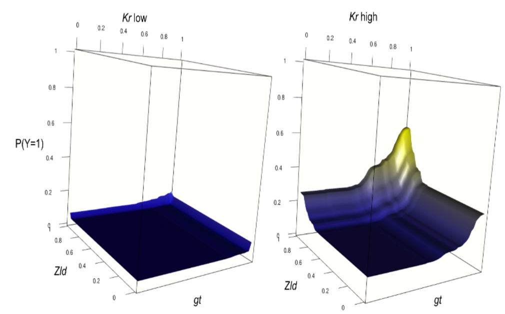
Making Random Forests interpretable

by adding (more) stability

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

Sumanta Basu[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



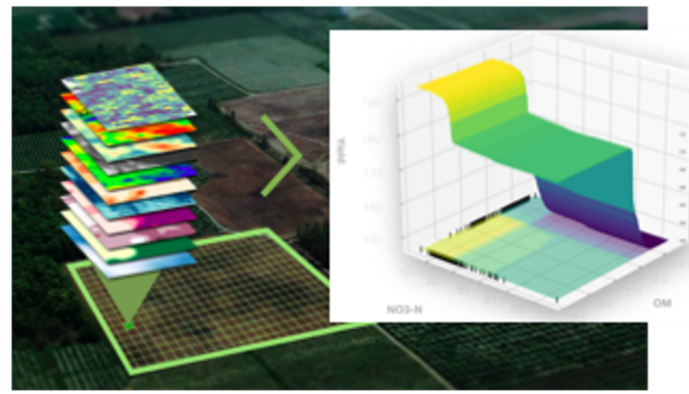
S. Basu K. Kumbier B. Brown

Culmination of 3+ years of work

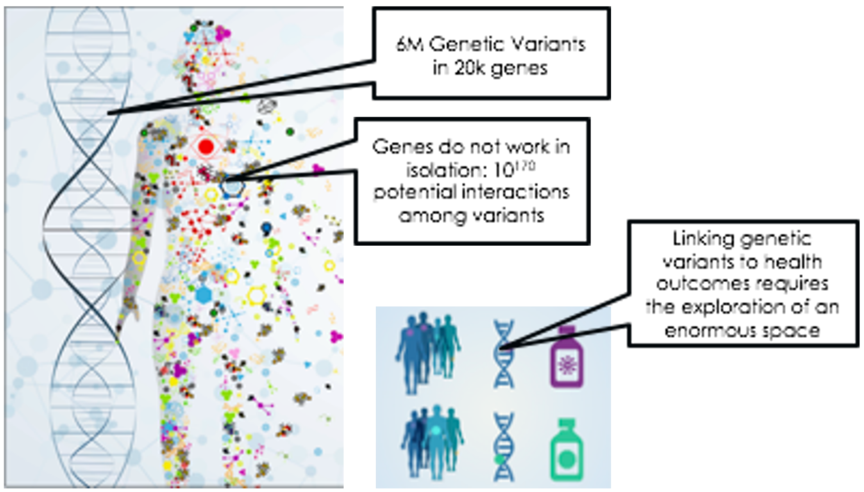
Pattern Recognition *vs.* Pattern Discovery

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

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



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.

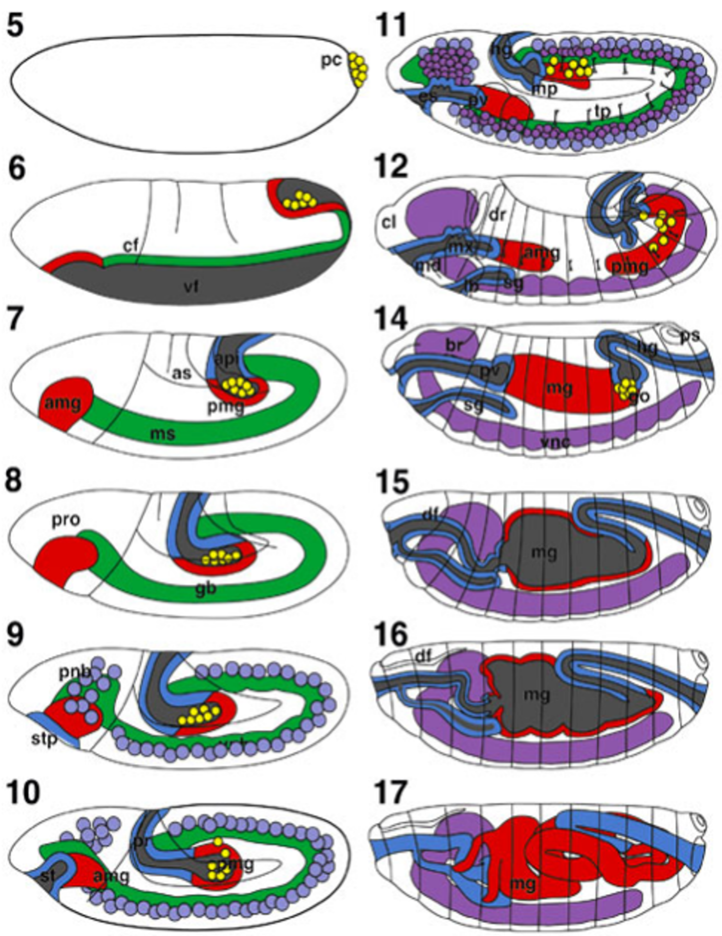
SOP is to leverage forward procedures:

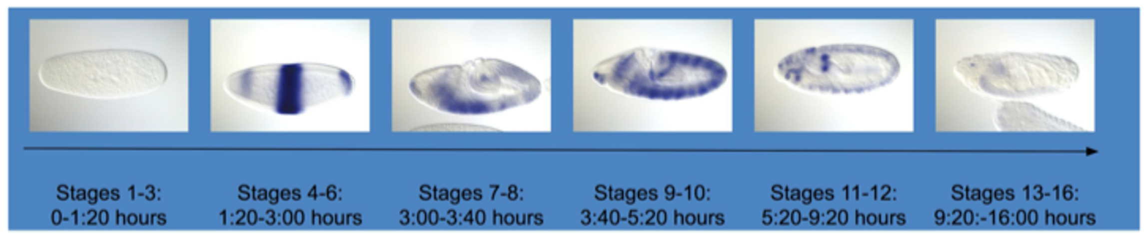
$$y \approx \sum_{j \neq i} \alpha_j x_j + \sum_{k,l \neq i} \beta_{k,l} x_k x_l + \cdots$$

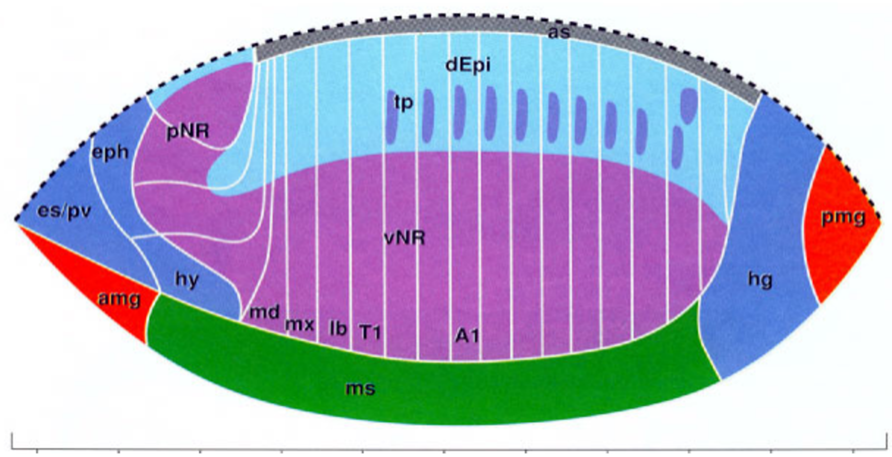
Polynomial interactions do not work well in genomics

Embryonic development in *Drosophila melanogaster*

Overview of the Stages of Development

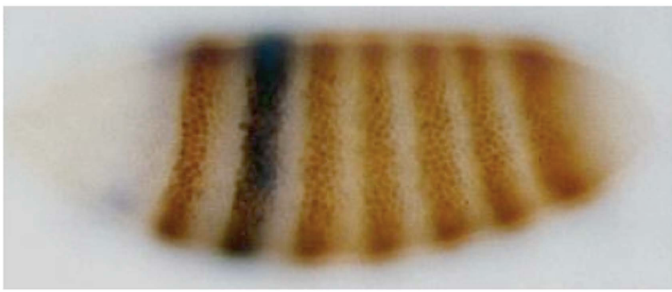


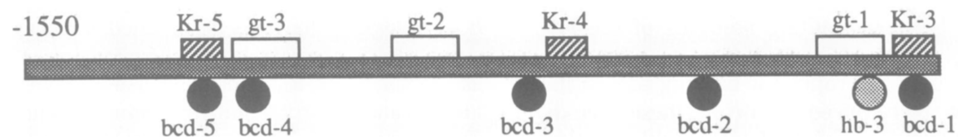




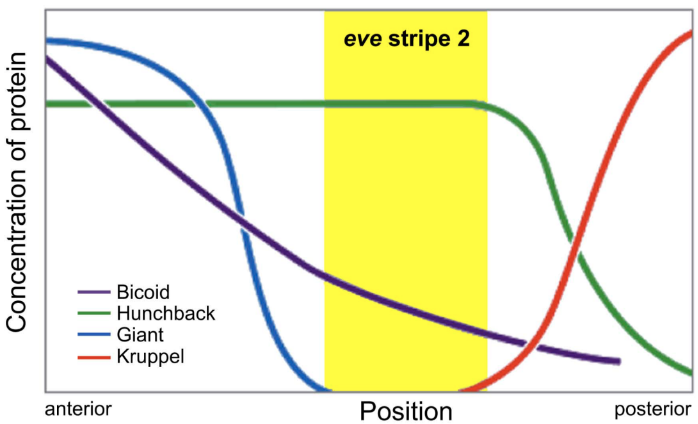
Order-4 interaction regulate *eve* stripe 2

