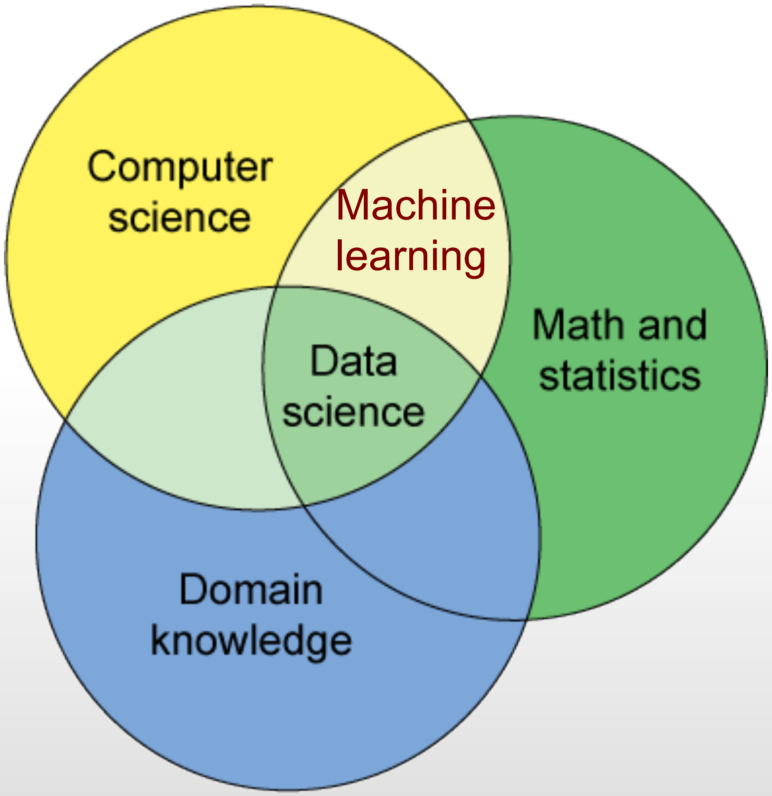
Three Principles of Data Science: predictability, computability, stability

Bin Yu
Statistics and EECS, UC Berkeley

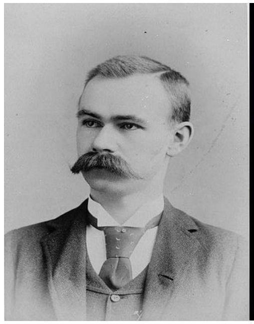
SIAM Conference on Uncertainty Quantification
Garden Grove, CA
April 18, 2018

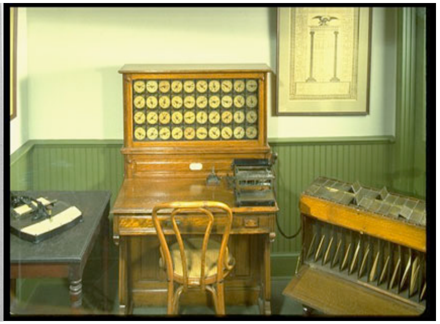
What is data science?

Conway's Venn Diagram



Statistician, Inventor **H. Hollerith**





1890's Hollerith Tabulating Machine

Founding father of modern statistics and statistical genetics, **R. A. Fisher**

Data science is the re-merging of **computational** and **statistical** thinking in the context of domain problems

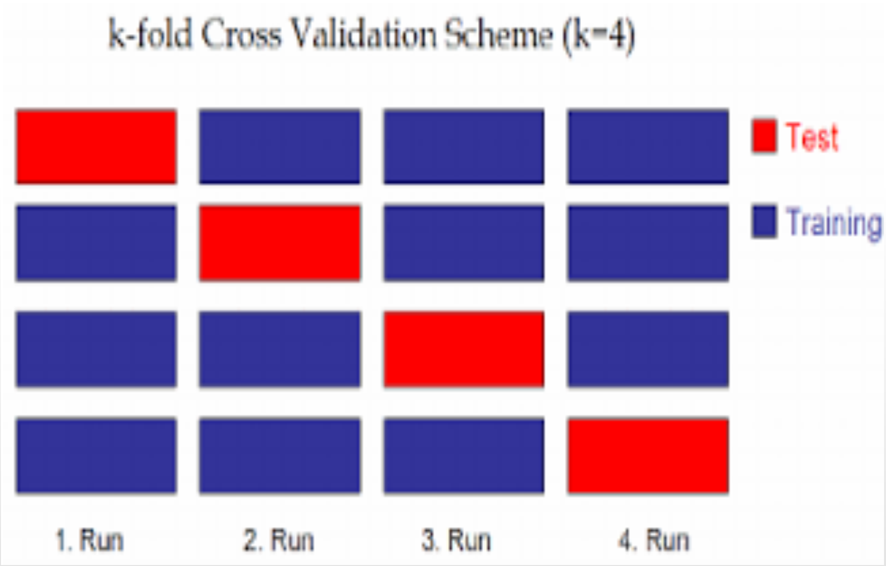
Machine learning (ML): part of statistics and CS

Prediction: part of statistics that also invented Cross-validation (CV) in the 70's.

