

Veridical Data Science towards Trustworthy AI

Bin Yu

Statistics, EECS, Computational Biology
UC Berkeley

Statistics Seminar, QUT
Jan.12, 2024

AI is part of modern life

Virtual assistants
(Siri, Alexa,
Cortana)

Online news

Bill Gates: A.I. is like nuclear energy — 'both promising and dangerous'

Published Tue, Mar 26 2019 8:45 AM EDT • Updated Tue, Mar 26 2019 11:40 AM EDT



Catherine Clifford
@CATCLIFFORD

Share [f](#) [t](#) [in](#) [✉](#)

Recommendation
systems
(YouTube, Facebook)

Online gaming

Self-driving cars

Sociology

Election campaigns

Precision medicine

Chemistry

Wearable health
devices
(FitBit, Apple watch)

Biology

Neuroscience

Materials Science

Economics

Political Science

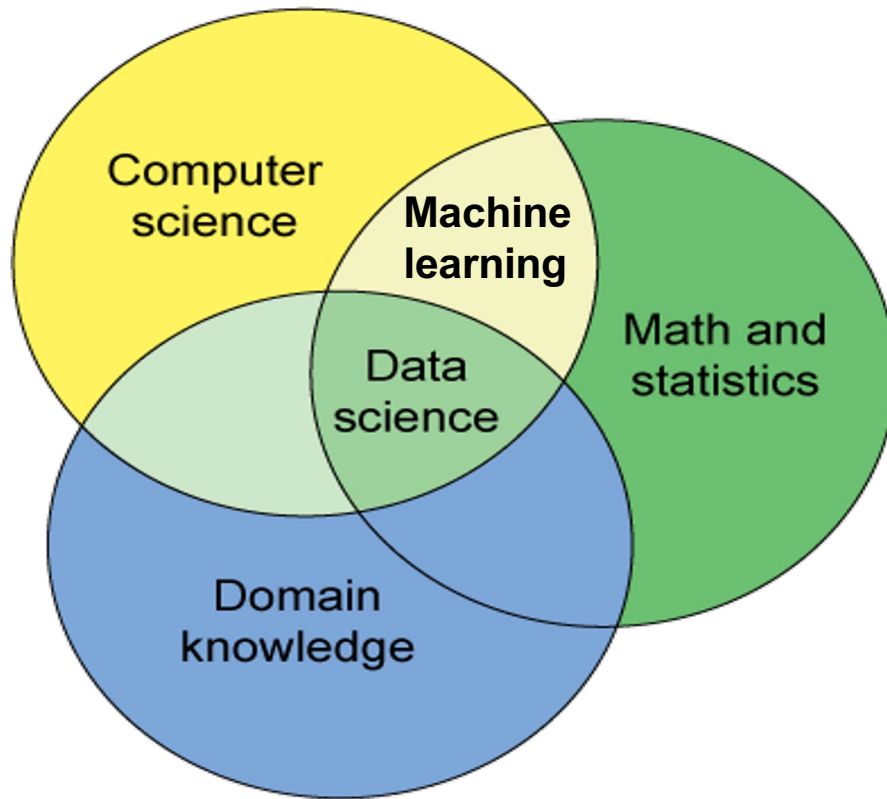
Cosmology

Law

... 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

CZ Biohub intercampus research award (2018-2021)



Stanford



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: Victoria Parikh, Chris Re, Deepak Srivastava



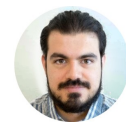
M. Behr



K. Kumbier



M. Aguirre



A. Cordova-
Palomera



Q. Wang



N. Youlton



C. Weldy



W. Hughes



A. Agarwal



T. Tang



O. Ronen



X. Li



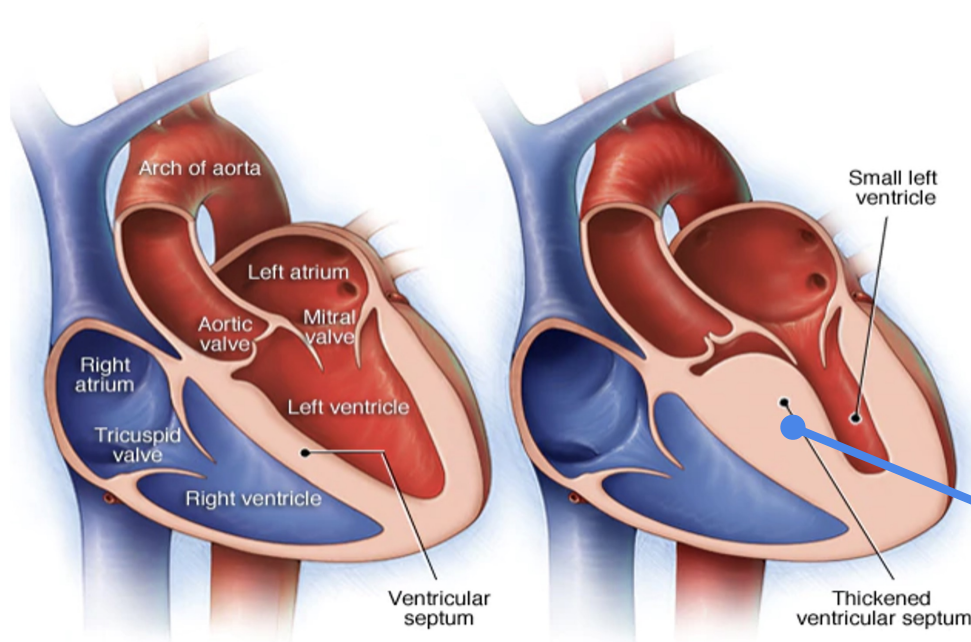
A. Kenney

Hypertrophic Cardiomyopathy (HCM)

HCM is a genetic heart disease, characterized by **thicker walls** of the heart chamber (left ventricle).

Normal Heart

HCM Heart

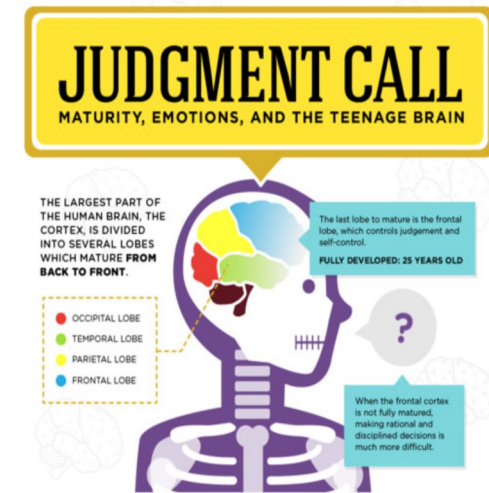


Past research has shown that large **cell size** is associated with HCM and there is an important genetic component to it.

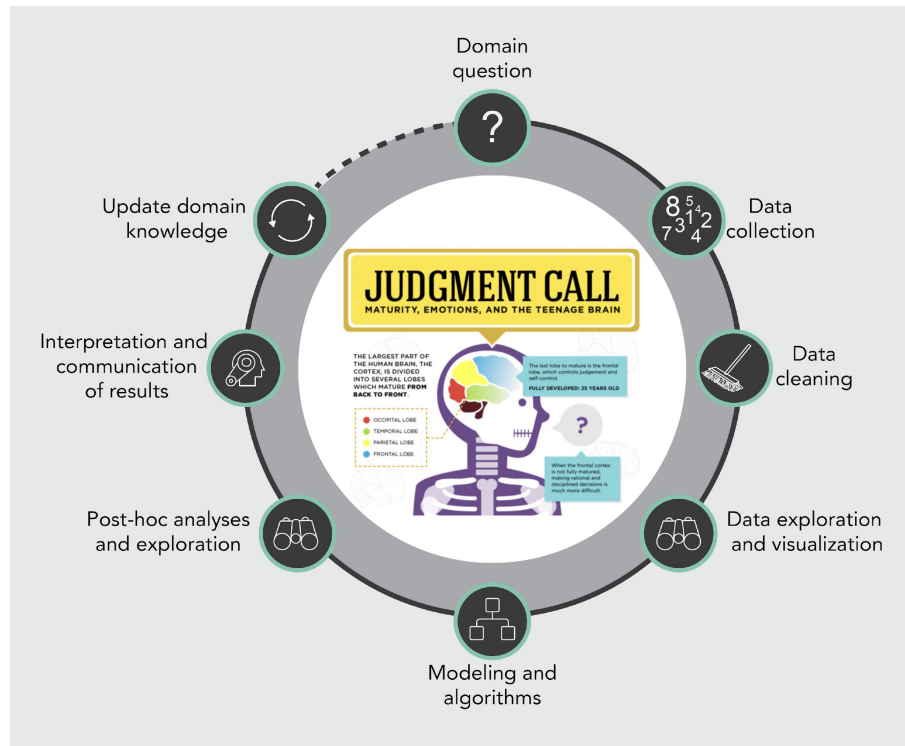
Thickened heart wall

Recommendation of genetic drivers of HCM as a data science problem

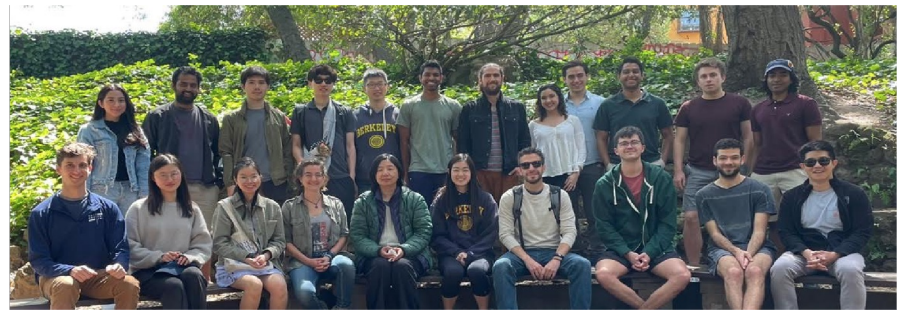
- Medical **question**: which genes interact to drive HCM?
- Which **data** to use? How to **clean**?
- **EDA**: summaries, plots, ...
- **Modeling**: Which algorithms to use to find nonlinear interactions?
- **Interpretation & evaluation** of recommendations for gene-silencing experiments in Ashley Lab



Data Science Life Cycle (DSLCL): A holistic view



Every step is a source of uncertainty due to data collection process, and human judgment calls.