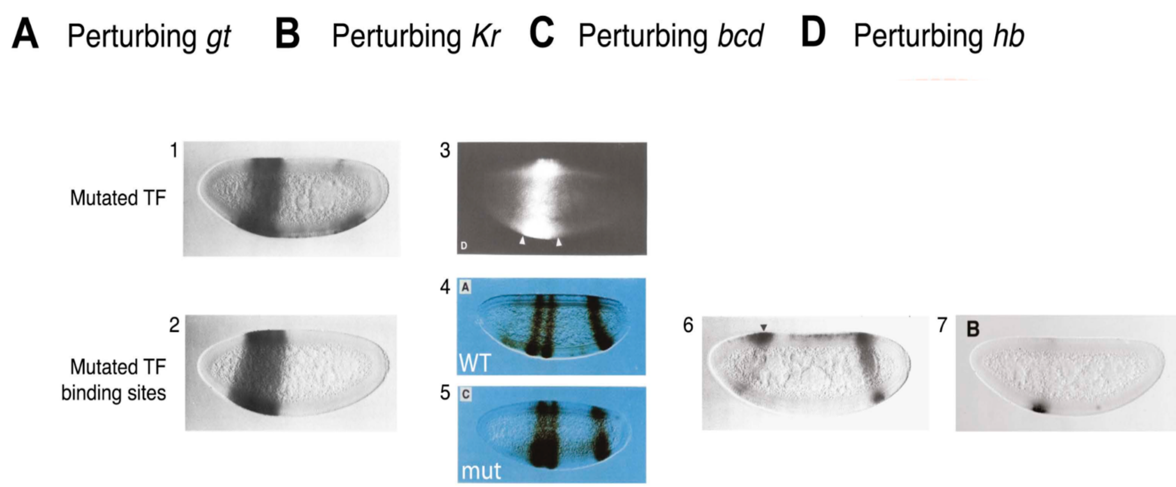
A





B

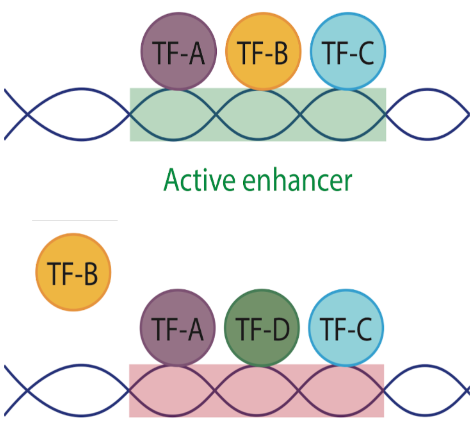




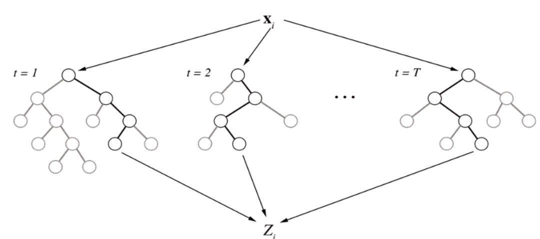
Goto et al. (1989), Harding et al. (1989), Small et al. (1992),
Isley et al. (2013), Levine et al. (2013)

Regulatory interactions through predictability and stability or PCS

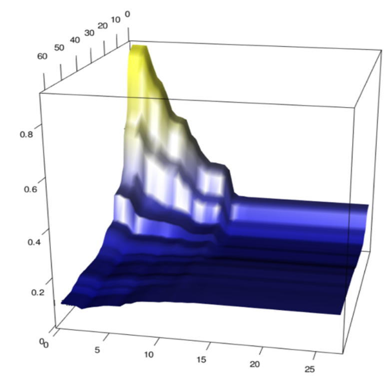
Natural phenomenon



Prediction



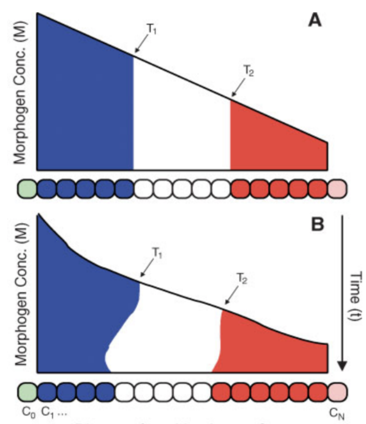
Interpretation



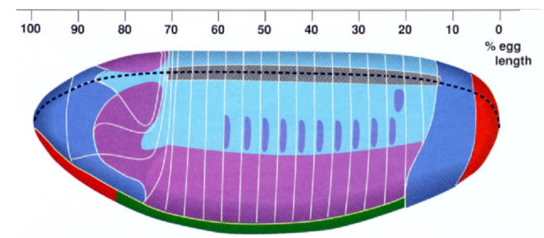
Transcription is initiated when activating transcription factors reach sufficient DNA occupancy

Capturing the form of genomic interactions

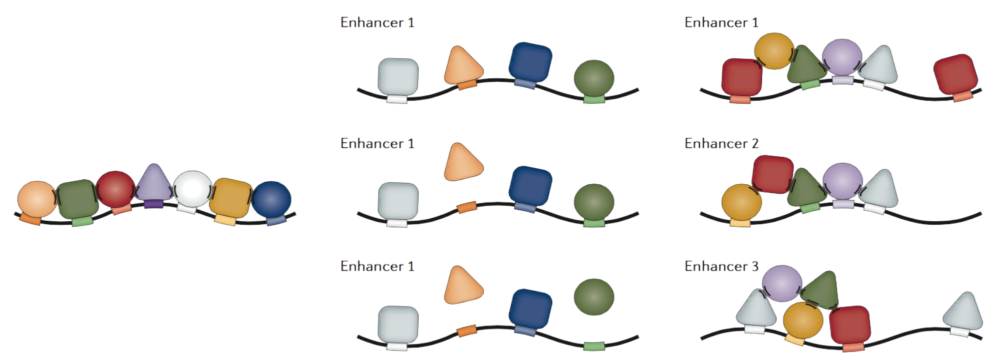
- Interactions are high-order and combinatorial in nature
- Interactions can vary across space and time as biomolecules carry out different roles in varied contexts
- **Interactions** exhibit **thresholding behavior,** requiring sufficient levels of constitutive elements before activating



(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

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

Order- s interaction,

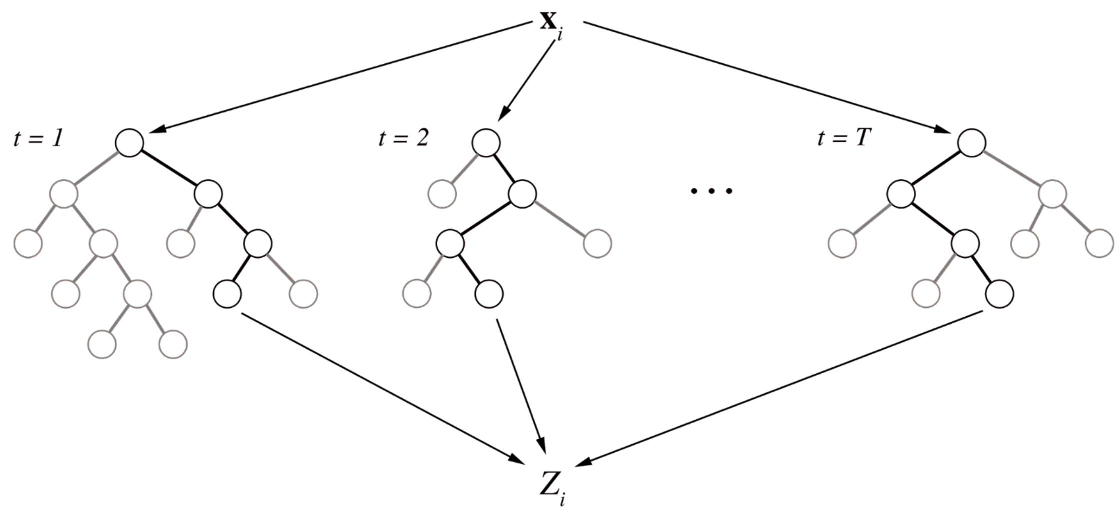
$S \subseteq \{1, \ldots, p\}, |S| = s$

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 `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

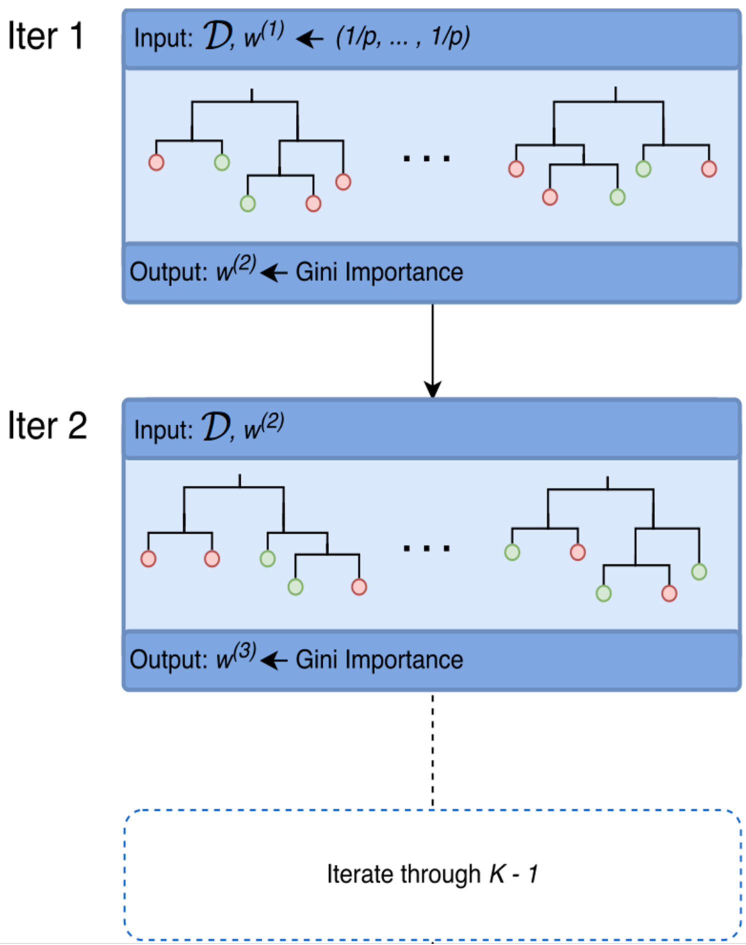
2. Random interaction trees to find intersections of paths

3. Outer-loop bagging assesses stability

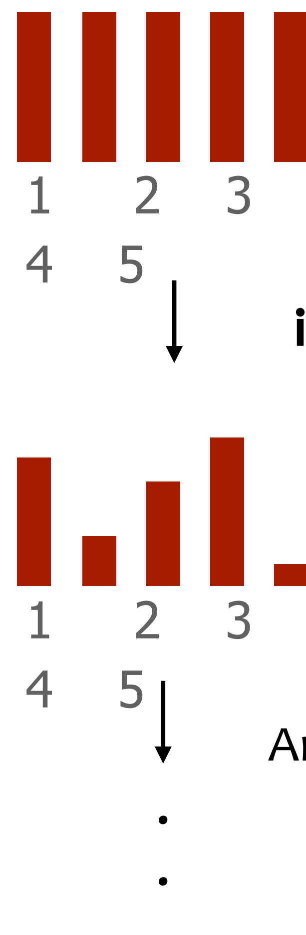
Similar computational and memory costs as RF