First generation ML: **prediction** + **optimization**, with a heavy use of CV

Cross-validation (CV):
to estimate prediction error within one data set

Given a prediction problem with an "exchangeable" data set, CV creates k "pseudo-replicated" prediction problems:



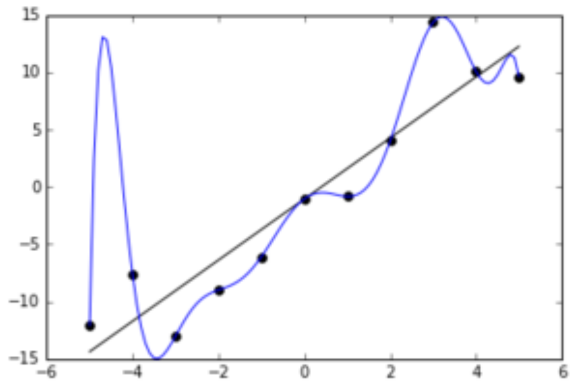
CV prediction error is the average over k-fold (not always a good estimate of the pred. error)

Reasons for ML success

- Prediction and cross validation are both natural and simple conceptually
- Data availability
- Computing resource availability
- Open-source software

ML/Stats Frontier: interpretation

CV avoids over-fitting for prediction purposes



CV can result in over-fitting for explanation purpose

EU's General Data Protection Regulation (2016) gives a “right” to explanation, and demands ML/Stats algorithms to be

human interpretable

Data Science Challenges

- Organize and develop ML/Stats/DS knowledge through first principles that take advantage of computing resources to increase accessibility and impacts

- Integrate better ML/Stats and other approaches not necessarily probabilistic to solve complex data problems

Guiding principles for data-intensive science

"Embedded" students/postdocs work on site,
in the wet lab

Seed scientific problem(s)

↓

Generalization

Generalization: workflow, algorithms, theory

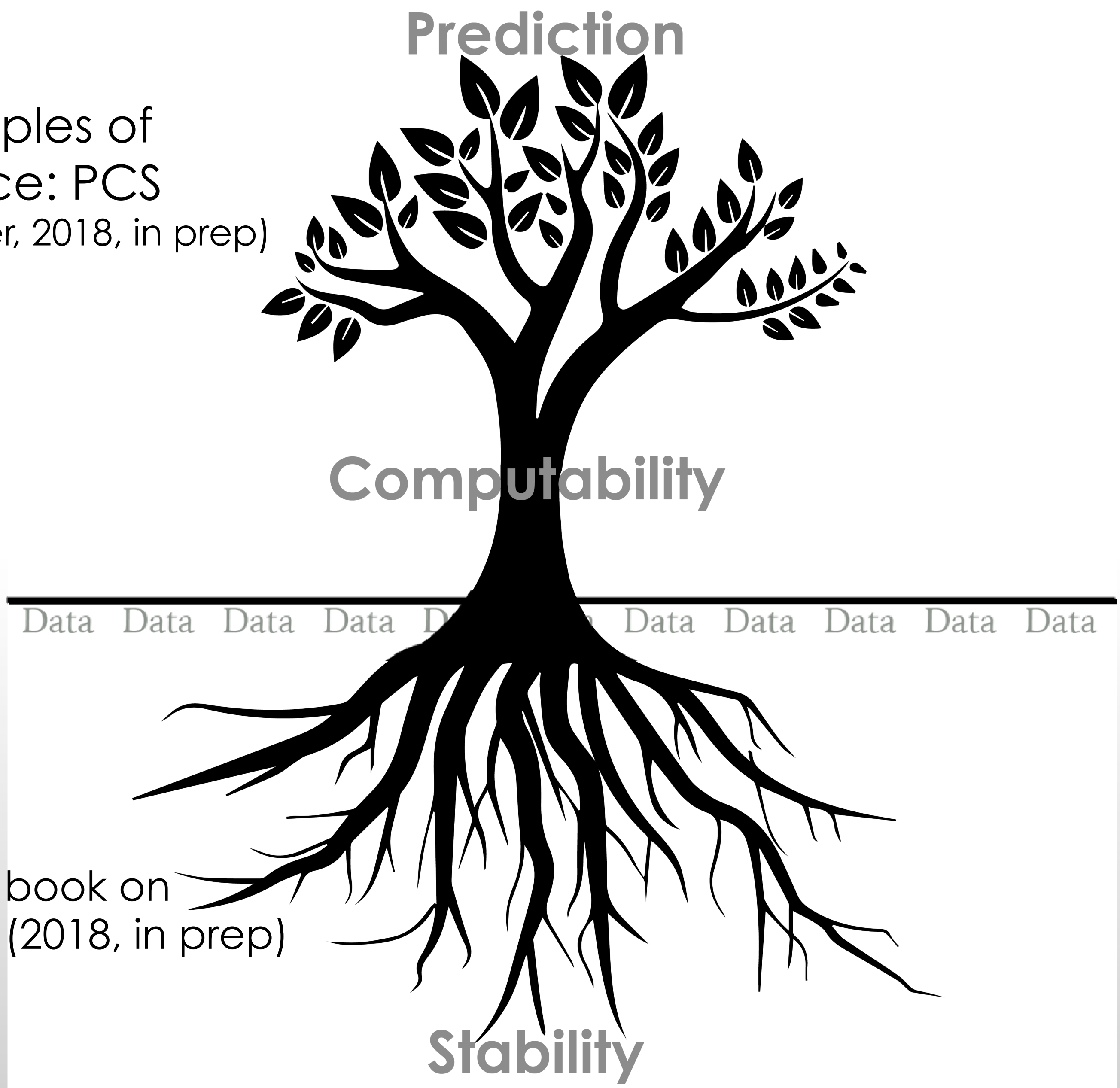
Current Framework: PCS workflow
(**PCS**=**P**redictability, **C**omputability, and **S**tability)

- Build on top of machine learning to have **predictability** as a first base - **check for reality**

- **Computability** as the second base

- **Stability** as the third base as a minimum requirement for **reproducibility** and **interpretability**, and as an extension of **uncertainty** assessment (stat inference)
  …

Three principles of
data science: PCS
(Y. and Kumbier, 2018, in prep)



Y. and Barter: book on
data science (2018, in prep)

Rebecca Barter

Stability of Knowledge

"That is why knowledge is prized higher than correct opinion, and **knowledge** differs from correct opinion **in being tied down**. . ."

-- Plato in *Meno*

Bernoulli **19**(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

A platform to integrate a myriad of works in the literature and to develop new methods …

It is a minimum requirement for reproducibility and interpretability, and intervention experiment design.