Box (1979). Cox and Snell (1981), Nelder (1991)....

Image credits: R. Barter and toronto4kids.com

Uncertainty quantification is central for building trust in AI

Current approach considers

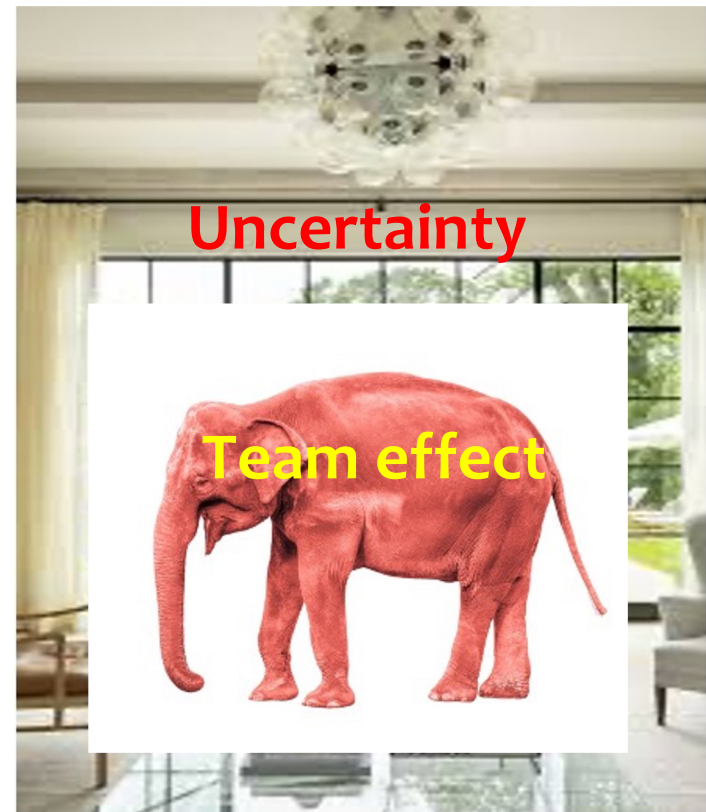
only uncertainty from a generative stochastic model,

which is often assumed, with limited empirical checking.

In a data science life cycle (DSLCL), there are many other important sources of uncertainty, due to human judgment calls.

Realistic/trustworthy uncertainty quantification is a must.

In our house of “uncertainty”



<https://www.housebeautiful.com/room-decorating/living-family-rooms/g715/designer-living-rooms/>

<https://www.forbes.com/sites/janicegassam/2019/11/27/the-pink-elephant-in-the-workplace-how-to-have-conversations-about-race-politics-and-religion-at-work/>

Applied Stats 215A Final Project in Fall 2021



TA: O.Ronen

- Students developed models to predict the risk of **Traumatic Brain Injuries (TBI)** for kids
- Three groups of students, each team with a UCSF medical doctor, worked on the problem **independently**, using the same raw data and with the same data cleaning guidelines

In terms of sensitivity,
uncertainty (10%) from data cleaning choices is similar to
uncertainty from bootstrap samples from each cleaned dataset.



C. Singh



A. Kornblith

“Team effect or algorithm choice” uncertainty

PNAS

RESEARCH ARTICLE

SOCIAL SCIENCES

 OPEN ACCESS

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022

“... [Seventy-three independent research teams](#) used identical cross-country survey data to test a prominent social science hypothesis: that more immigration will reduce public support for government provision of social policies. Instead of convergence, [teams’ results varied greatly, ranging from large negative to large positive effects of immigration on social policy support.](#) The choices made by the research teams in designing their statistical tests explain very little of this variation; a hidden universe of uncertainty remains....”

To gain trust in and maximize promise of DS or AI

Data conclusions must **capture reality** and be **stable** to human judgment calls throughout an integrated data science life cycle (DSLCL).

A **quality control** protocol is necessary, which is built on successful empirical practice.

Rest of the talk

- Predictability-Computability-Stability (PCS) framework/documentation
- PCS-based method development: iRF
- Finding genetic drivers of HCM: iRF recs and intervention experiments
- Another PCS case study: pancreatic cancer risk prediction
- Current PCS directions: PCS inference, softwares, document template, extensions of PCS by others, ...

Predictability, Computability, Stability (**PCS**) framework and documentation for veridical data science

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

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: **one culture**

Yu and Kumbier (PNAS, 2020)



Three principles of data science:

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

(**C**)omputability [ML]

(**S**)tability [Stats, control theory, numerical analysis]

Veridical Data Science

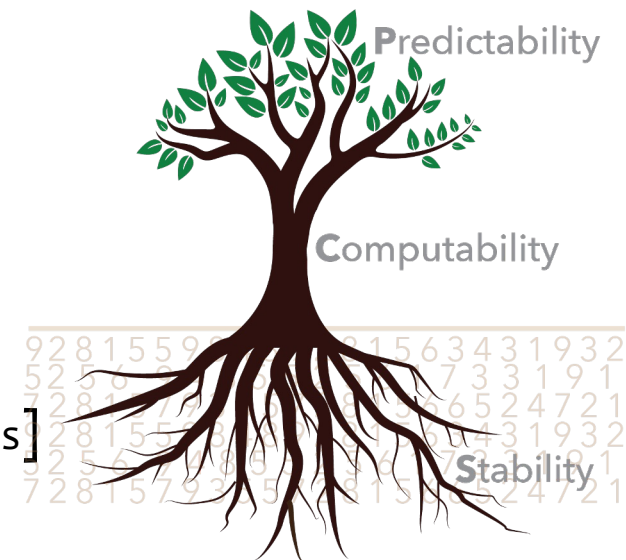


Image credit: R. Barter

PCS

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

Builds a **platform** for further developments.

The stability principle

Stability

BIN YU

*“**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.”*

Ioannidis, 2005; Kraft et al., 2009, Donoho, 2010; Casadevall and Fang, 2011; Nosek et al., 2012; Gelman and Loken, 2014,...

Stability is also a prerequisite for interpretable ML or explainable AI.

Stability is key to statistical and ML theory

Central limit theorem

Concentration inequalities

Random matrix results

Uniform stability for generalization

...

How to shake DSLC **reasonably**: possible data perturbations

- Bootstrap
- Subsampling
- Adding small noise to data
- Bootstrapping residuals
- Block-bootstrap

Understanding data collection process and domain knowledge help to choose to capture uncertainty in data collection.

Document decisions.

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

How to shake DSLC **reasonably**: possible model/algorithm perturbations

- Robust statistics
 - Semi-parametric models
 - Lasso and Ridge
 - Modes of a non-convex empirical minimization
 - Sensitivity analysis in Bayesian modeling
 - DL models from different initializations and optimization algorithms
 - ...
- Domain knowledge and P-check help to choose based on reality-check.
- Document decisions.

How to choose perturbations in PCS?

For **each step** of DSLC, there are **multiple choices**, possibly favored with different weights based on prior/expert knowledge.

To accommodate limited human/computing resources, can **randomly choose reasonable actions** at each step for a total of **N perturbation “paths.”** (“Forking” in Gelman and Loken, 2014) or use weighing based on domain knowledge.

Record all human reasoning and judgment calls using **PCS documentation**.

PCS documentation [on GitHub (Jupyter Notebook R Markdown)]



Reality



Stability formulation

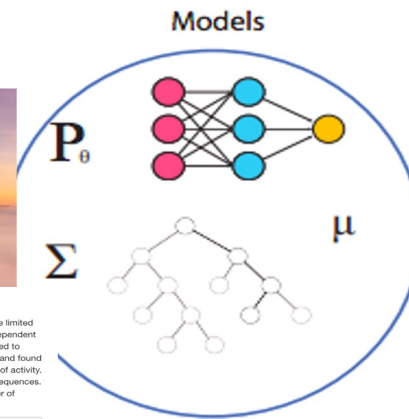
Bootstrap sampling is a widely accepted perturbation understanding of the dependencies. However, sequer behavior that is possible to account for. In particular, confer robustness to regulatory processes (Hong, Her that over 70% of loci they examined have anywhere fi To account for this potential dependency along the ge We define the stability of an interaction to be the prop bootstrap samples using the 3 proposed perturbation

JUDGMENT CALL



seful baseline for data where we have limited ce (i.e. nearby on the DNA) exhibit dependent in as "shadow enhancers" are believed to studied shadow enhancers in detail and found 316) with highly overlapping patterns of activity. urbations using blocks of 5 and 10 sequences. 7 = 100 RFs trained on an outer layer of

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



Mental
Construct

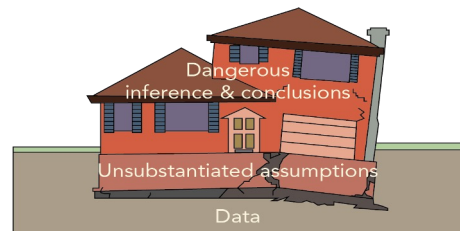


Image credits: Rebecca Barter

PCS documentation template: <https://yu-group.github.io/vdocs/PCSDoc-Template.html>

PCS is a research program for data science

- Philosophical, conceptual, practical, and standing on basic principles PCS
- A systems approach integrating steps in DSLC via PCS, with expanded uncertainty quantification
- Indispensable PCS documentation

Necessarily vague to require domain knowledge and critical thinking to devise P-checks and **reasonable perturbations** for S-checks “in-context”.