Iteratively re-weighted RF stabilize decision paths

Iteratively re-weighted
Random Forests



Feature weights

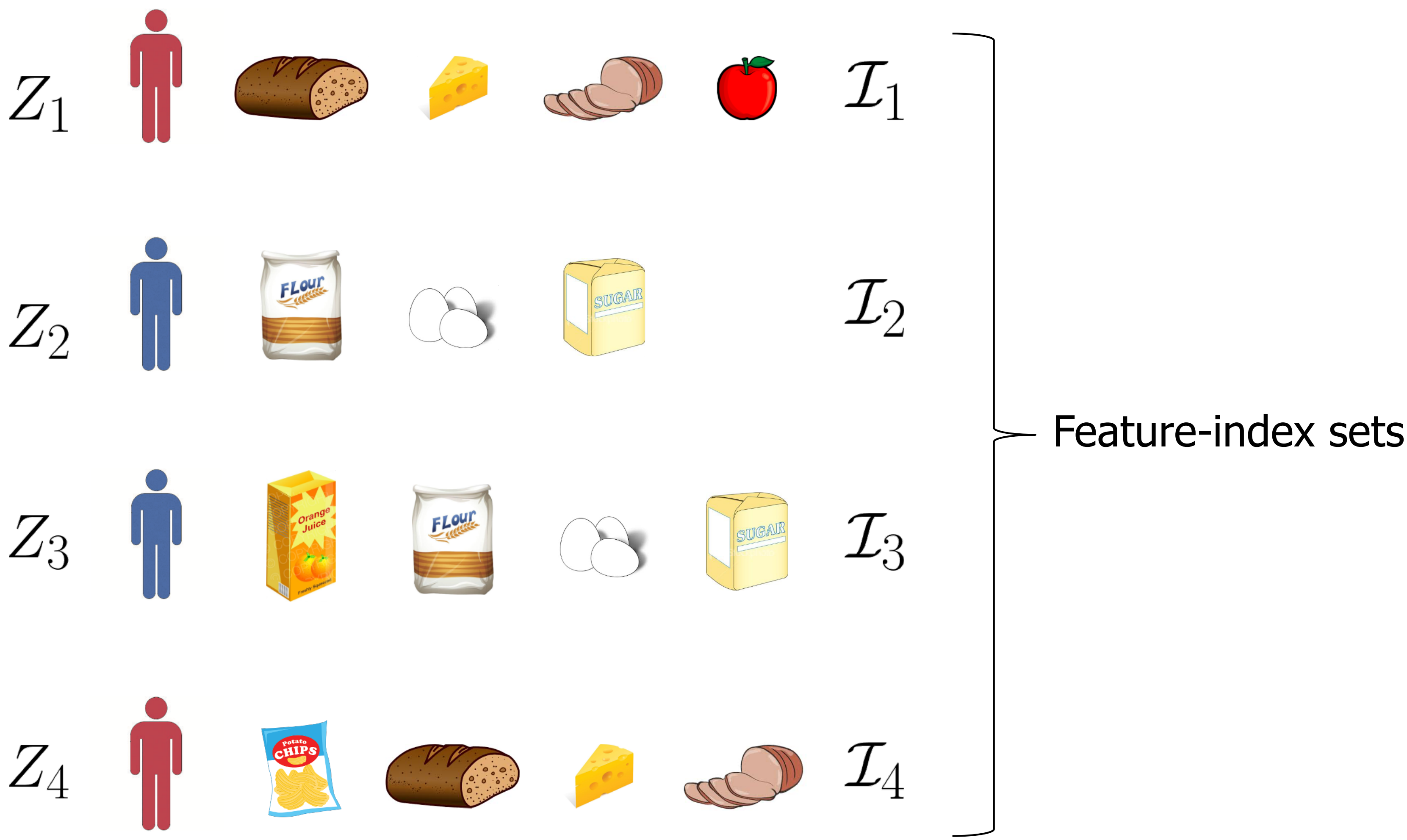


importance index

Re-weighting
Amaratunga et al. (2014)

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

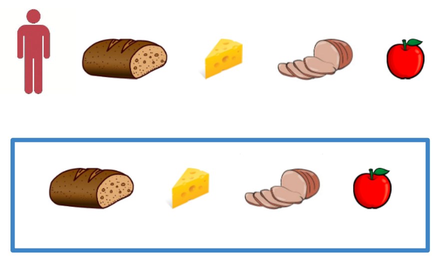
Digression: Interactions in market baskets



Random Intersection Trees (RIT)

Shah and Meinshausen (2014): fast computation uses sparsity

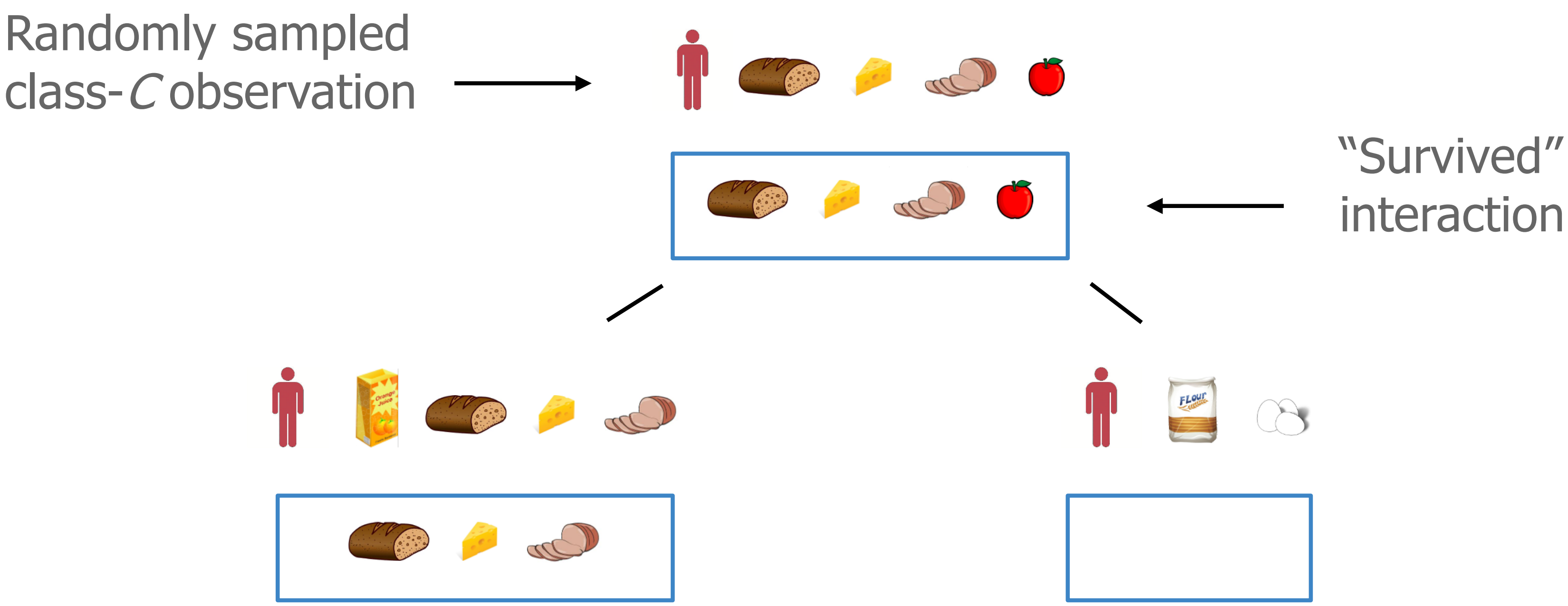
Randomly sampled class-*C* observation



“Survived” interaction

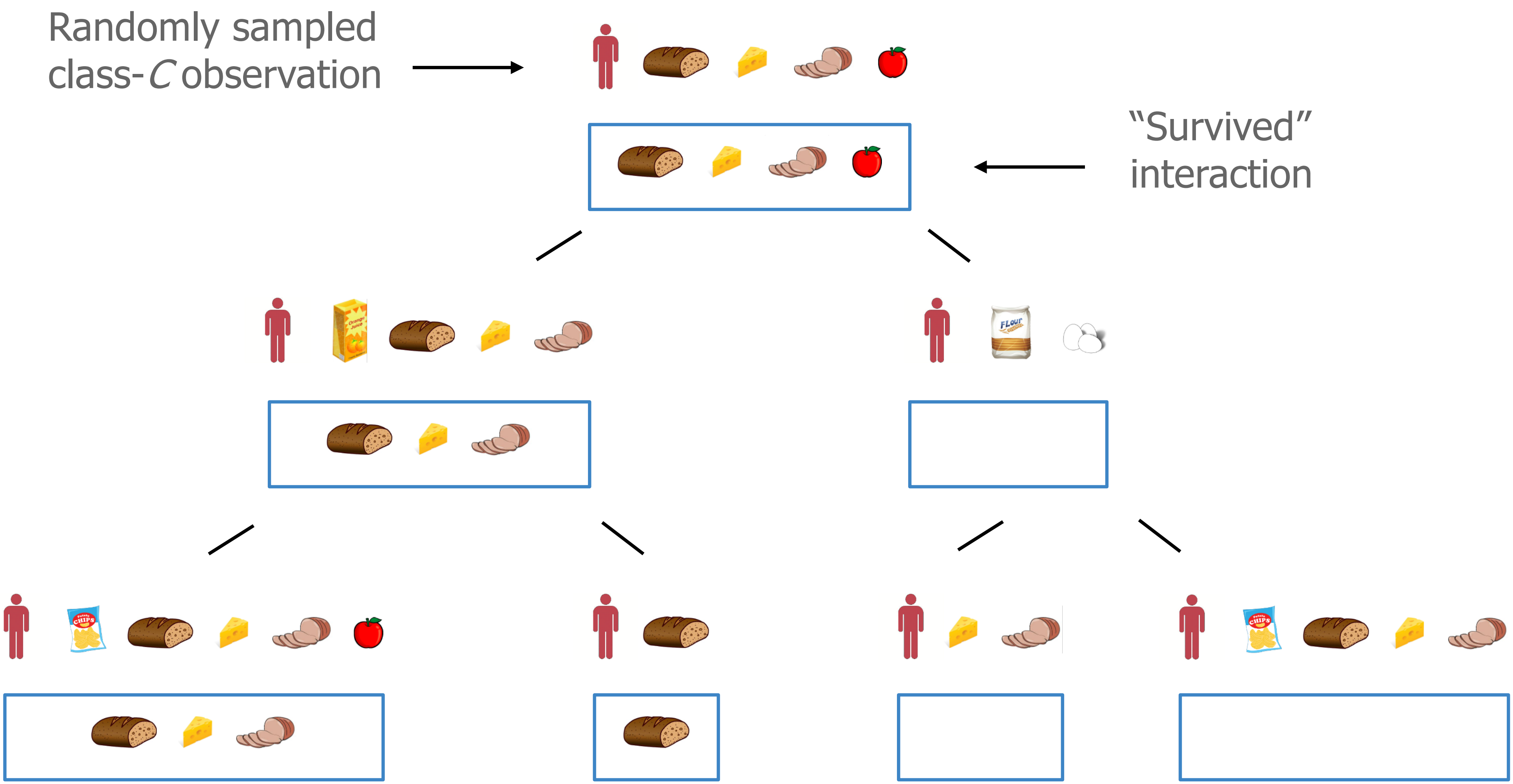
Random Intersection Trees (RIT)

Shah and Meinshausen (2014)



Random Intersection Trees (RIT)