Stability Principle

Application of **Stability Principle** needs clearly defined

1. **Target(s) of interest**

2. **Appropriate perturbation(s)** to inputs to the DS cycle, including to **pre-processing** methods, **EDA**, **data** and/or **models/algorithms,** and **ad-hoc human decisions**

3. **Stability measure(s) on the target(s) after perturbation**

Appropriateness of perturbations and stability metrics is determined based on subject knowledge, experience, judgment, and data collection process, resource, regulation, interpretability, …

“Stability Principle” in the literature

Algorithmic stability: Devroye and Wagner (1979), Kearns and Ron (1999), Bousquet and Elisseeff (2002), Kutin and Niyogi (2002), Mukherjee et al (2006)….

Model selection: Stone (1973), Allen (1973), Shao (1995), Breiman (1996), …

Sensitivity analysis in Bayesian modeling: Box (1980), Berger (1984), Smith (1984), …

Causal inference: Leamer (1982), Athey and Imbens (2015), Ding and VanderWeele (2015), …

Lasso or sparse modeling: Bach (2008), Meinshausen and Bühlmann (2010), Liu, Roeder and Wasserman (2010), Haury et al (2011), Li et al (2011)…

Clustering: Meilia (2006), von Luxburg (2010), Bubeck (2012),…

Differential privacy: Dwork (2006), Dwork et al (2015),…

…

Many UQ considerations seem to be stability considerations…

Stability is as fundamental as predictability

- Stability deals with perturbations well beyond sampling perturbations – it embodies general robustness

- It allows us to go beyond “true” distribution postulation

- CLT and other limiting results are stability results

Examples of **data** perturbation

- **Cross-validation partition**
- Bootstrap
- Subsampling
- Adding small amount of noise to data
- Bootstrapping residuals in linear regression and liner time series models
- Block-bootstrap
- *Data perturbations through mechanistic simulation models
- *Adversarial examples in deep learning
- …

Examples of **model/algorithm** perturbation

- Robust statistics models
- Semi-parametric models
- **Lasso and Ridge models**
- Different modes of a non-convex empirical minimization
- **Different versions of Deep Learning algorithms**
- Different kernel machines
- Sensitivity analysis of Bayesian modeling
- …

Causality evidence spectrum

Mechanistic Individual level	…	Average effect Group level
Stable, replicable		Effect depends on the group Stability implicit in causal inference: e.g. SUTVA

PCS workflow: Prediction + stability (+ computability)

↓

interpretation + intervention design

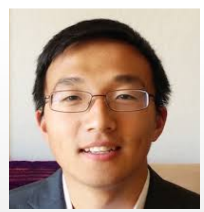
First example of PCS

Deep nets meet real neurons: transfer learning and neuron functions

Abbasi-Asl, Chen, Bloniarz, Oliver, Willmore, Gallant, and Y. (in prep, 2018)