If your lab is doing ideas in PCS already, great! **PCS can still help the beginners to speed up, and even veterans to be more systematic and thorough, esp. through PCS documentation.**

PCS success stories, “in-context”

- New methods: iterative random forests (iRF), staNMF, staDISC, staDRIP, ...

for non-linear gene-gene interaction and fate-mapping in developmental biology, subgroup discovery for RCTs, and drug discovery for cancer, ...

- Case studies: PCS-based epistatic gene recommendations for experiments in cardiology; PCS stress-tests for CDRs in pediatric emergency medicine, ...

lo-siRF with experimental validation to find epistasis genetic drivers of heart disease HCM; Stress-testing clinical decision rules in pediatric emergency medicine

Extensions to spatial stats, network analysis, and reinforcement learning by others.

PCS-based method development

iterative random forests (iRF)

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}

PNAS, 2018

Co-authors



S. Basu

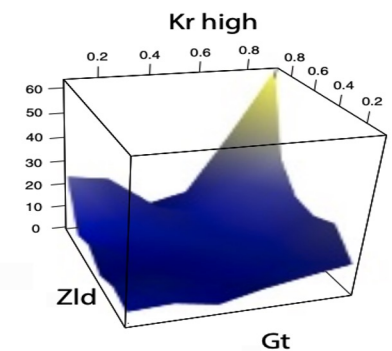
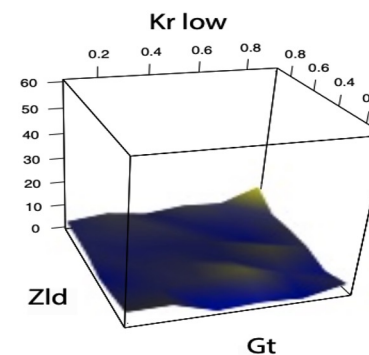


K. Kumbier



B. Brown

Culmination of 3+ years of work



Pattern Recognition vs. Pattern Discovery

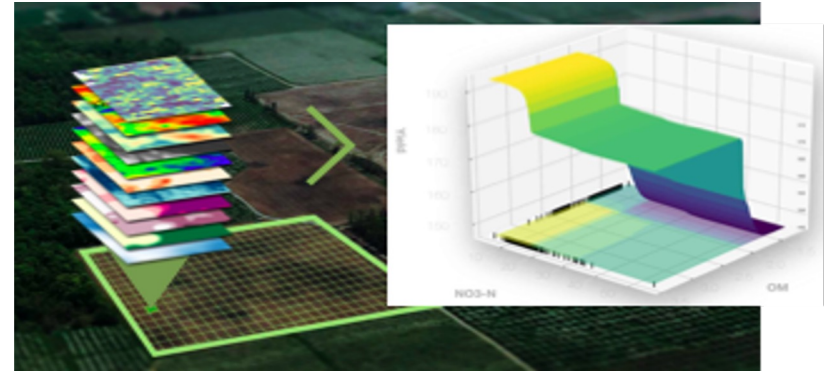
Pattern Recognition:

Finding something for which you already know to look

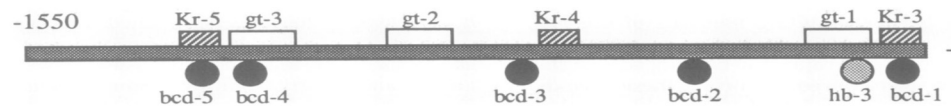
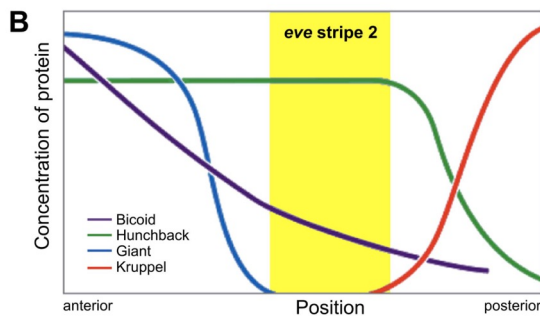
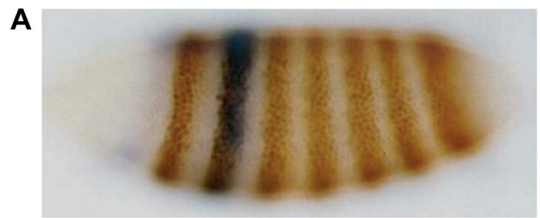


Pattern Discovery:

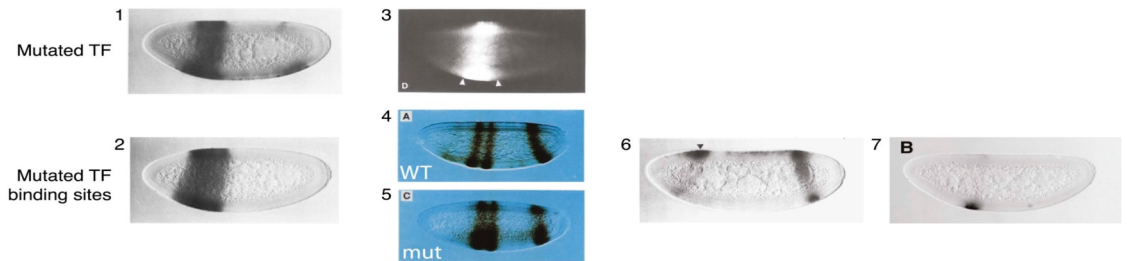
Identifying structure that hasn't been seen before



Order-4 interaction regulates eve stripe 2



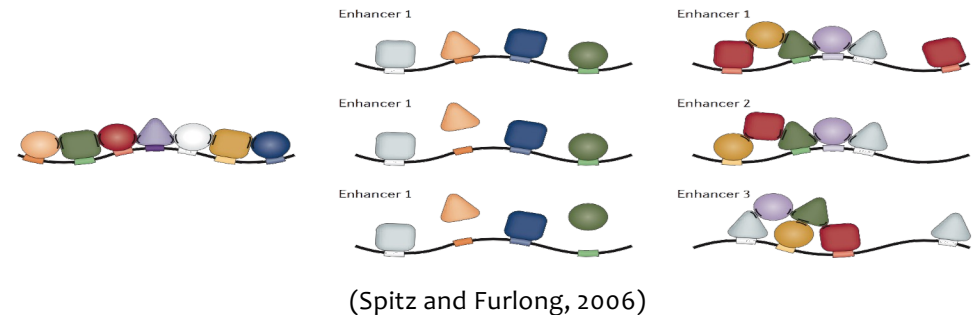
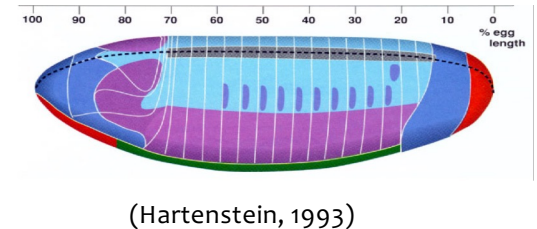
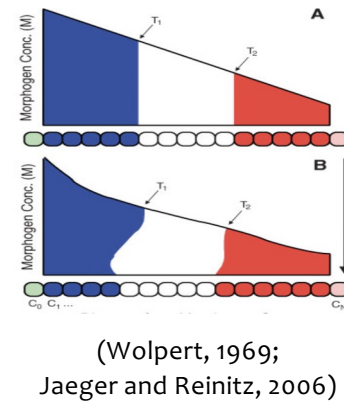
A Perturbing *gt* **B** Perturbing *Kr* **C** Perturbing *bcd* **D** Perturbing *hb*



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



From genomic to statistical interactions

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



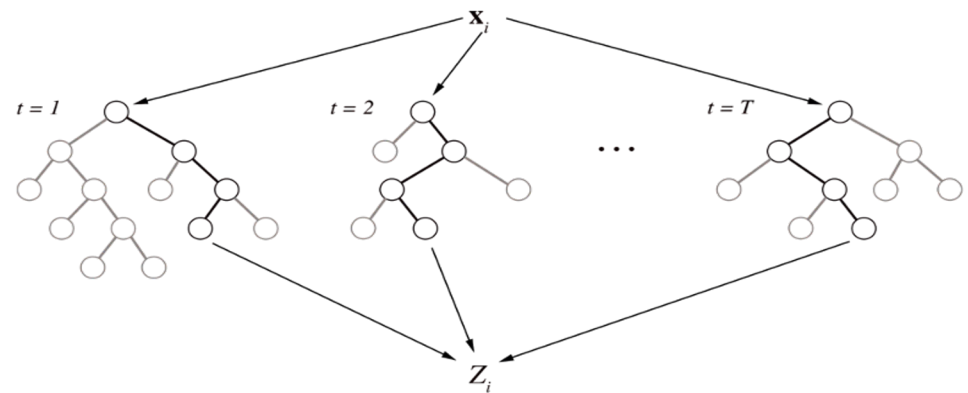
$$R(\mathbf{x}) = \prod_{i \in S} 1\{x_i > t_i\}$$