Shah and Meinshausen (2014)



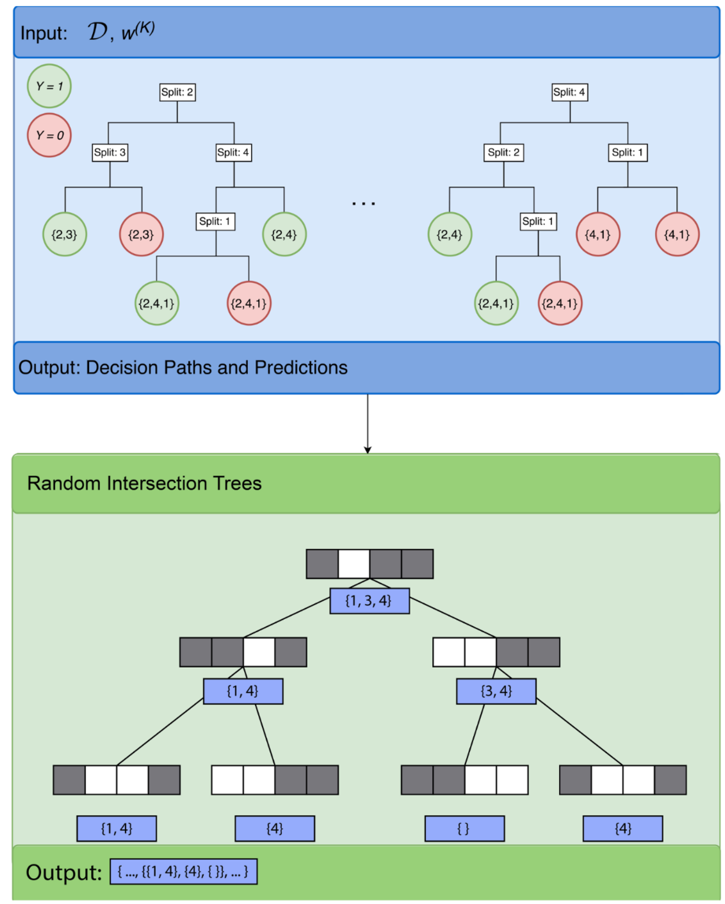
Generalized RIT for Decision Trees

fast computation uses sparsity

$\mathcal{I}_{i_t} \subseteq \{1, \ldots, p\}$ *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 $t = 1, \ldots, T$

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



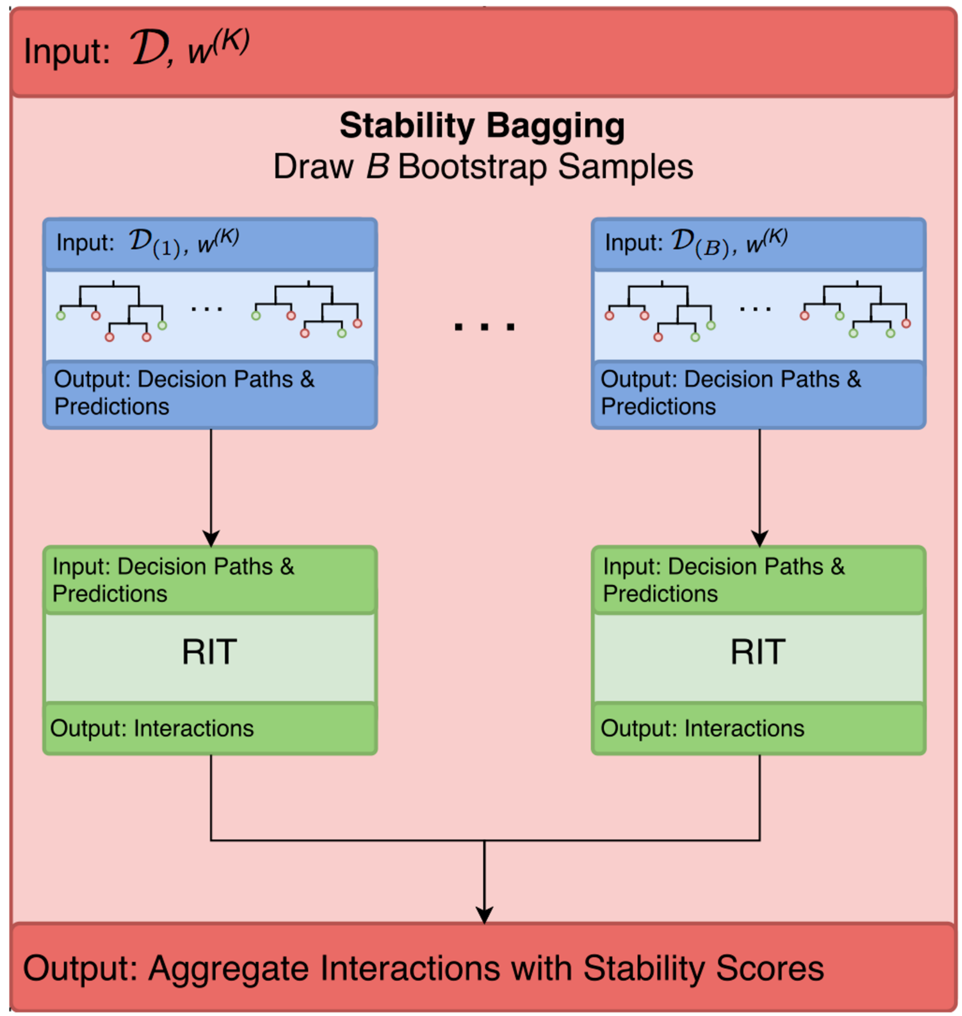
Stability bagging

Output feature interaction sets with stability scores:

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

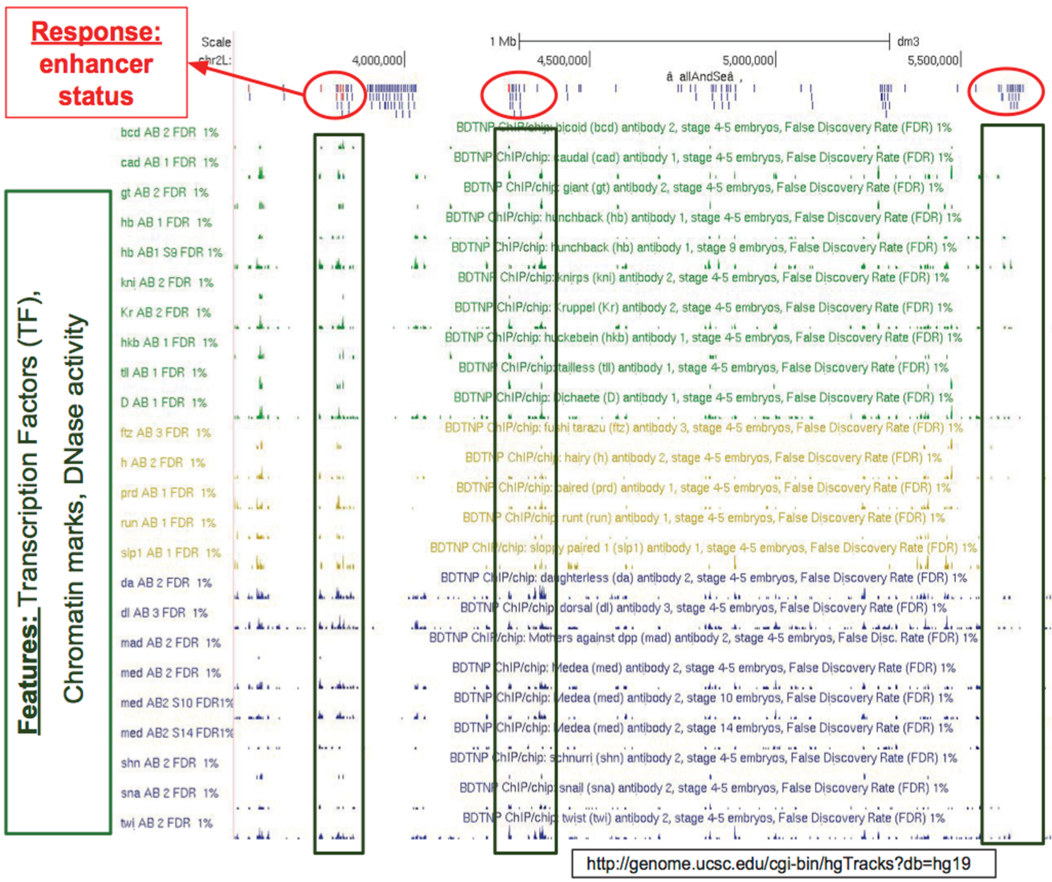
$$S \subseteq \{1, \ldots, p\}$$

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



Reference: (Breiman, 1996)

Example: Enhancer activity in *Drosophila*

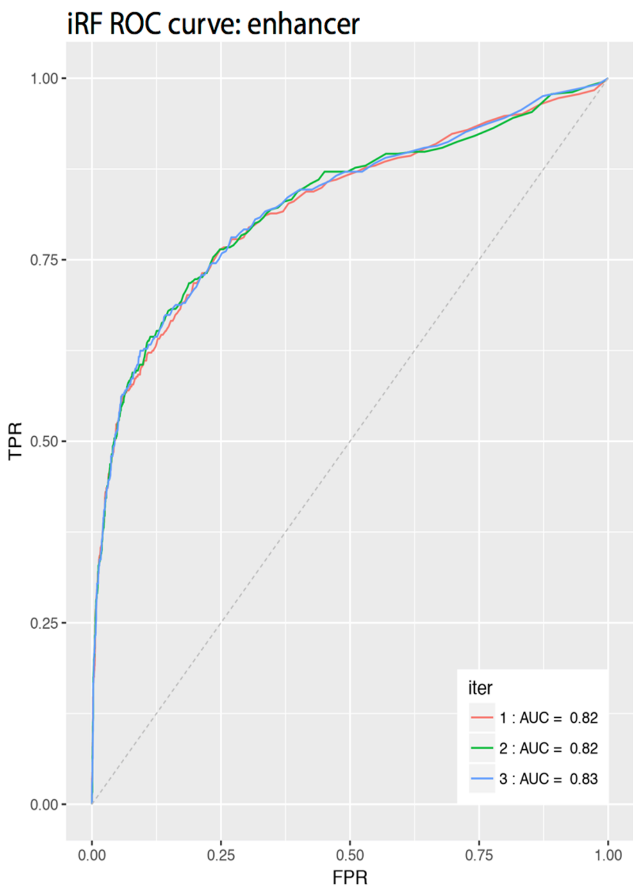


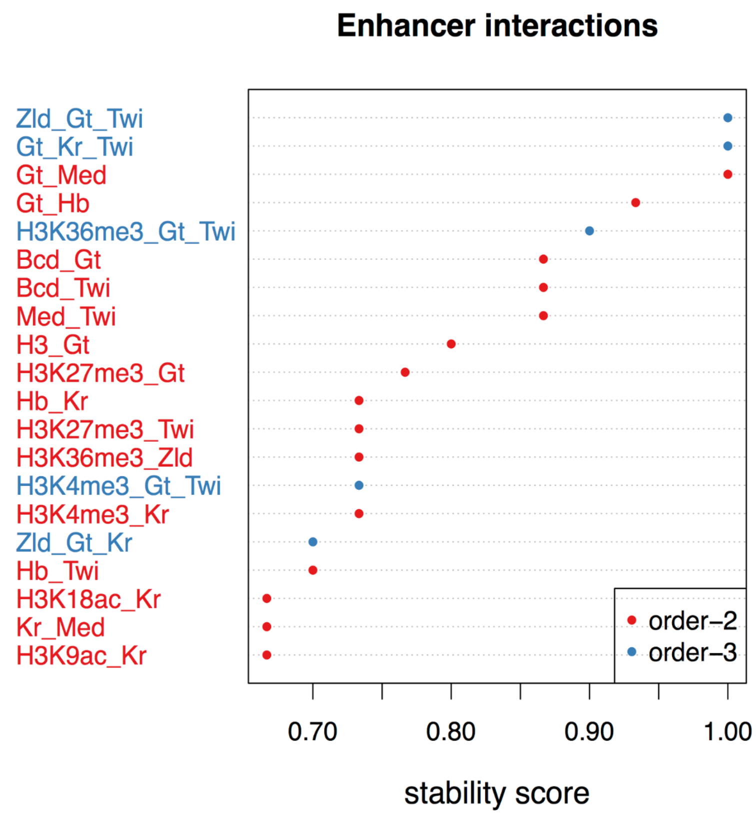
Drosophila blastoderm embryos:

- 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

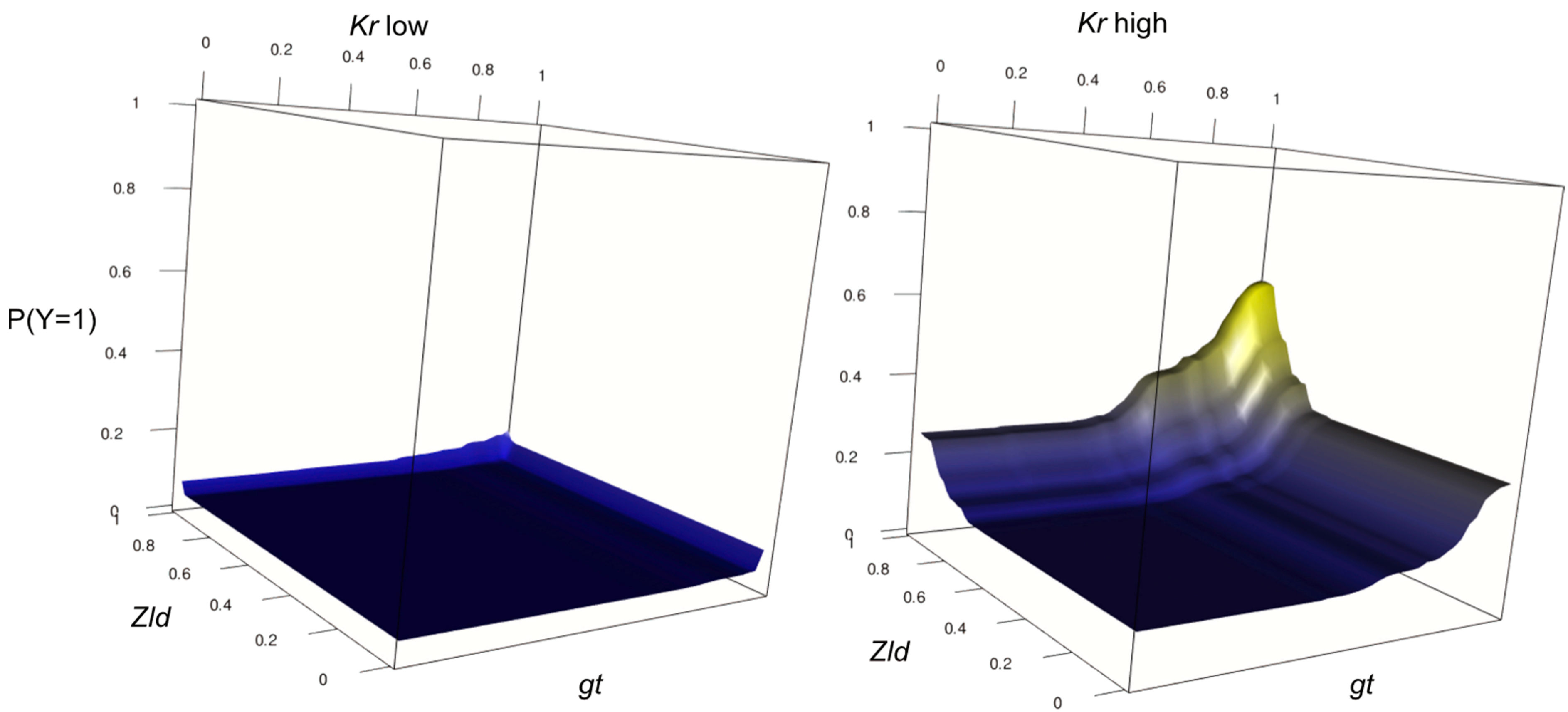




80% *of pairwise interactions are validated*

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



3rd or 4th or higher order interactions are suggestions for Crispr experiments

2018: Chan Zuckerberg Biohub Intercampus Award
iRF is a cornerstone


CHAN ZUCKERBERG
BIOHUB
PEOPLE PROJECTS

CHAN ZUCKERBERG BIOHUB AWARDS $13.7 MILLION TO FUND NEW INTERCAMPUS COLLABORATIVE RESEARCH PROGRAMS TO ADVANCE HUMAN HEALTH

SAN FRANCISCO — Sept. 26, 2018

One of the 6 awards

Project leaders:
Rima Arnout and Atul Butte (UCSF)
James Priest and Euan Ashyley (Stanford)
Ben Brown and **Bin Yu** (UC Berkeley)
Collaborators:
Chris Re (Stanford), Deepak Srivastava (UCSF)

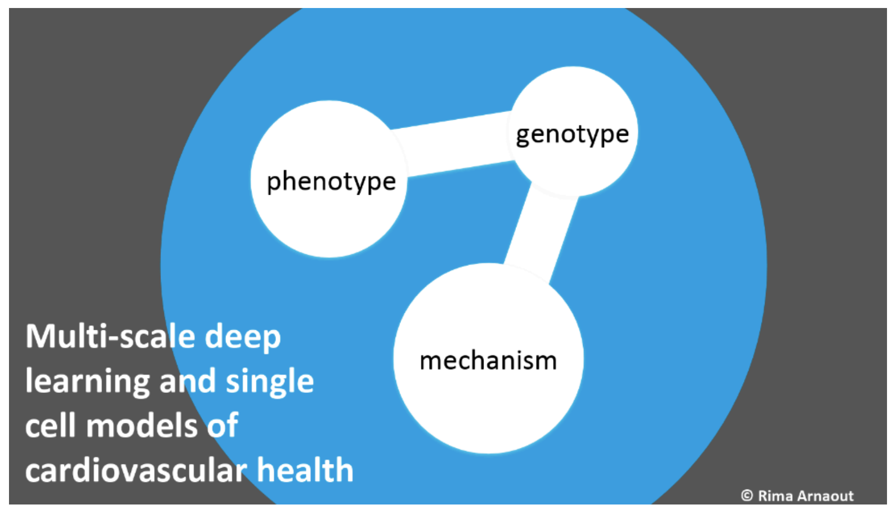
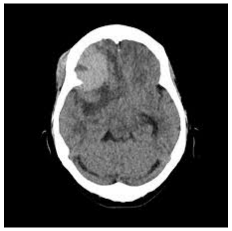


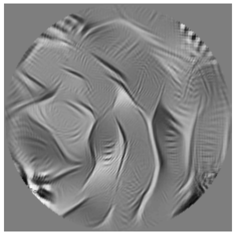
Image credits: Rima Arnout.

Interpreting iRF results generates biological hypotheses

Other examples of interpretation need

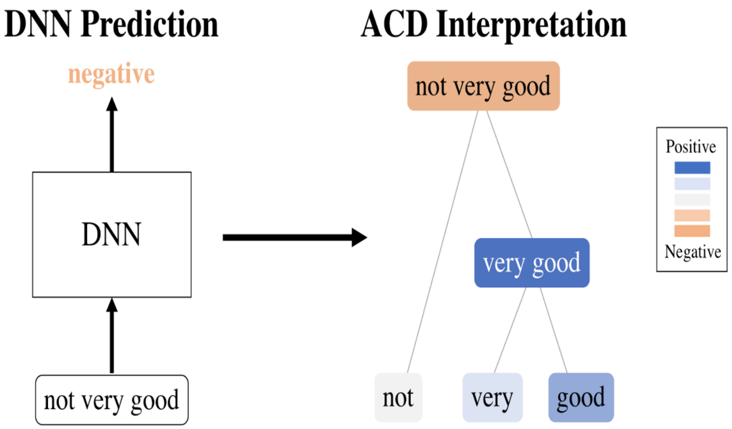
- FDA wants interpretation of DL algorithms for radiology





- 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**

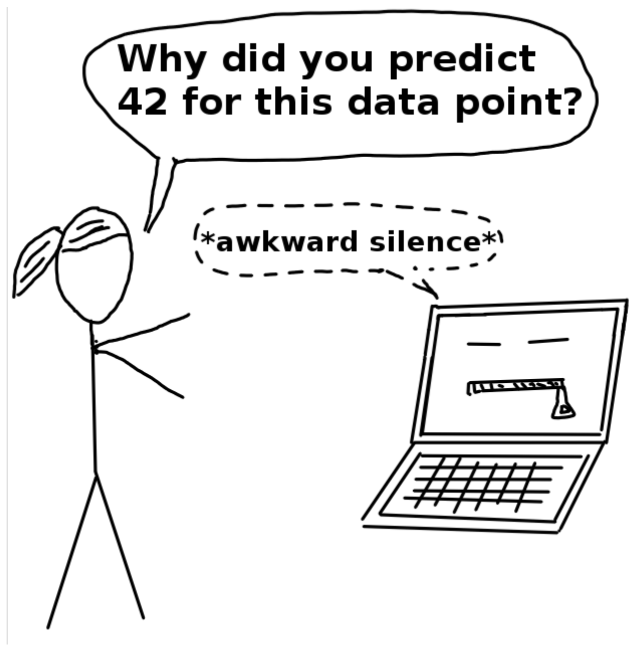


Image credit: https://christophm.github.io/interpretable-ml-book/

Some related work

- Lipton (2017)
- Doshi-Velez and Kim (2017)
- Molnar (2019) book

“Definitions, Methods and Applications in Interpretable Machine Learning”