Culmination of 3+ years of work

Reza Abbasi-Asl

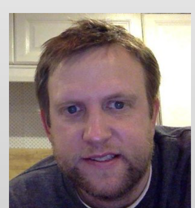


Yuansi Chen

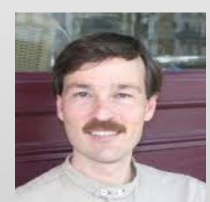
Adam Bloniarz

In collaboration with

Mike Oliver



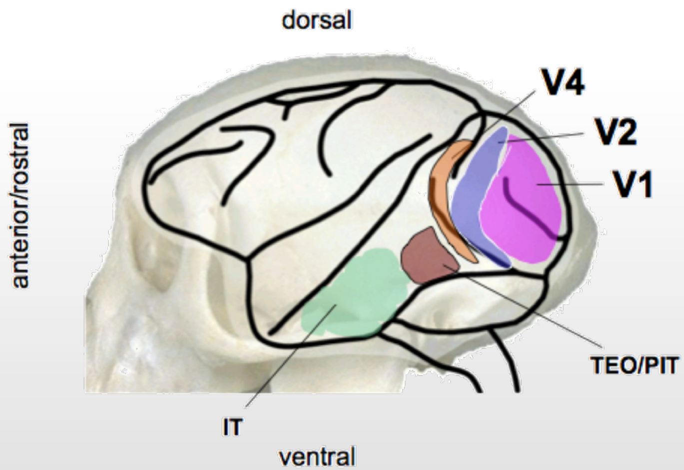
Ben Willmore



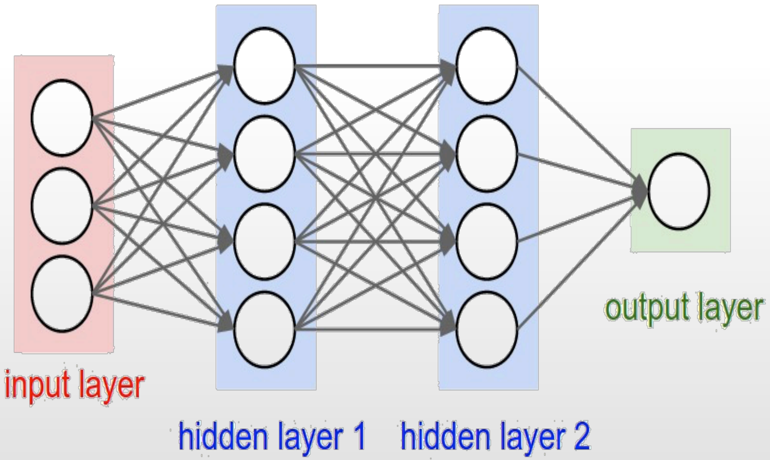
Jack Gallant

Interface between Neuroscience and Deep Learning

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



- Deep convolutional neural networks

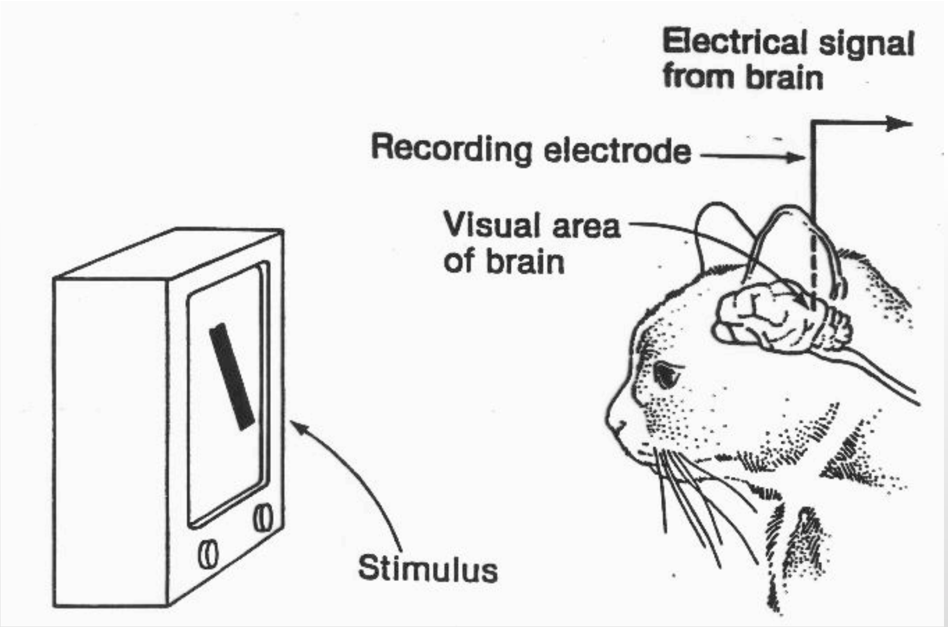


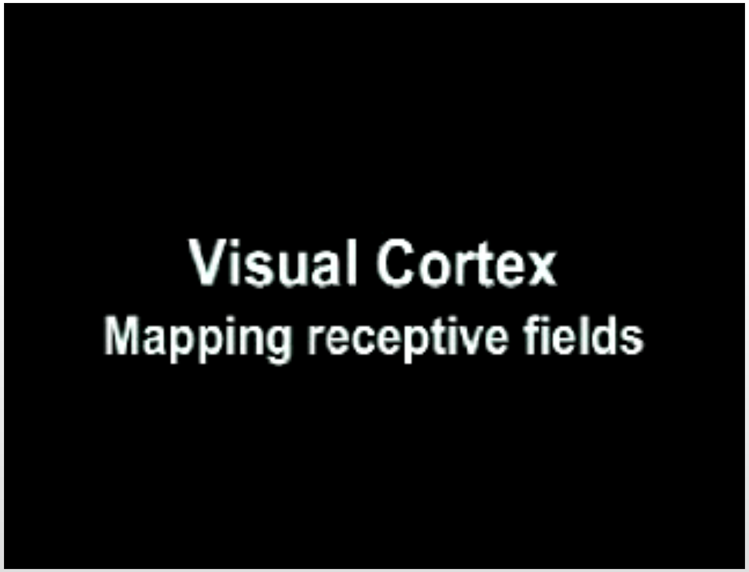
http://cs231n.github.io/assets/nn1/neural_net2.jpeg

V1 decoded by Hubel and Wiesel (1959)