Order- s interaction,
 $S \subseteq \{1, \dots, p\}, |S| = s$

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 m features uniformly at random and optimize CART criterion over subsampled features.



iterative Random Forests (iRF)

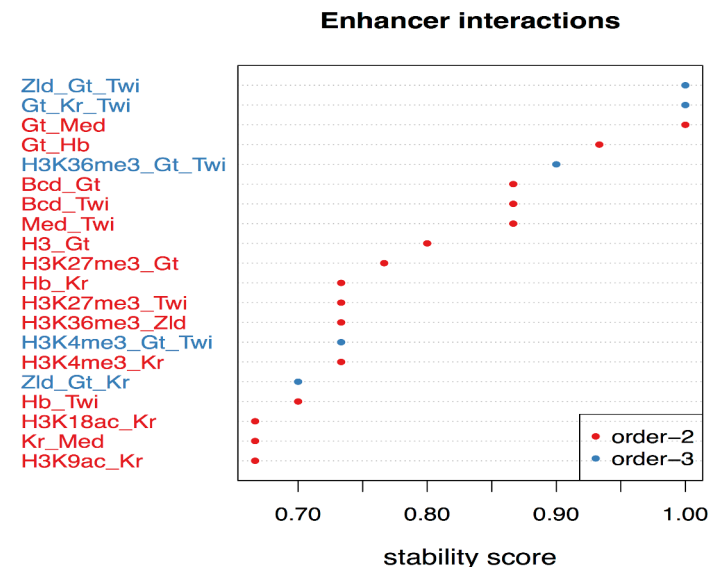
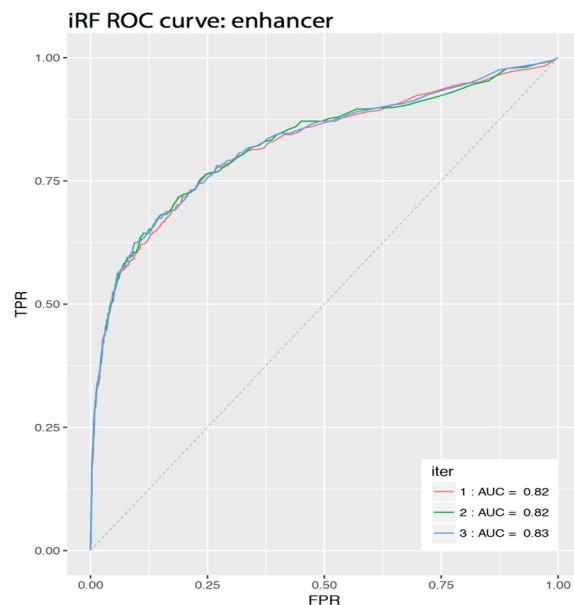
Basu, Kumbier, Brown and Yu (2018)

Core idea: **add stability** to random forests (RF)

1. **Soft dim reduction** using importance index to sample features
1. Random interaction trees (RIT) to find intersections of paths
1. Outer-loop bagging assesses **stability**

Similar computational and memory costs as RF

iRF keeps predictive accuracy, and finds stable interactions for a Drosophila enhancer prediction problem

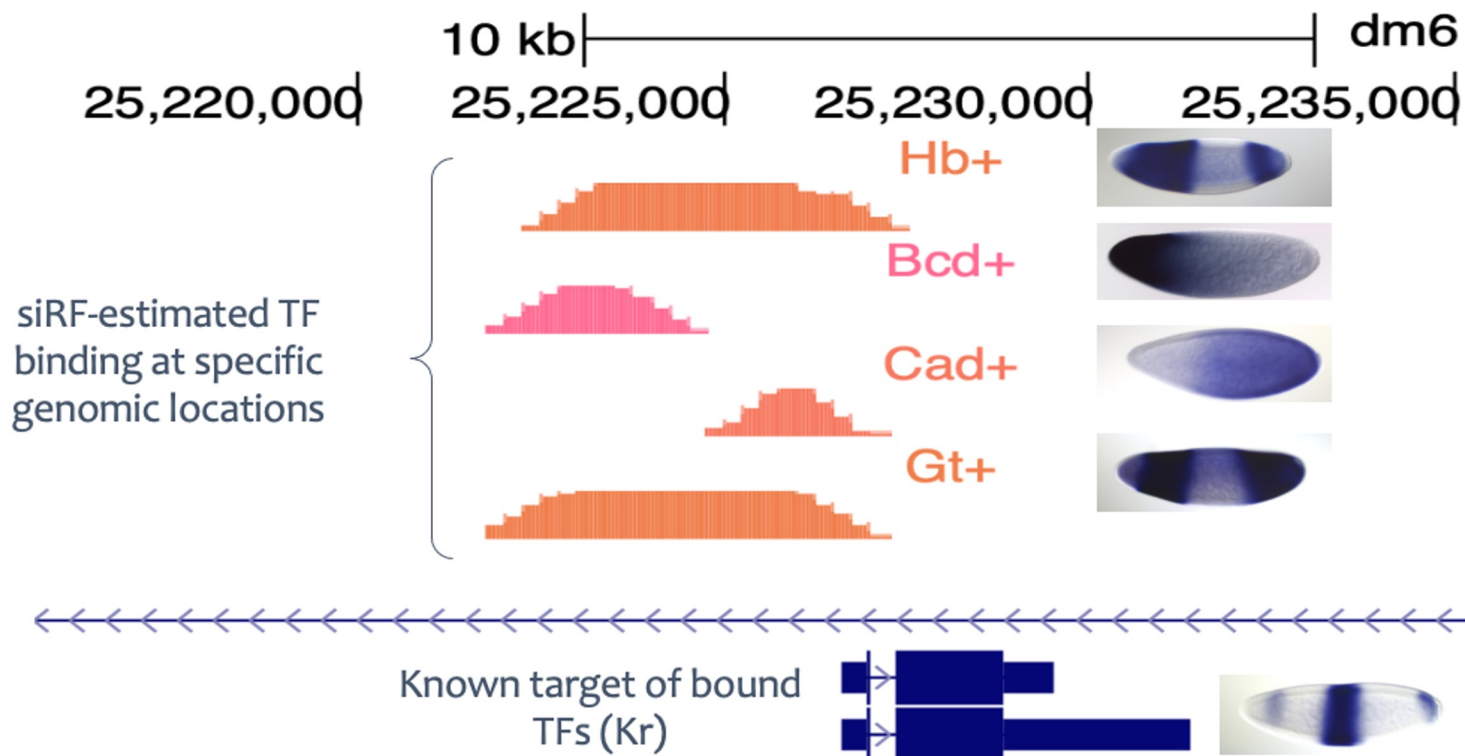


External validation: 80% of pairwise interactions are validated
by past biological experiments in the literature

siRF-estimated TF binding will be made available as a **UCSC genome browser track**
siRF: signed iRF (2018 arXiv, to be submitted soon)



K. Kumbier



Other co-authors: S. Celniker, B. Brown, E. Frise, S. Basu

Causality Spectrum and PCS

Mechanistic
Individual level

...

Average effect
Group level

Stable, replicable

Effect depends on the group
Stability implicit in causal
inference: e.g. SUTVA

PCS works towards causality:

Predictability + stability (+ computability)



interpretable hypothesis generation
recommendations for experiment

Comparing CI-ob with PCS-rec. system

- Similarities: both use observational data, with the ultimate goal of causal relationship discovery, and rely on human judgment calls
- Differences:

	CI-ob	PCS-rec
Model checking	no (?)	yes via P-screening
Causality evidence	assumptions based on domain knowledge	stability analyses to hopefully weed out confounding factors
Intervention experiment needed	no	yes

M. Behr




Y. Wang



X. Li



Provable Boolean interaction recovery from tree ensemble obtained via random forests

Merle Behr^{a,1} , Yu Wang^{a,1}, Xiao Li^a, and Bin Yu^{a,b,c,2}

- New Local Spiky Sparse (LSS) model: linear combination of Boolean interactions as regression function
$$E(Y|X) = \beta_0 + \sum_{k=1}^K \beta_k \prod_{j \in S_k} \mathbf{1}(X_j \leq \gamma_j)$$
- 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

Finding genetic drivers of HCM: a low SNR problem

Problem formulation

iRF recommendations

Gene-silencing (intervention) experiments

HCM data choices within UKBB

UK Biobank database: covariates SNPs

Use HCM labels in the database – **no signal found** (too many false negatives) – **a couple of months**

New phenotype response Left Ventricular Mass (LVM)

extracted by Weston from Ashley's lab from MRI images
– continuous variable



W. Hughes

No benchmark for noise level – so can't tell whether we are capturing any reality or passing “P”-screening in PCS

HCM: formulation choices



T. Tang

UK Biobank database: covariate SNPs, LVM response

Breakthrough via **binarization** - forced better signal to pass "P"