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

“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

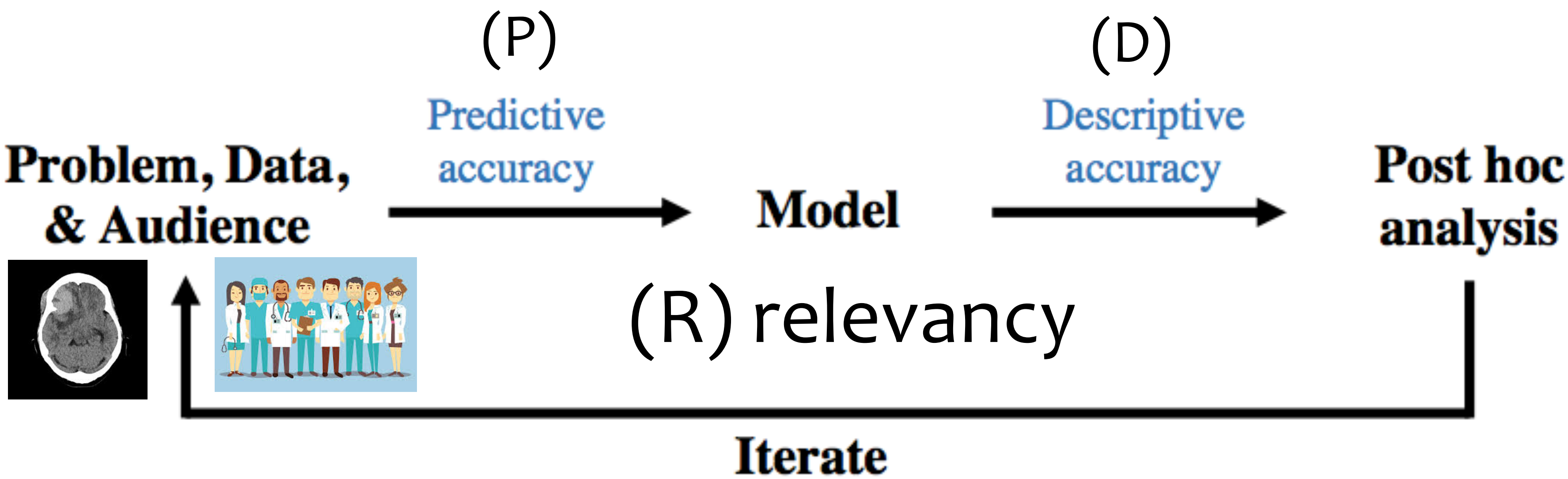
  average (global) and point-wise (local)

- **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



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)

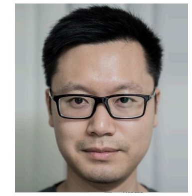
Previous work (post-hoc interpretation)

- 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)

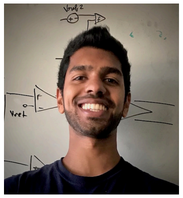
- Given a LSTM with weights, CD gives a prediction-level score for **each part of the input** to "explain" the prediction

The movie was **not good** → CD → Negative / Positive / Neutral

$$\mathrm{LSTM}(w_1, \ldots, w_T) = \mathrm{SoftMax}(\gamma_T + \alpha_T)$$

- γ_T corresponds to contributions solely from the phrase, α_T other factors

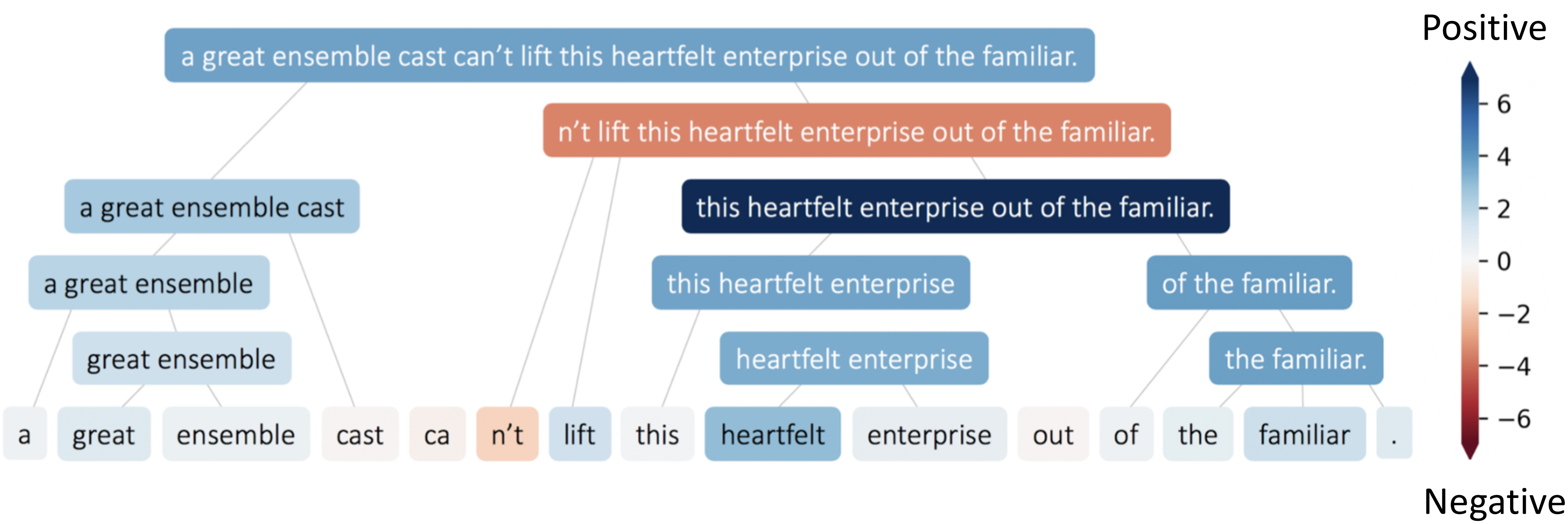
Agglomerative Contextual Decomposition (ACD)