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

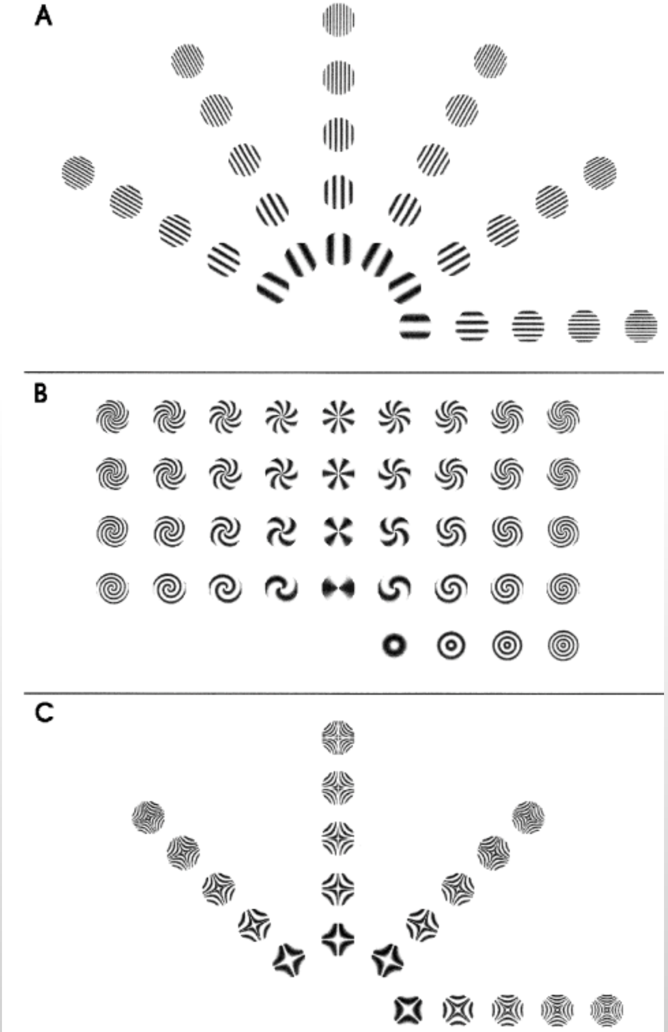
Nobel Prize in 1981





V4: synthetic polar and hyperbolic gratings and complex shape stimulus

Gallant et al. 1993, 1996

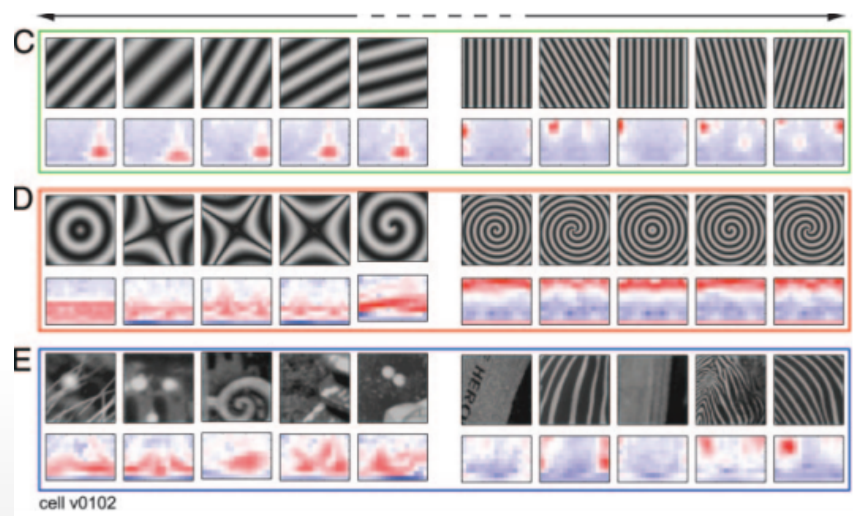


Cartesian

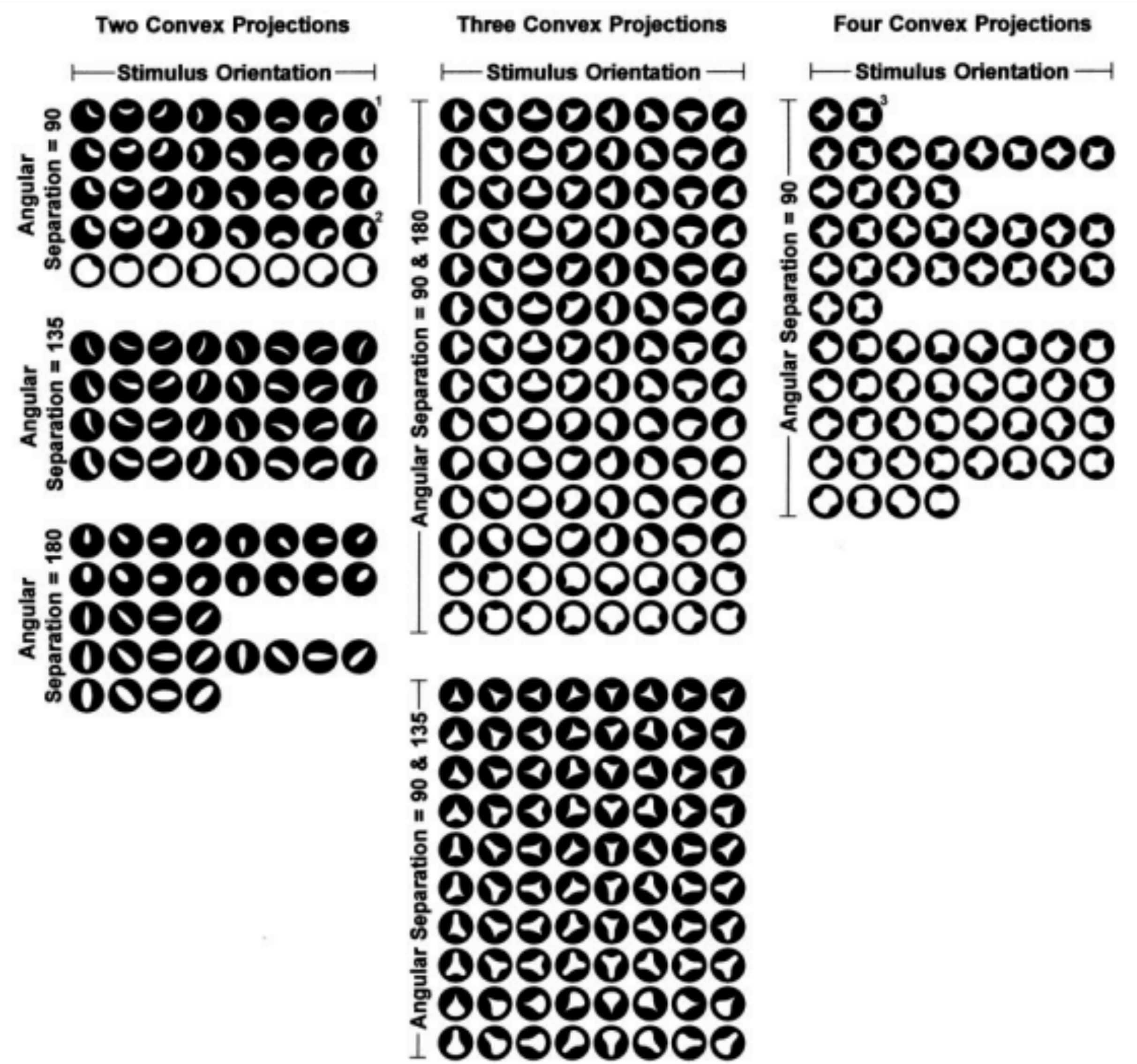
Polar

Hyperbolic

David et al (2006)



V4: synthetic convex and concave boundary stimulus



Pasupathy and Connor 1999, 2002

The stimuli were created by systematically combining convex and concave boundary elements.

Our data collection: 71 V4 neurons

(from the Gallant Lab at UC Berkeley)

Well-isolated visual neurons

Neuronal behavior is probed using sequences of natural images

Related works

Mairal et al (2013-18, in prep): earlier work from us

Parallel developments in the DiCarlo Lab at MIT :
Yamins et al (2014, 2016) and Cadieu et al (2014)
(**semi-natural** images, predictive modeling)

We replicate their predictive results and
aim at interpretation and understanding.

Questions to answer

What do **V4** neurons do?

How much do Convolutional Neural Networks (CNNs) resemble brain function?