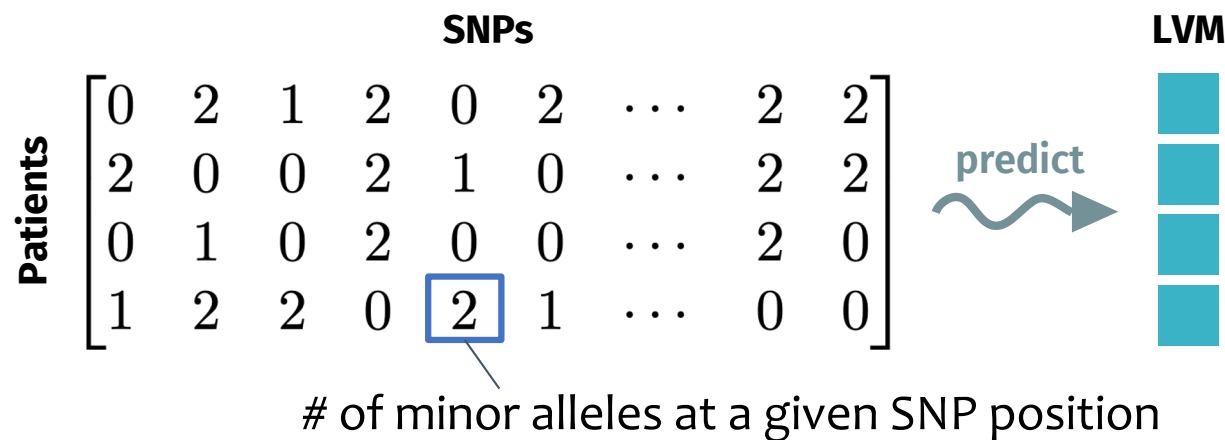
Binarized LVM with the top and bottom 20% into a balanced binary classification problem

↳ Stability analysis done using 15% and 25%, showing stable results

50% is a good benchmark on prediction error

Back to HCM: UK Biobank Data

n ~ 30K white British unrelated population with MRI data
p ~ 15 million imputed SNPs!!



Dimensionality reduction is necessary for our recommendations

Step 1: HCM gene/interaction recommendation pipeline



Dimension reduction



Fit iRF on *binarized* iLVM



Rank gene (interactions)

Step 1: HCM gene/interaction recommendation pipeline



Dimension reduction

- Run **GWAS** using BOLT-LMM and PLINK *and* ordinary linear regression
- Select union of top 1000 SNPs from each GWAS run: 1500 SNPs



Fit iRF on *binarized* iLVM



Rank gene (interactions)

Step 1: HCM gene/interaction recommendation pipeline



Dimension reduction

- Run **GWAS** using BOLT-LMM and PLINK and ordinary linear regression
- Select union of top 1000 SNPs from each GWAS run: 1500 SNPs



Fit iRF on *binarized* iLVM

- **Binarize** iLVM phenotype into high and low groups to “denoise” (using multiple thresholds: 15%, 20%, 25%)
- Fit **iRF** on SNP data to extract candidate gene interactions



Rank gene (interactions)

Step 1: HCM gene/interaction recommendation pipeline



Dimension reduction

- Run **GWAS** using BOLT-LMM and PLINK and ordinary linear regression
- Select union of top 1000 SNPs from each GWAS run: 1500 SNPs



Fit iRF on *binarized* iLVM

- **Binarize** iLVM phenotype into high and low groups to “denoise” (using multiple thresholds: 15%, 20%, 25%)
- Fit **iRF** on SNP data to extract candidate gene interactions



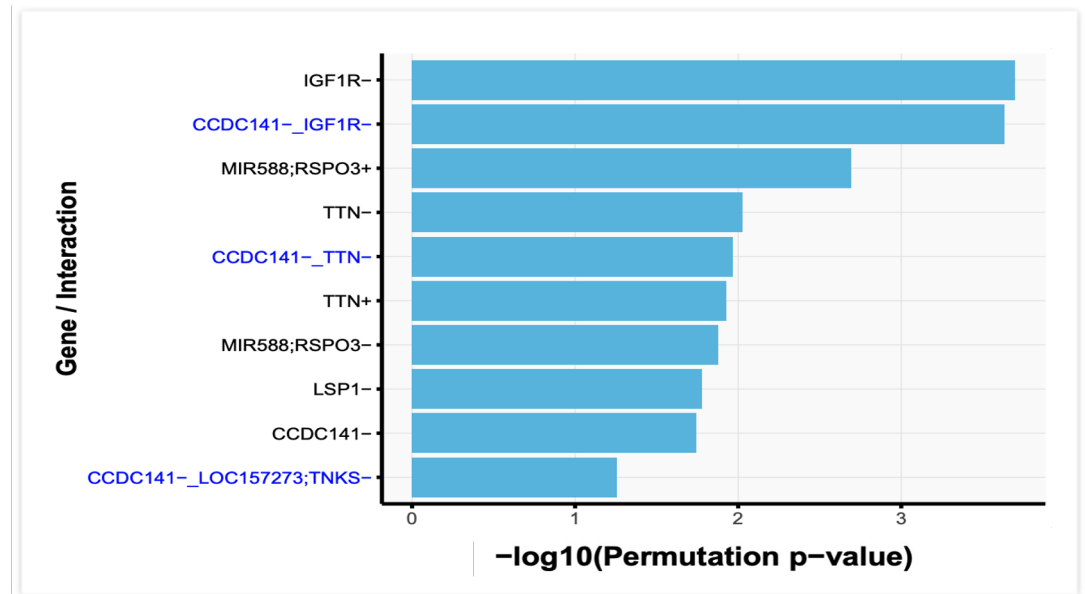
Rank gene (interactions)

- Using a **new stability-based importance score** to aggregate SNP-level importances from iRF into a **gene-level** score

Top LVM gene and interaction recommendations

Our iRF pipeline identifies

- **genes** (*TTN*, *IGF1R*) that are **well-known** to impact LVM,
- **promising candidate genes** (*CCDC141*, *RSPO3*, *LSP1*) that are known to be associated with the heart, and
- **interesting interactions** (*CCDC141-IGF1R*, *CCDC141-TTN*, *CCDC141-TNKS*)



* p-values averaged across multiple runs using different binarization thresholds (15%, 20%, 25%)



T. Tang



A. Agarwal



X. Li

Step 2.1 **Support to our findings by experts**

“Domain expert opinion solicitation with negative controls”

Three lists presented to cardiologists

1. Top-ranked findings
1. Mid-range ranked findings
1. Random findings

Collaboators (Chad and Euan) passed our test... :)



T. Tang



C. Weldy

Step 2.2: Biological evaluations using annotated databases on top genes

<u>Top Genes</u>	<u>Description / Supporting Evidence</u>
TTN	A well-known heart muscle gene that plays a key role in the sarcomere (a basic unit of muscle contraction)
IGF1R	A well-known gene that has been previously implicated in cardiac hypertrophy (i.e increased heart cell size)
CCDC141	A mostly unknown gene that neighbors TTN and is highly expressed in the heart
RSPO3	Wnt signaling gene, known to be associated with BMI and body height
LSP1	Known to associated with blood pressure and hypertension
TNKS	Wnt signaling gene, known to associated with hypertension, vascular and heart problems

* **Annotated databases searched:** Uniprot, gnomAD, Stanford Global Biobank Engine, GTEx, Tabula Muris