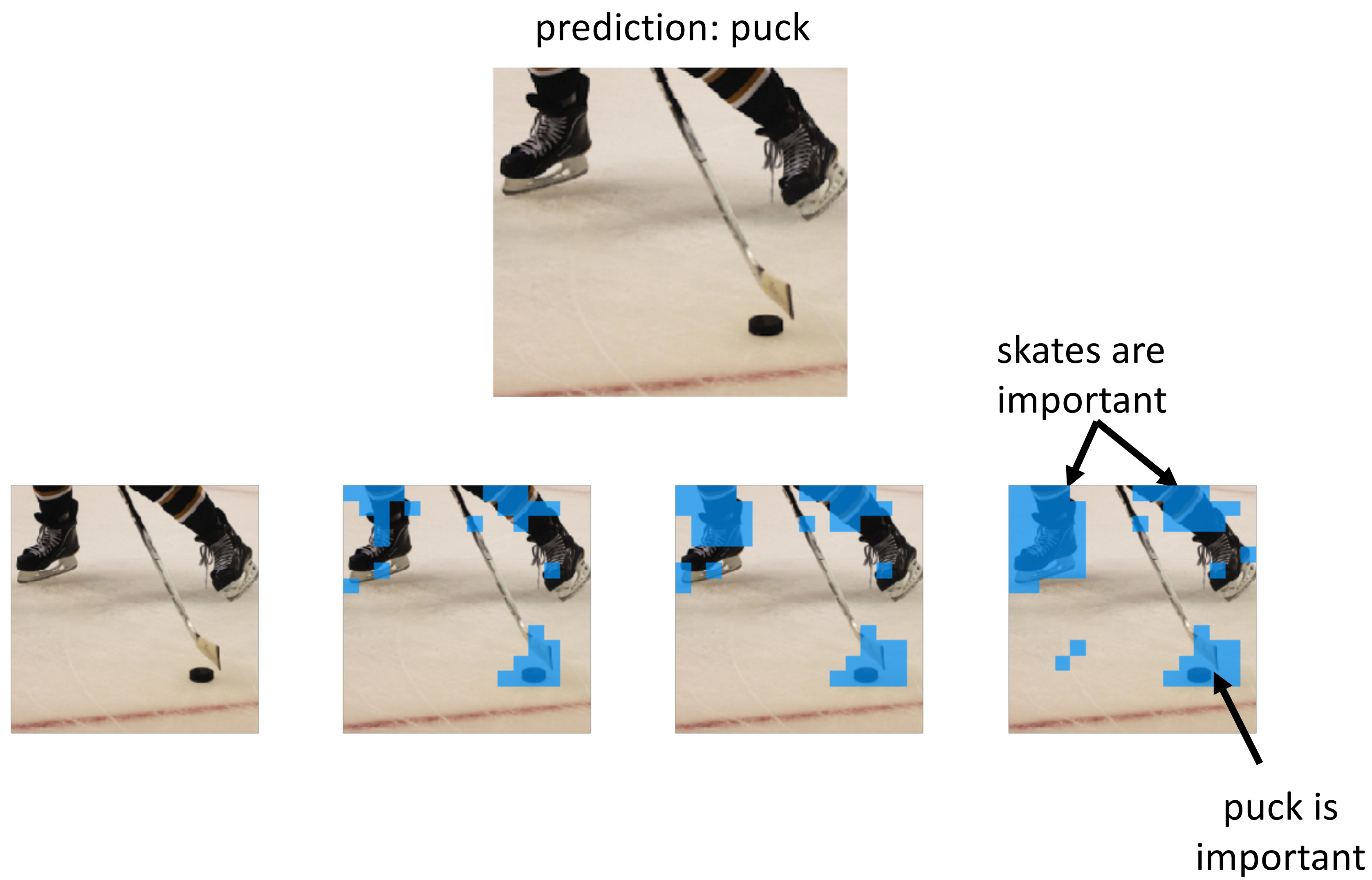
*Singh, *Murdoch, Y. (2019). ICLR

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/csinva/acd

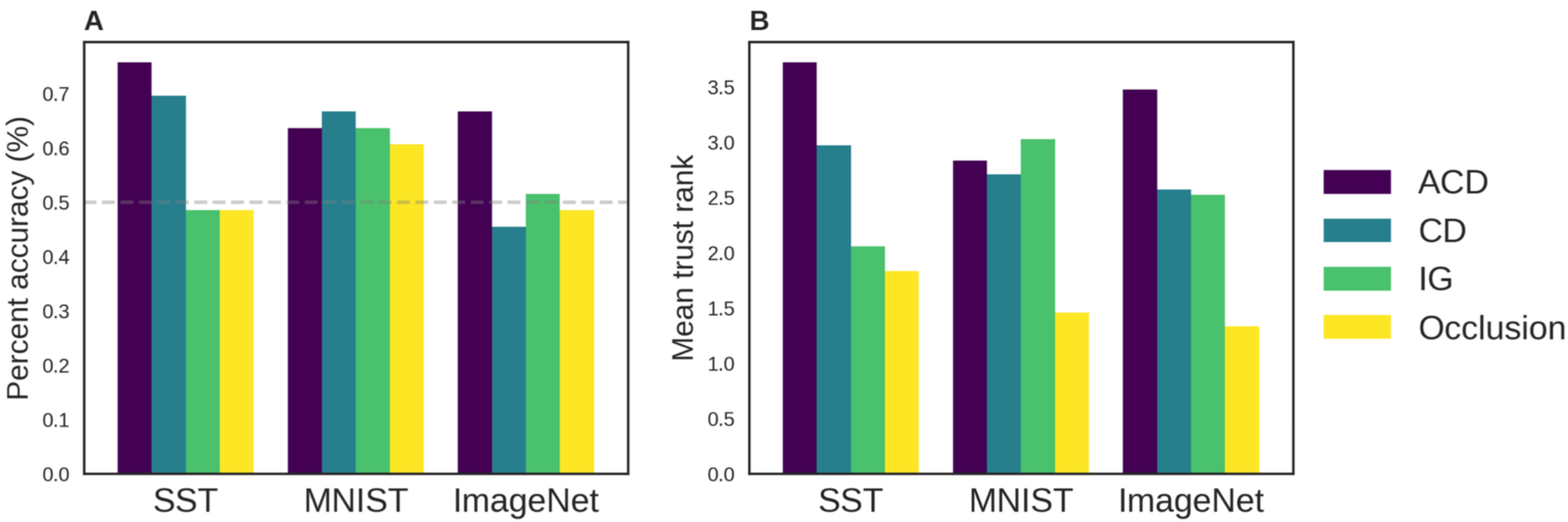


a great ensemble cast can't lift this heartfelt enterprise out of the familiar.
n't lift this heartfelt enterprise out of the familiar.
a great ensemble cast
this heartfelt enterprise out of the familiar.
a great ensemble
this heartfelt enterprise
of the familiar.
great ensemble
heartfelt enterprise
the familiar.
a
great
ensemble
cast
ca
n't
lift
this
heartfelt
enterprise
out
of
the
familiar
.
Positive
6
4
2
0
−2
−4
−6
Negative



prediction: puck
skates are important
puck is important

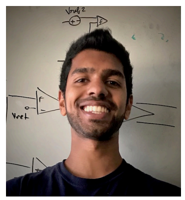
Human experiments



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

Whether Interpretation instills trust or not

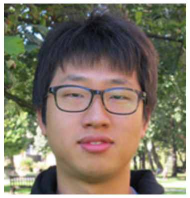
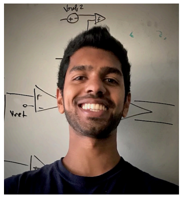
Improving models by regularizing ACD explanations

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

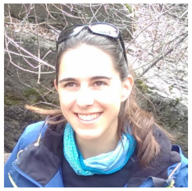
github.com/laura-rieger/deep-explanation-penalization

Using CD to identify fundamental cosmological parameters of the universe


Yu group

In Progress

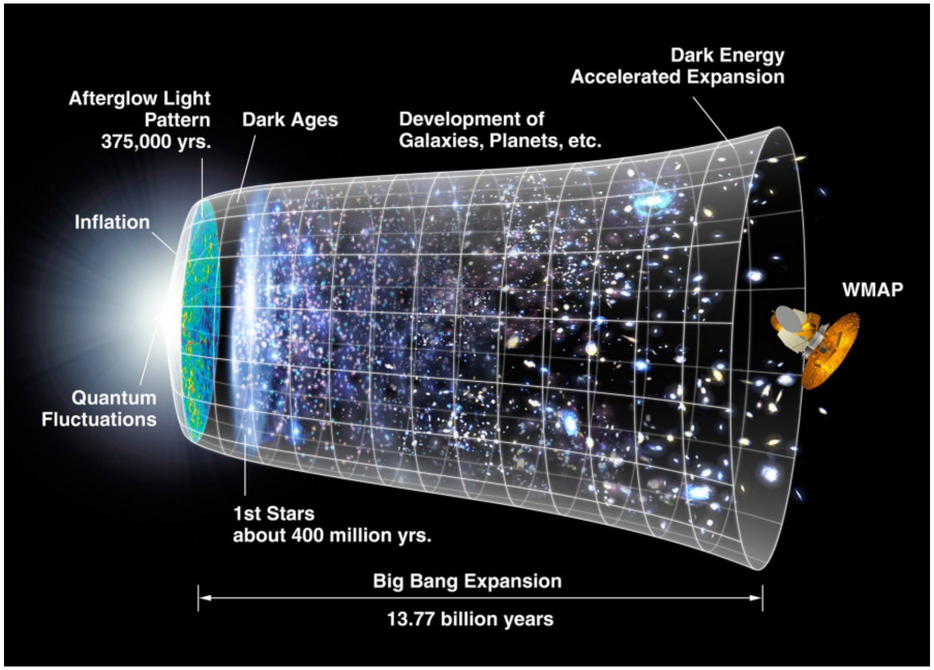



@ Berkeley Center for Cosmological Physics

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

Cosmological parameters such as Ω_M, determine evolution of universe

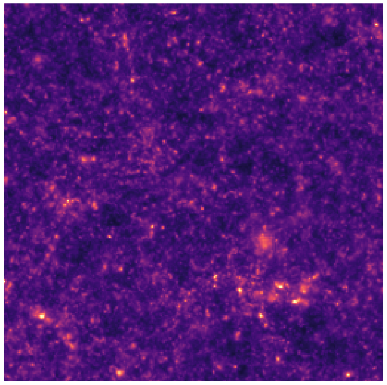
$\Omega_M \rightarrow$



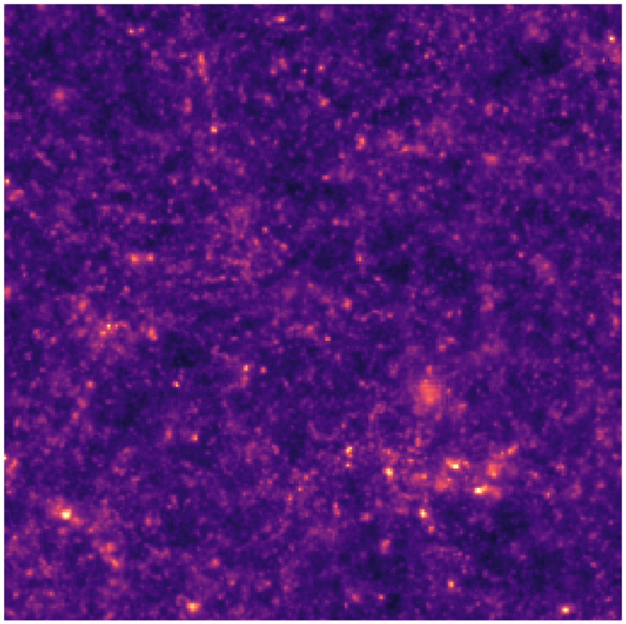
Adaptation of NASA WMAP Science Team Image

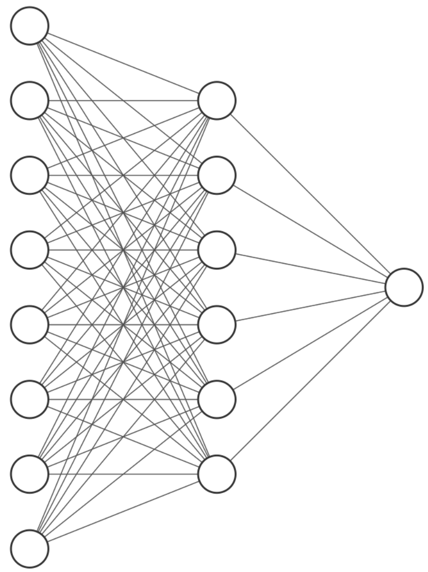
$\rightarrow$

Map of mass in the universe



CNN predicts well, but what does it learn?



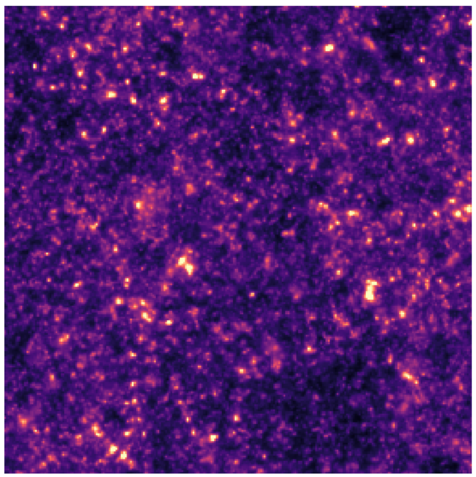


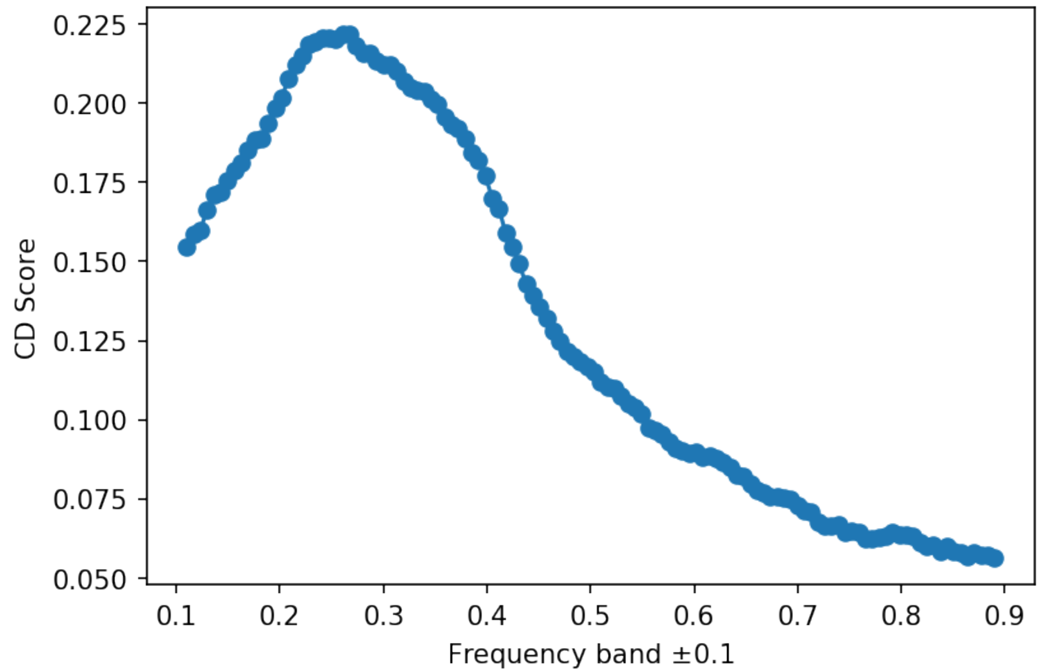
$\rightarrow \Omega_M$

Need to go beyond just identifying important pixels...