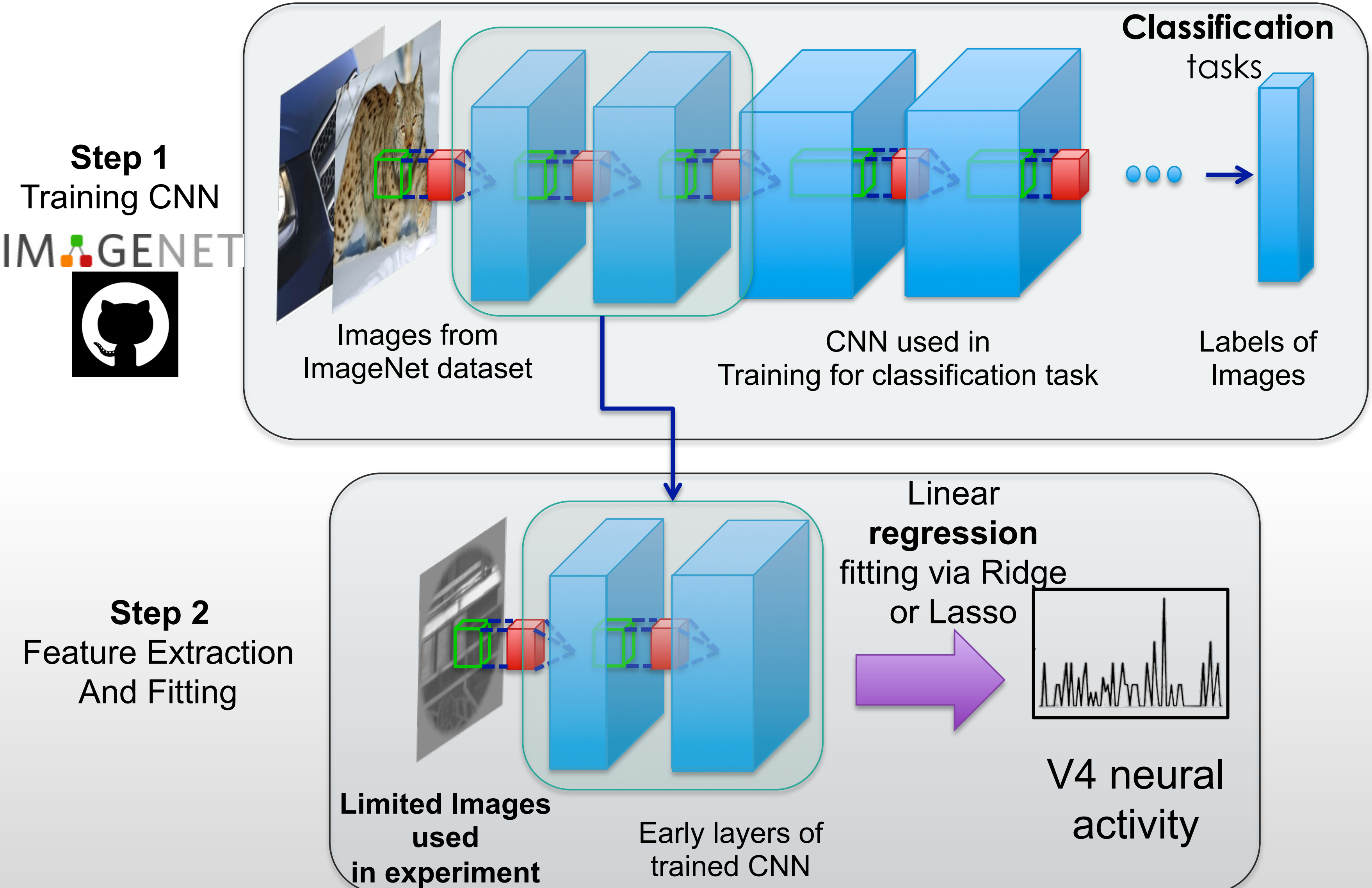
Our aims are two-fold

Transfer predictive learning to derive **state-of-art prediction** model for our V4 neurons

System neuroscience insights into neurons through **stable interpretation** of predictive models to suggest what V4 neurons do