T. Tang



C. Weldy

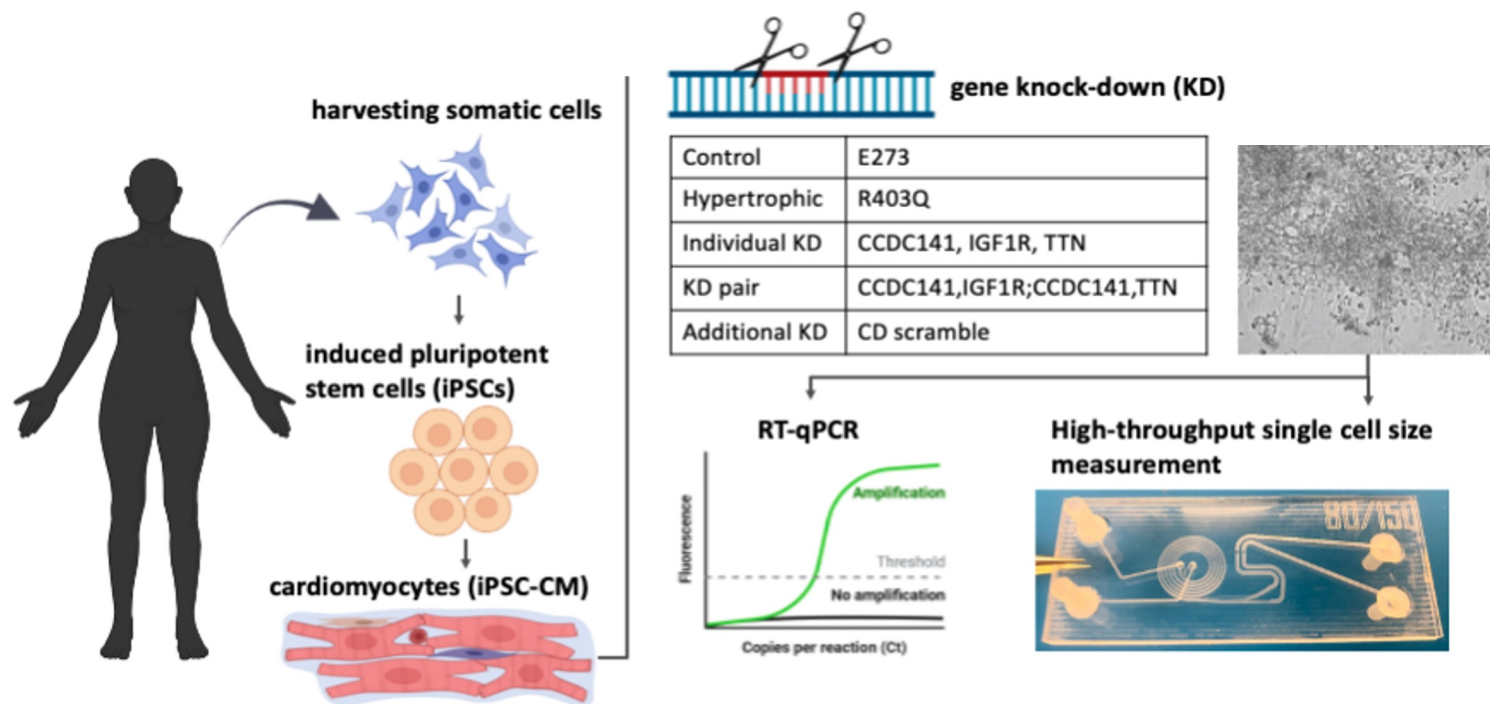
Step 3: Gene Silencing experiments



Qianru Wang



Nate Youlton



Stress-testing image processing pipeline with manual annotations



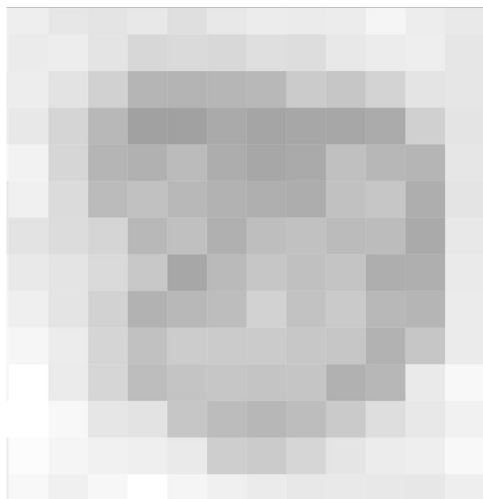
O. Ronen



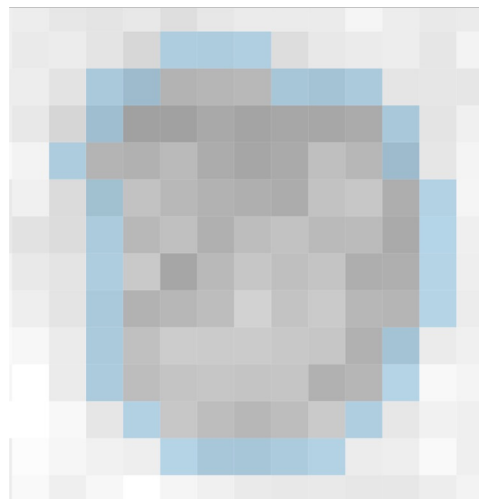
A. Kenney

We **manually annotated** 20 random images from the top outlet and 10 random images from the bottom outlet – **new experiments were run based on our investigations.**

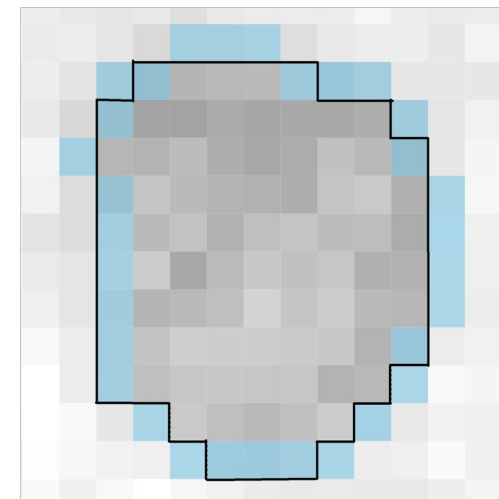
Original image



Algorithmic masking



Manual masking (in black)



Stress-testing revealed many challenges with the experimental data

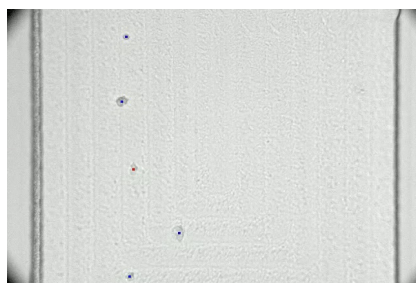


O. Ronen

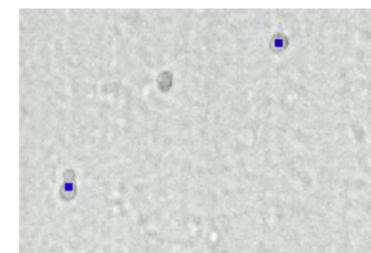
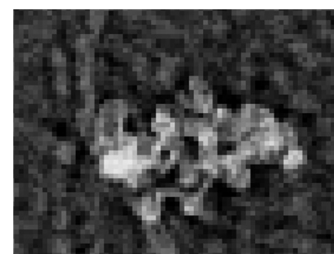


A. Kenney

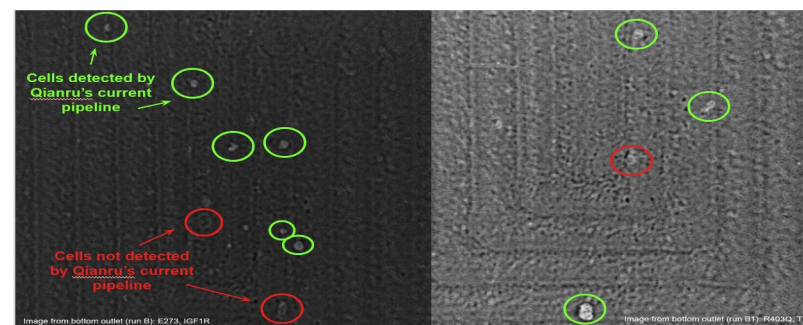
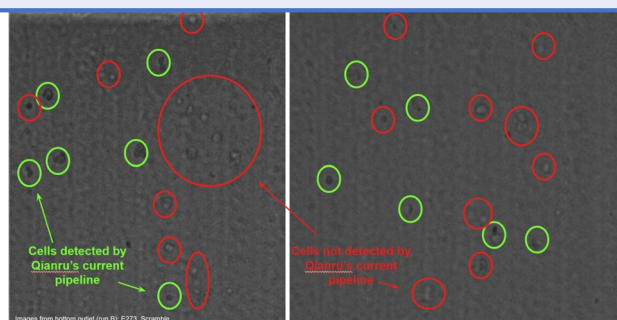
Repeated measurements



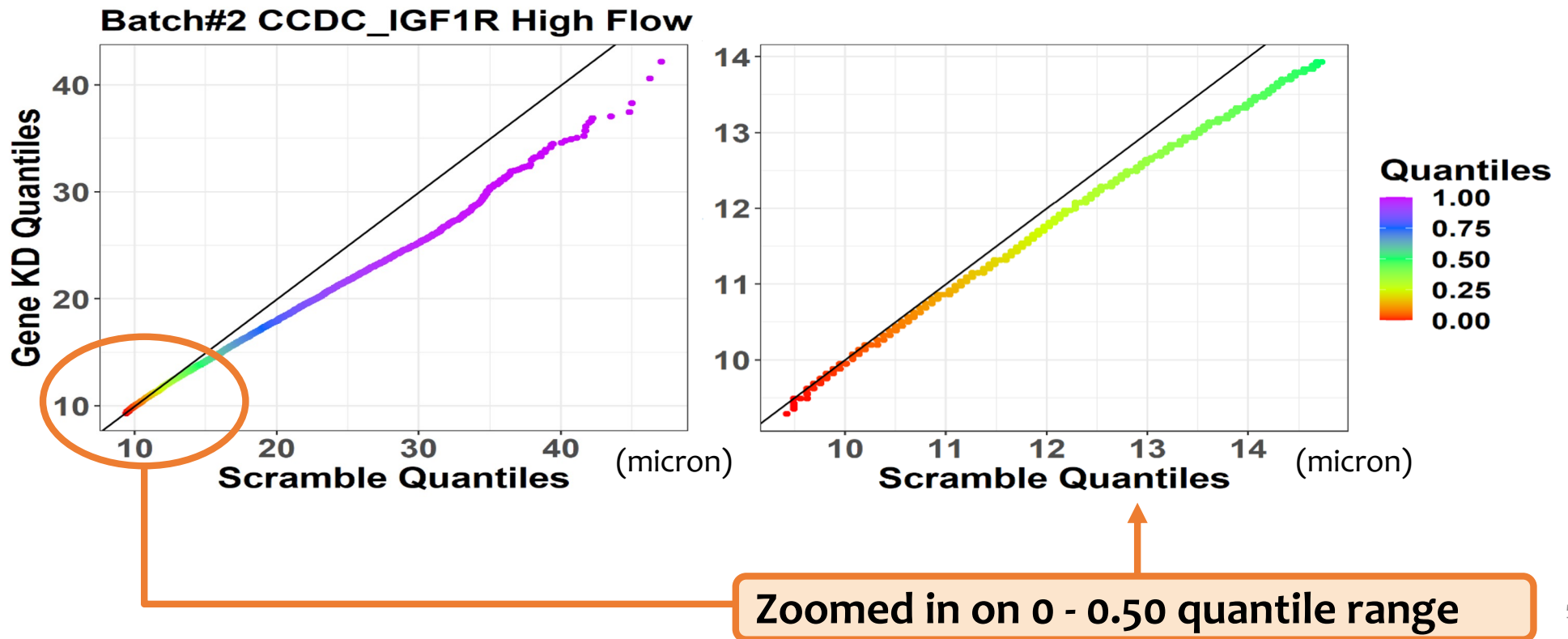
Clumps and fragments



Undetected cells in the Bottom outlet



Cell size comparisons revealed a consistent impact when silencing genes in an HCM cell line: **CCDC-IGF1R** interaction



Gene silencing experiment result summary

- 5 sets of experiment conducted and 4 found strong causal evidence
- Possible mechanistic explanation for found gene-gene epistatic interactions (CCDC141-IGF1R and CCDC141-TTN): **CCDC141 may interact with IGF1R and TTN through mediating transcription factor-DNA binding** (a huge amount of work here again)

Paper to be submitted to *Nature Medicine* with **Tiffany Tang** and **Qianru Wang** as first co-authors (and E. Ashley and me as senior co-authors), and many other collaborators.