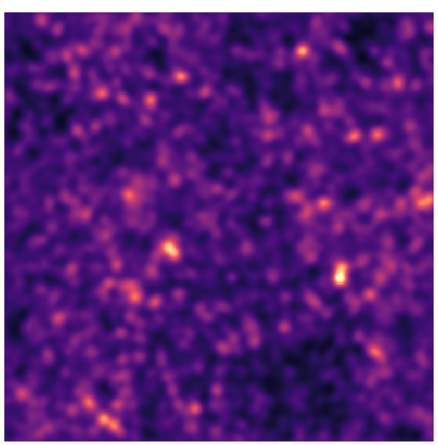
CD can measure the importance of different **frequencies** in the image to the model’s prediction

Original image

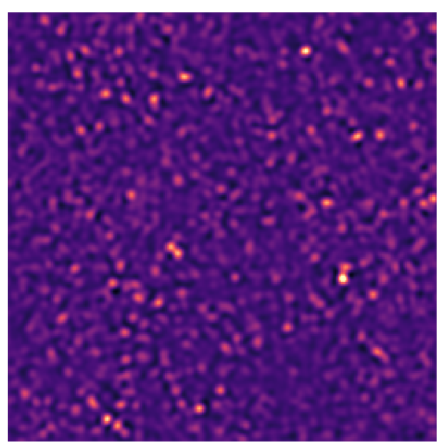




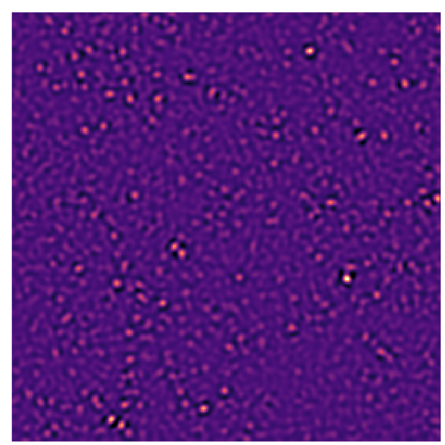
0.1



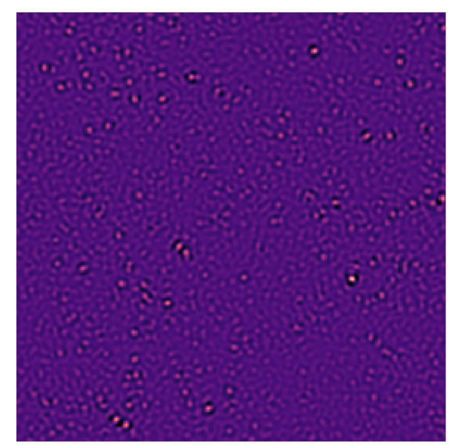
0.2



0.3



0.4



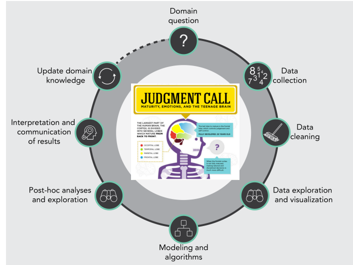
Goals of (faithful) interpretation

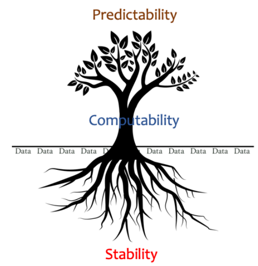
- Save on data collection
- understand which features drive the predictions
- give trust to using deep learning
- distill the DL model into a simple model (e.g**. generative and mechanistic**)

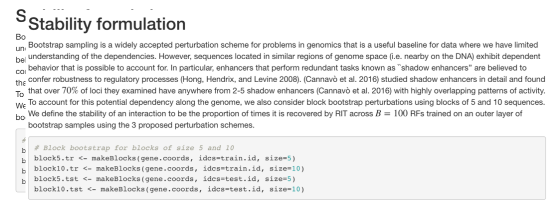
Success of these goals serves as validation

"Data science process: one culture"

Summary







Stability formulation

Bootstrap sampling is a widely accepted perturbation scheme for problems in genomics that is a useful baseline for data where we have limited understanding of the dependencies. However, sequences located in similar regions of genome space (i.e. nearby on the DNA) exhibit dependent behavior that is possible to account for. In particular, enhancers that perform redundant tasks known as "shadow enhancers" are believed to confer robustness to regulatory processes (Hong, Hendrix, and Levine 2008). (Cannavò et al. 2016) studied shadow enhancers in detail and found that over 70% of loci they examined have anywhere from 2-5 shadow enhancers (Cannavò et al. 2016) with highly overlapping patterns of activity. To account for this potential dependency along the genome, we also consider block bootstrap perturbations using blocks of 5 and 10 sequences. We define the stability of an interaction to be the proportion of times it is recovered by RIT across $B = 100$ RFs trained on an outer layer of bootstrap samples using the 3 proposed perturbation schemes.

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

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

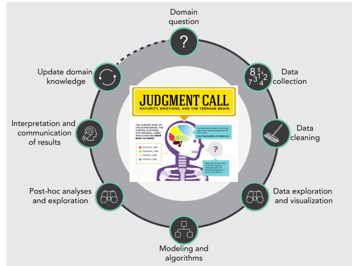
- **PDR** interpretation framework guides selection and evaluation of interpretation methods

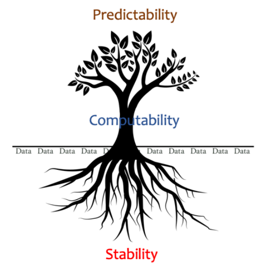
- Case studies: iRF (siRF), ACD (*DeepTune omitted)

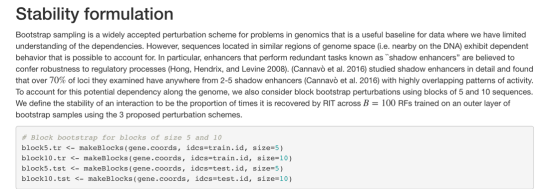
- Domain knowledge is important and **PCS** generates testable hypotheses towards causality

Hope PCS and PDR are useful for your projects

PCS next steps







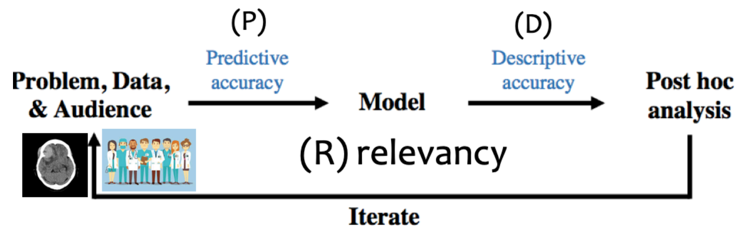
Stability formulation

Bootstrap sampling is a widely accepted perturbation scheme for problems in genomics that is a useful baseline for data where we have limited understanding of the dependencies. However, sequences located in similar regions of genome space (i.e. nearby on the DNA) exhibit dependent behavior that is possible to account for. In particular, enhancers that perform redundant tasks known as ``shadow enhancers" are believed to confer robustness to regulatory processes (Hong, Hendrix, and Levine 2008). (Cannavò et al. 2016) studied shadow enhancers in detail and found that over 70% of loci they examined have anywhere from 2-5 shadow enhancers (Cannavò et al. 2016) with highly overlapping patterns of activity. To account for this potential dependency along the genome, we also consider block bootstrap perturbations using blocks of 5 and 10 sequences. We define the stability of an interaction to be the proportion of times it is recovered by RIT across $B = 100$ RFs trained on an outer layer of bootstrap samples using the 3 proposed perturbation schemes.

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

- **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)

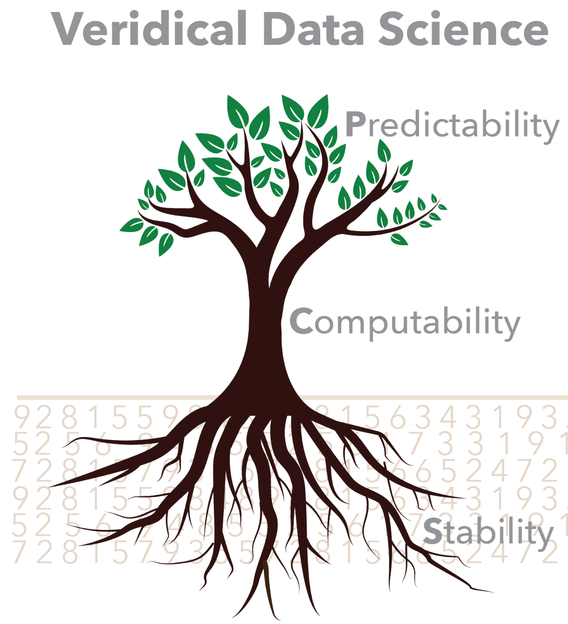
For veridical data science, academic/industry/government leadership and funding agencies need to incentivize

- **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

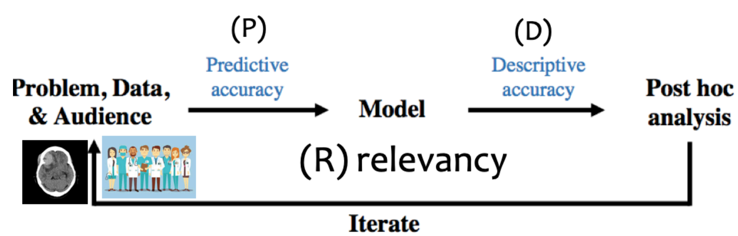
1. Veridical data science
   B. Yu and K. Kumbier (2020), PNAS

(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

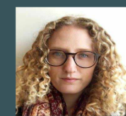
Coming at the end of 2021 with a free on-line interactive version in the spring

Veridical Data Science: A Book

Bin 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	Algorithmic	Stability-based inference
Data cleaning	Dimension reduction	Inference
Exploratory Data Analysis	Clustering	Causal Inference
Data merging	Least Squares & ML	Perturbation Intervals
	Regularization	Trustworthiness Statements

Communication

Exploratory Visual Summaries

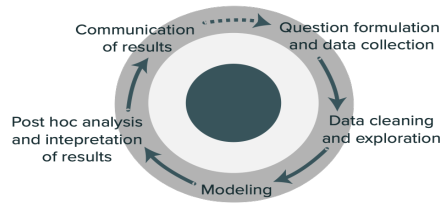
Preparing explanatory visual and numeric summaries for explaining data and findings to an external audience

Written reports

Preparing written analytic reports for case studies based on real, messy data

Core guiding principles for the book

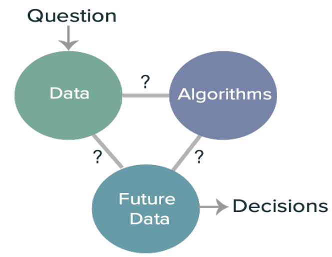
The DS Lifecycle



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.

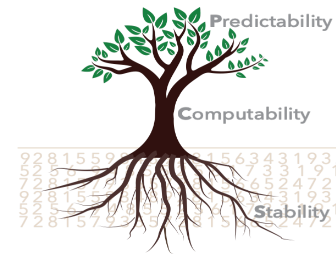
Blending together concepts from statistics, computer science and domain knowledge, the data science life cycle is an iterative process that involves human analysts learning from data and refining their project-specific questions and analytic approach as they learn.

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

Thank you!

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