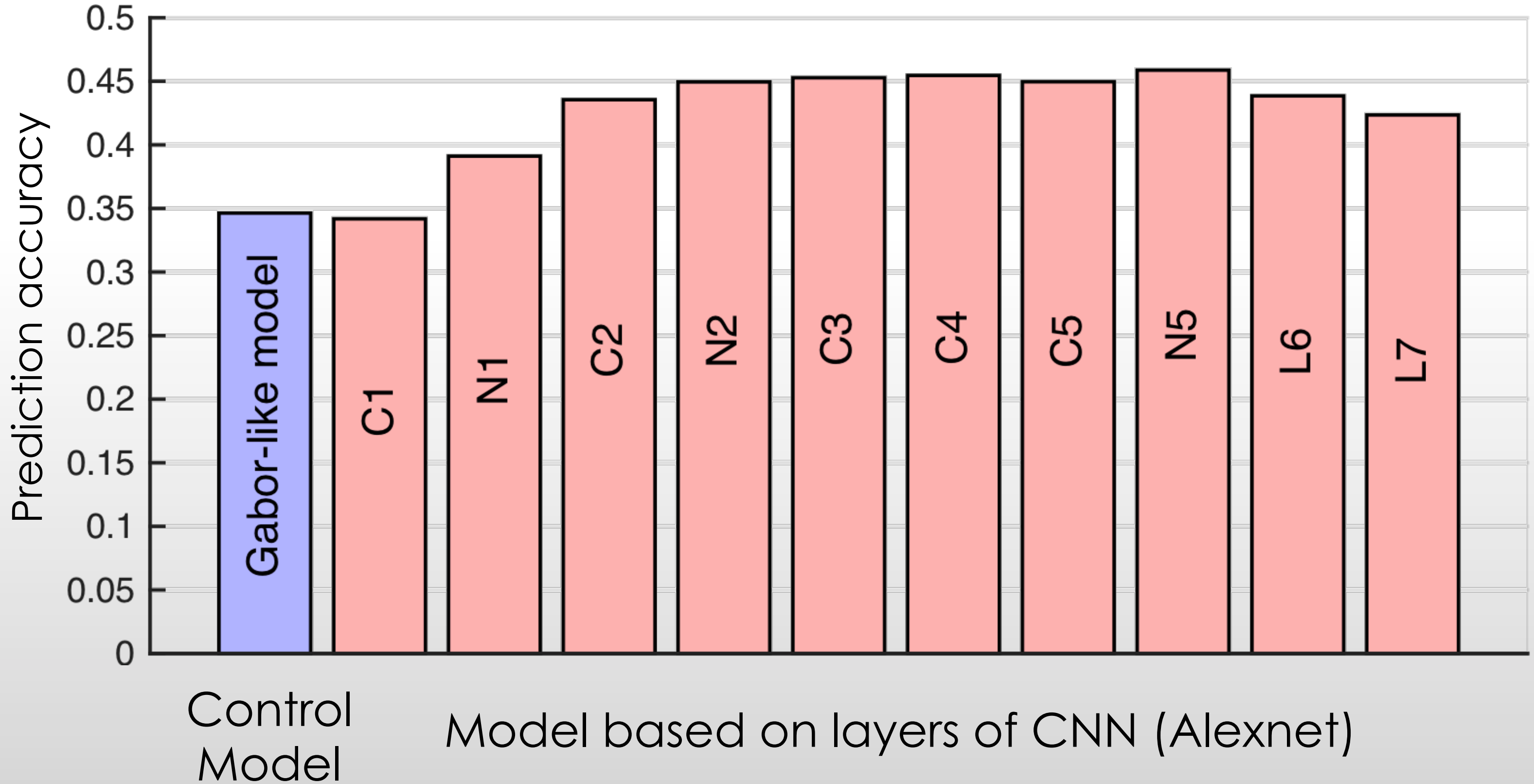
As a result, we provide some support for resemblance of CNNs to primate brain

Transfer learning…

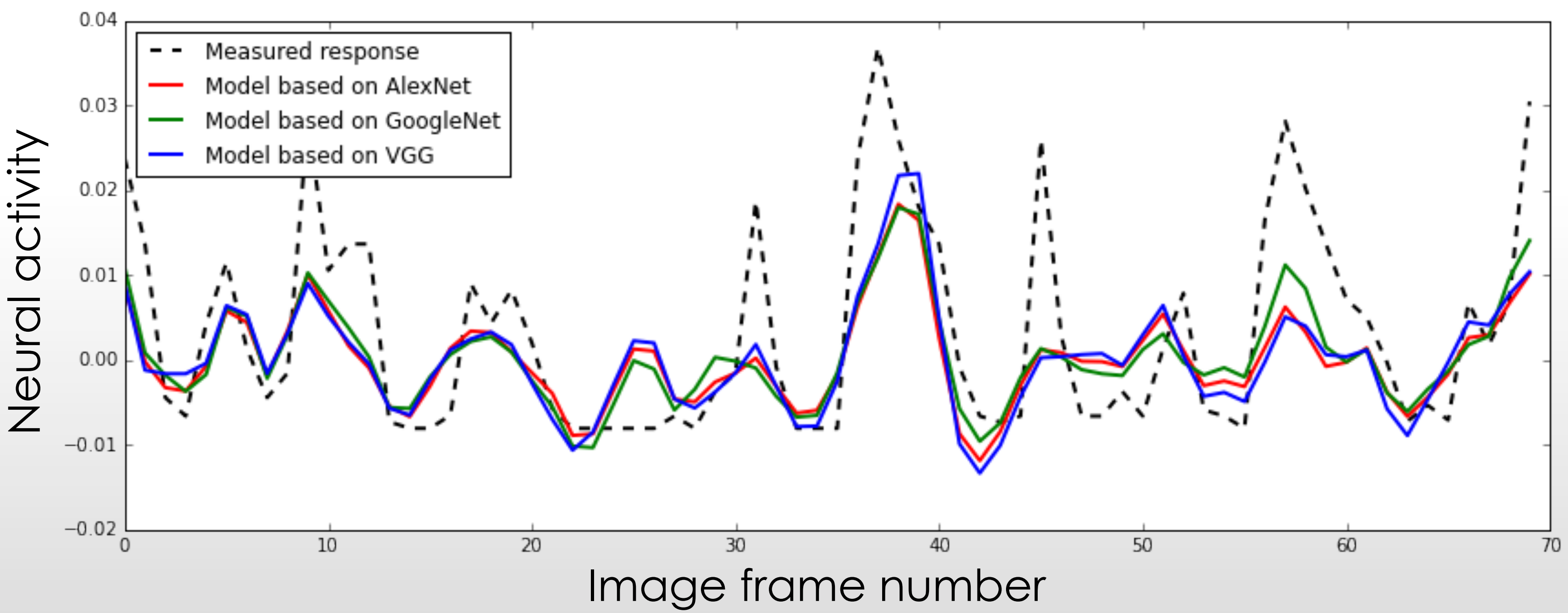


Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Prediction performance across different layers of CNN(AlexNet): N2 works well for V4



Stable predicted neuron activity from three deep nets +Lasso for a particular neuron



Deep nets meet real neurons

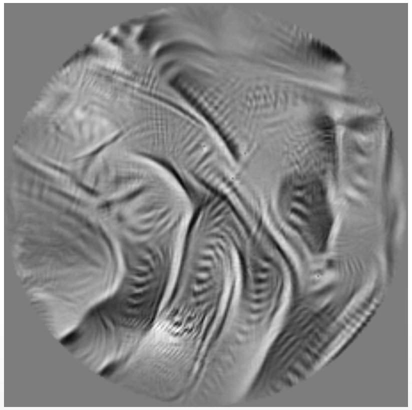
CNN (e.g. AlexNet) + regression gives state-of-art prediction for V4 neurons – 18 such models