Main co-authors:



Qianru Wang



Tiffany Tang



Euan Ashley

Biohub pipeline adhering to the PCS framework

Split 3 ways (random split)

Training (n = 15K) Validation (n = 5K) Test (n = 10K)

Domain-inspired dimension reduction to prioritize SNPs + binarize phenotype

Binarization: think like a scientist – why is one's heart much much bigger than another?

Prediction-check multiple methods

*siRF, RF, L2-regularized logistic,
L1-regularized logistic, SVM*

Very weak signal (siRF: ~55% balanced classification accuracy)

But siRF, on average, yielded the highest prediction performance (accuracy, AUROC, AUPRC) across all binarization thresholds, relative to other methods

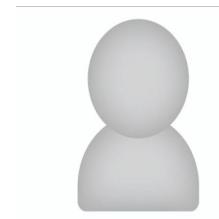
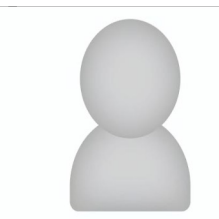
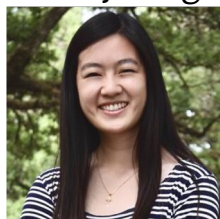
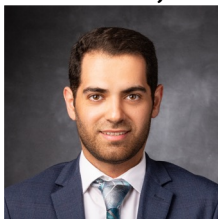
Stability check across binarization thresholds

Only recommend interactions that are stably important for all binarization thresholds

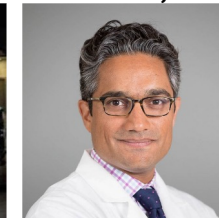
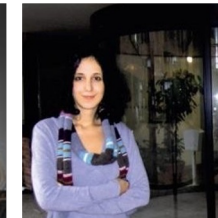
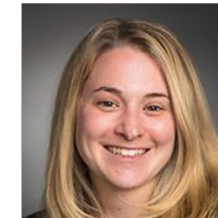
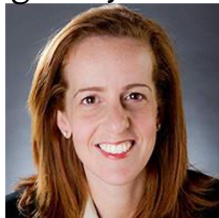
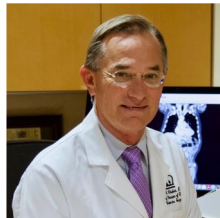
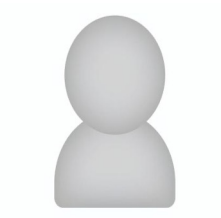
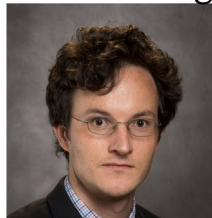
Validate prioritized interactions via **wet-lab gene-silencing experiments**

Another PCS case study:
pancreatic cancer risk prediction

Ehsan Irajizad Ana Kenney Tiffany Tang Jody Vykoukal Ranran Wu Eunice Murage Jennifer Dennison Marta Sans



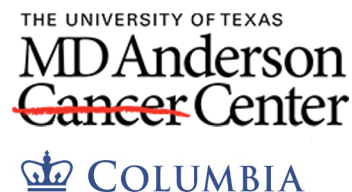
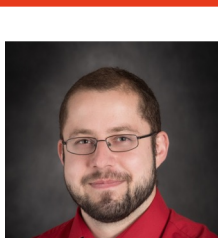
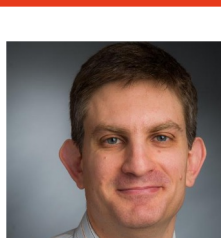
James P. Long Maureen Lottus John A. Chabot Michael D. Kluger Fay Kastrinos Lauren Brais Ana Babic Kunal Jajoo



Linda S. Lee Thomas E. Clancy Kimmie Ng Andrea Bullock Jeanine Genkinger Anirban Maitra Kim-Anh Do



Bin Brian M. Wolpin San Harash Johannes F. Fahrman



Constructing a metabolite panel predicting pancreatic cancer risk following the PCS framework

Cell Reports Medicine



Article

A blood-based metabolomic signature predictive of risk for pancreatic cancer

Ehsan Irajizad,^{1,2} Ana Kenney,³ Tiffany Tang,³ Jody Vykoukal,² Ranran Wu,² Eunice Murage,² Jennifer B. Dennison,² Marta Sans,² James P. Long,¹ Maureen Loftus,⁴ John A. Chabot,² Michael D. Kluger,² Fay Kastrinos,^{6,9} Lauren Brais,⁴ Ana Babic,⁴ Kunal Jajoo,⁵ Linda S. Lee,⁵ Thomas E. Clancy,² Kimmie Ng,⁴ Andrea Bullock,⁷ Jeanine M. Genkinger,^{9,10} Anirban Maitra,¹¹ Kim-Anh Do,¹ Bin Yu,³ Brian M. Wolpin,⁴ Sam Hanash,^{2,12,*} and Johannes F. Fahrmann^{2,12,13,*}

¹Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

²Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

³Department of Statistics, University of California, Berkeley, CA, USA

⁴Dana-Farber Brigham and Women's Cancer Center, Division of Gastrointestinal Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

⁵Division of Gastroenterology, Hepatology and Endoscopy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

⁶Dana-Farber Brigham and Women's Cancer Center, Division of Surgical Oncology, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

⁷Division of Hematology/Oncology, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

⁸Division of Digestive and Liver Diseases, Columbia University Irving Medical Center and the Vagelos College of Physicians and Surgeons, New York, NY, USA

⁹Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA

¹⁰Department of Epidemiology, Columbia Mailman School of Public Health, New York, NY, USA

¹¹Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

¹²These authors contributed equally

¹³Lead contact

*Correspondence: shanash@mdanderson.org (S.H.), jfahrmann@mdanderson.org (J.F.F.)

<https://doi.org/10.1016/j.xcrm.2023.101194>

Recently published in **Cell Reports Medicine**, with an **editorial** on our paper.

Cell Reports Medicine



Preview

Where the metabolome meets the microbiome for pancreatic cancer detection

Lucy Oldfield¹ and Eithne Costello^{1,*}

¹Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, UK

*Correspondence: ecostell@liverpool.ac.uk

<https://doi.org/10.1016/j.xcrm.2023.101011>

Risk prediction tools for pancreatic cancer are urgently sought to facilitate screening. Irajizad et al.¹ describe the performance of a risk prediction model based on circulating microbial- and non-microbial metabolites for assessment of 5-year pancreatic cancer risk.

Reviewer #1: "I consider this to be the best of the many papers exploring metabolites as cancer predictors that I've reviewed."

Analysis pipeline adhering to the PCS framework

Split 3 ways (group-based split)

Training (5 centers) Validation (2 centers) Test (3 centers)

Use domain knowledge to prioritize microbiome

14 microbial-related metabolites analyzed

Prediction check multiple methods

*Logistic, logistic with L2, logistic with L1,
iRF, deep learning, GBM, auto ML*

Microbial-panel: logistic regression with L1 (3 metabolites selected)

Non-microbial-panel: logistic regression (all 5 metabolites)

Stability check on perturbations and subgroups

Microbial-panel stable!

Validate microbial-panel on test set and independent/external cohort.
Combine with the non-microbial-panel for final results

Takeaways

- The panel with 3 microbiome-related metabolites is predictive of cancer risk across a **test set and independent cohort**
- The model choice consideration via prediction-check step **improves prediction by 6%**
- Stability-checks demonstrate consistency across perturbations and subgroups. Importantly, **diabetes status did not confound** the panel with microbiome-related metabolites
- The full metabolite panel (3 microbial-related and 5 non-microbial related) **improves on the standard CA19-9 marker** for risk prediction by **18%**

Current directions of PCS

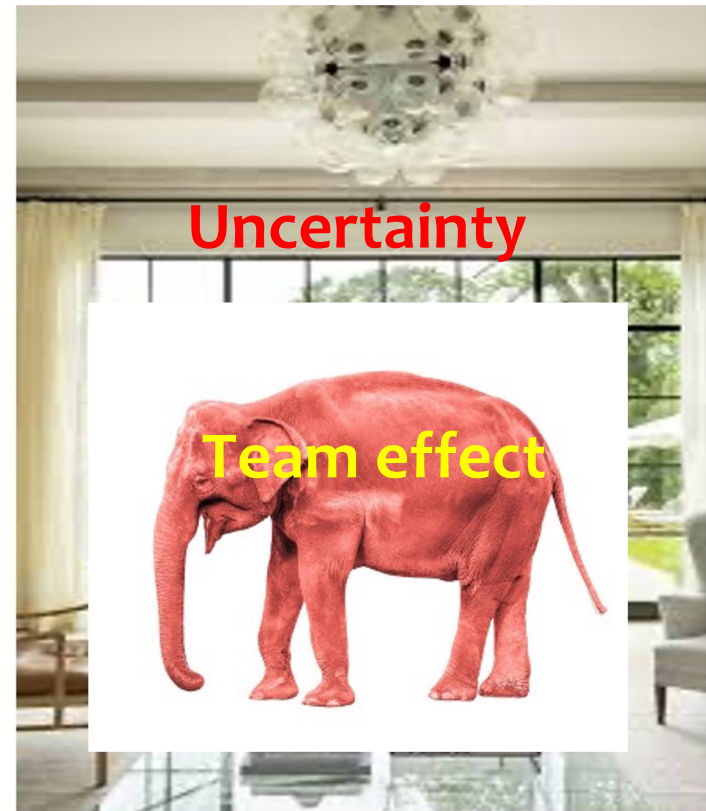
PCS Inference

Software: Veridical Flow (v-flow), simChef

Documentation template

Extensions by others

In our house of “uncertainty”



<https://www.housebeautiful.com/room-decorating/living-family-rooms/g715/designer-living-rooms/>

<https://www.forbes.com/sites/janicegassam/2019/11/27/the-pink-elephant-in-the-workplace-how-to-have-conversations-about-race-politics-and-religion-at-work/>

PCS recommendation on data cleaning

Multiverses analysis (Steegen et al, 2016):

keep multiple copies of cleaned/processed data
(at least two)

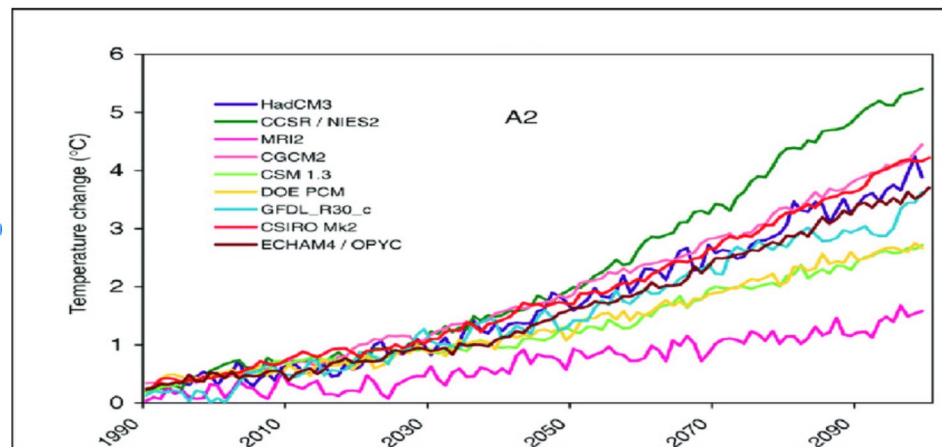
Document as much as one can to vet human decisions.

S. Steegen, F. Tuerlinckx, A. Gelman, and W. Wanpaemel (2016): Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.*

Global mean temp. change: 9 model choices

Researcher to researcher (or team to team) perturbation has been addressed by climate scientists.

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

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 **trustworthy** 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.

To meet this challenge, PCS inference requires model checking or reality check under “P” and accounts for important sources of uncertainty under “S” with documentation.

PCS Inference I (Yu and Kumbier, 2020)

Predictability: Use prediction error (and domain knowledge) for model checking or reality check.

Stability: Assessed across data and model perturbations through different in-context aggregation methods (e.g. worse case or perturbation interval) (with data perturbation broadly interpreted)

Computability: Implicitly required by P and S

Includes data-inspired simulation to assess coverage (on-going)

PCS Inference II (Yu and Kumbier, 2020)

Prediction perturbation interval (Yu and Barter, 2024): it formally takes into account two more uncertainty sources from data cleaning method and model/algorithm choices; coverage assessed in validation set and adjusted to the correct level by multiplier of length of interval.

Parameter perturbation interval (on-going): it not assume a probabilistic generative model -- similar to bootstrap-based inference when the probabilistic model approximates reality well, while taking into account additional uncertainty sources and using multiple vetted data-driven simulation models.

Software to address “C” in PCS



Veridical Flow: (v-flow) PCS-style data analysis made easy!



A. Agarwal



J. Duncan



R. Kapoor



C. Singh



simChef: PCS-style simulations made easy!



J. Duncan



C. F. Elliott



T. Tang



M. Behr



K. Kumbier

PCS documentation



T. Tang

A. Kenney

Template at my website: <https://yu-group.github.io/vdocs/PCSDoc-Template.html>

1 Domain problem formulation

2 Data

3 Prediction Modeling

4 Main Results

5 Post hoc analysis

6 Conclusions

1 Domain problem formulation

What is the real-world question? This could be hypothesis-driven or discovery-based. ⓘ

This should be very high level, providing the big picture behind the study. Often this takes the form of a pre-existing hypothesis (e.g., individuals with a specific genetic mutation are more likely to have a given characteristic) or more open-ended discovery (e.g., identify mutations that are related to a given characteristic).

Insert narrative here.

↩ ↪ T ¶ +

Why is this question interesting and important? What are the implications of better understanding this data? ⓘ

Summary: Multi-roles of PCS, and expanding

- **Internal validity** with prediction-checks, extensive stability-checks, and expanded uncertainty
- **Recommendation** for **external causality validation**, when combined with domain knowledge
- **Evaluation or stress-test** of existing data driven procedures, e.g. clinical decision rules
- **New Stats/ML/DS algorithms** by adding stability to “unstable” algorithms
- **Extensions** to veridical spatial data science, veridical network analysis, and reinforcement learning by others. Beginning theory work...

<https://binyu.stat.berkeley.edu> for PCS related papers, software, and doc. template

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

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)”**, <https://arxiv.org/abs/1810.07287>.

M. Behr, Y. Wang, X. Li and B. Yu (2022). **“Provable Boolean Interaction Recovery from Tree Ensemble obtained via Random Forests”**, PNAS.

E. Irajizad, A. Kenny, T. Tang, ..., B. Yu, B. Wolpin, S. Hanash, J. Fahmann (2023). **A blood-based metabolomic signature predictive of risk for pancreatic cancer**. Cell Reports Medicine (with an editorial).

Q. Wang, T. Tang, ..., B. Yu, and E. Ashley (2023). **“Epistasis regulates genetic control of cardiac hypertrophy”**. (to be submitted soon)

Other works: ESCV, staNMF, staDISC, staDRIP, DeepTune, ...

Softwares: v-flow, simChef, iRF, siRF, epiTree, ... **imodels (tree-based methods) also on AWS AutoGluon**

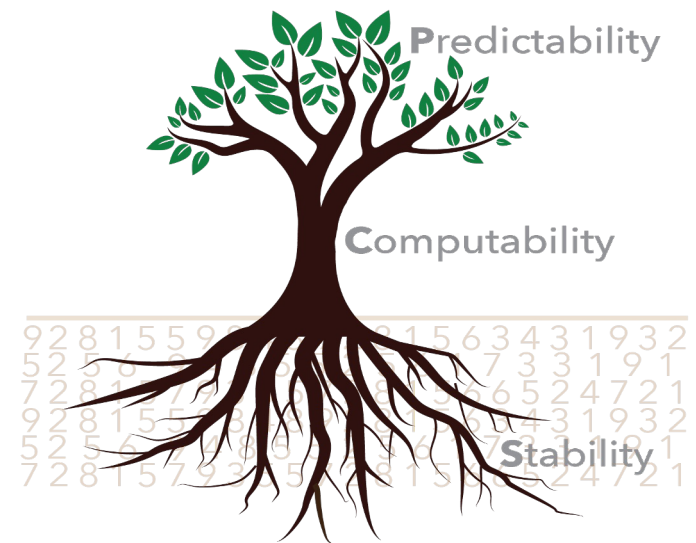
My hope

More adoption of **PCS** in the stats/ML and data science communities, with relevant and insightful theory

“At the dawning of the AI era, it is extremely valuable to have a uniform language and framework to talk about stress tests in the analysis pipeline.”

– Dr. Ehsan Irajizad, MD Anderson

Veridical Data Science



Berkeley-Stanford Joint Workshop on Veridical Data Science

BIDS, UC Berkeley, May 31, 2024

Bin Yu and Rebecca Barter (MIT Press) (based on PCS for DSLC) free online copy (Feb, 2024), hard copy (2024)

Veridical Data Science: A Book

Bin Yu^{1,2} and Rebecca Barter¹

¹Department of Statistics, UC Berkeley

²Department 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

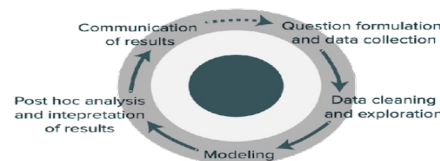
Preparing explanatory visual and numeric summaries for explaining data and findings to an external audience

Written reports

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

Core guiding principles for the book

The DS Lifecycle



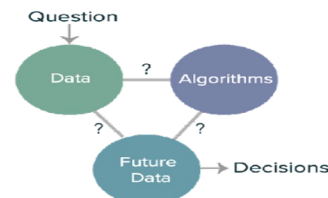
The Data Science Lifecycle is an iterative process that takes the analyst from problem formulation, data cleaning, exploration, algorithmic analysis, and finally to obtaining a verifiable solution that can be used for future decision-making.

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.

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.

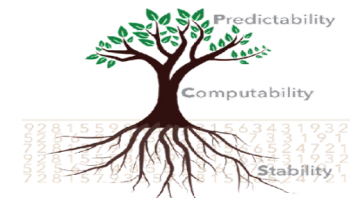
Three realms



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.

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 all!

<https://www.qut.edu.au/research/why-qut><https://www.qut.edu.au/research/why-qut>