Stability of excitatory images over 18 models and several compressed models provides testable (prescriptive) characterizations of V4 neurons

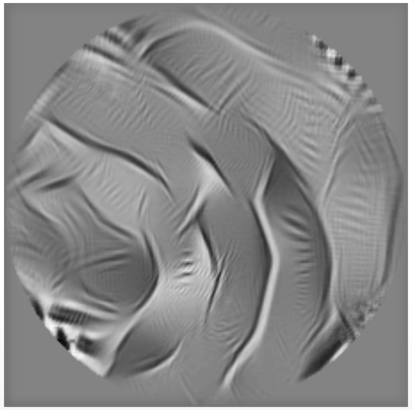
We combat "**model-hacking**" via "stability principle"

Neuron E
Excitatory patterns/images

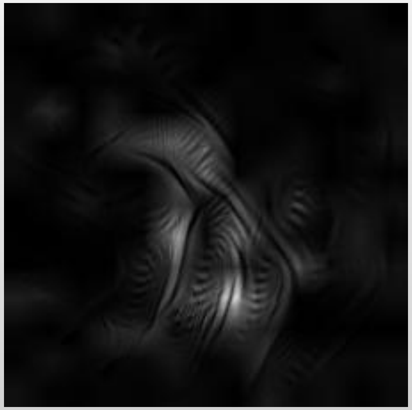
Lasso
CC = 0.63



Ridge
CC = 0.64

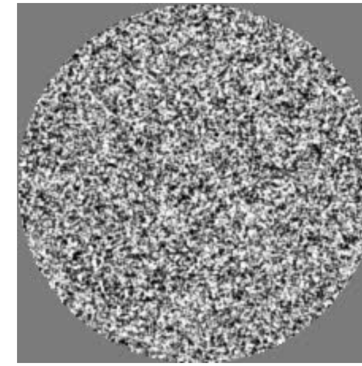


DeepTune
patterns/images
to characterize
Neuron E

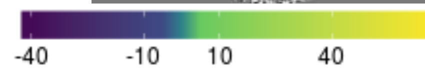


Masked DeepTune
patterns

Superheat plot of **DeepTune** optimization process

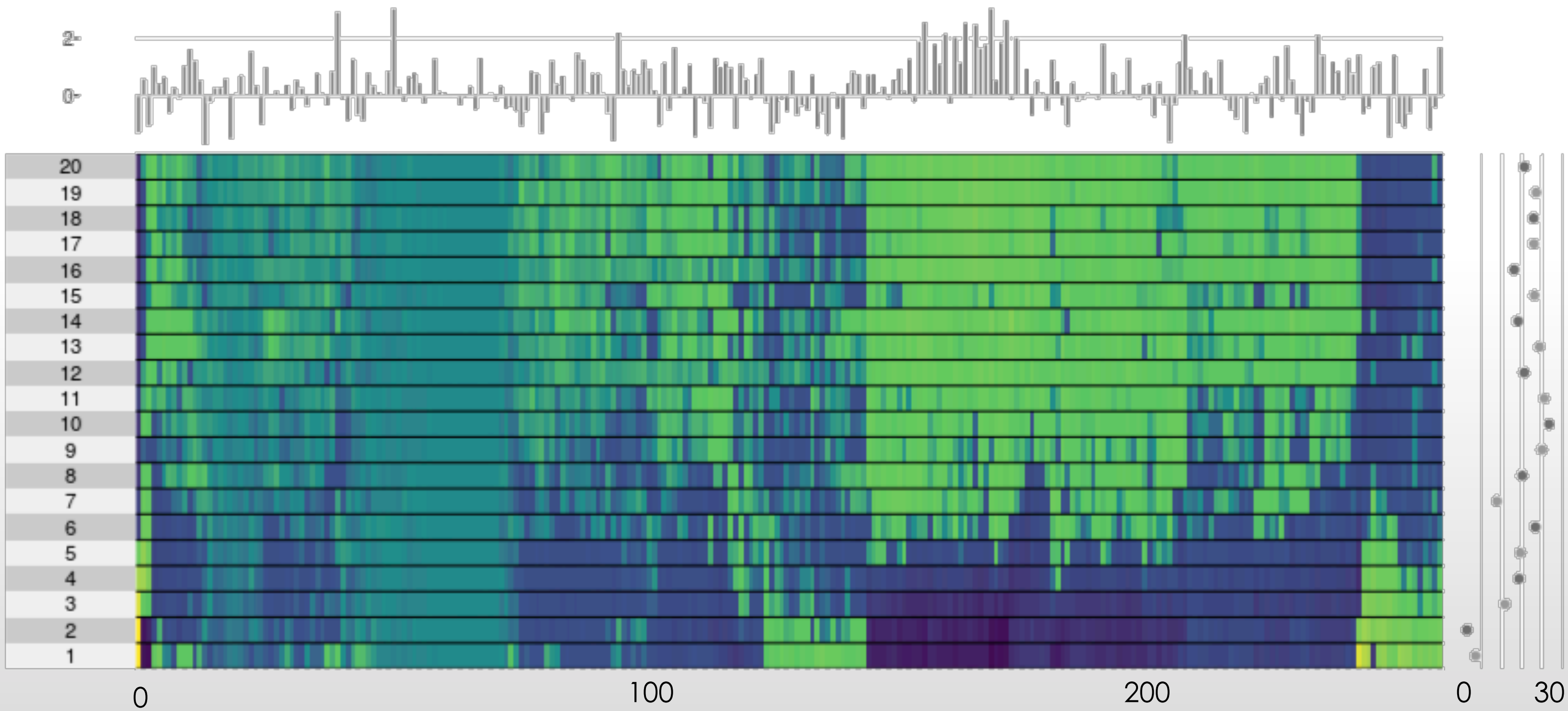


Neuron E



Regression coefficients

Deep Dream iteration



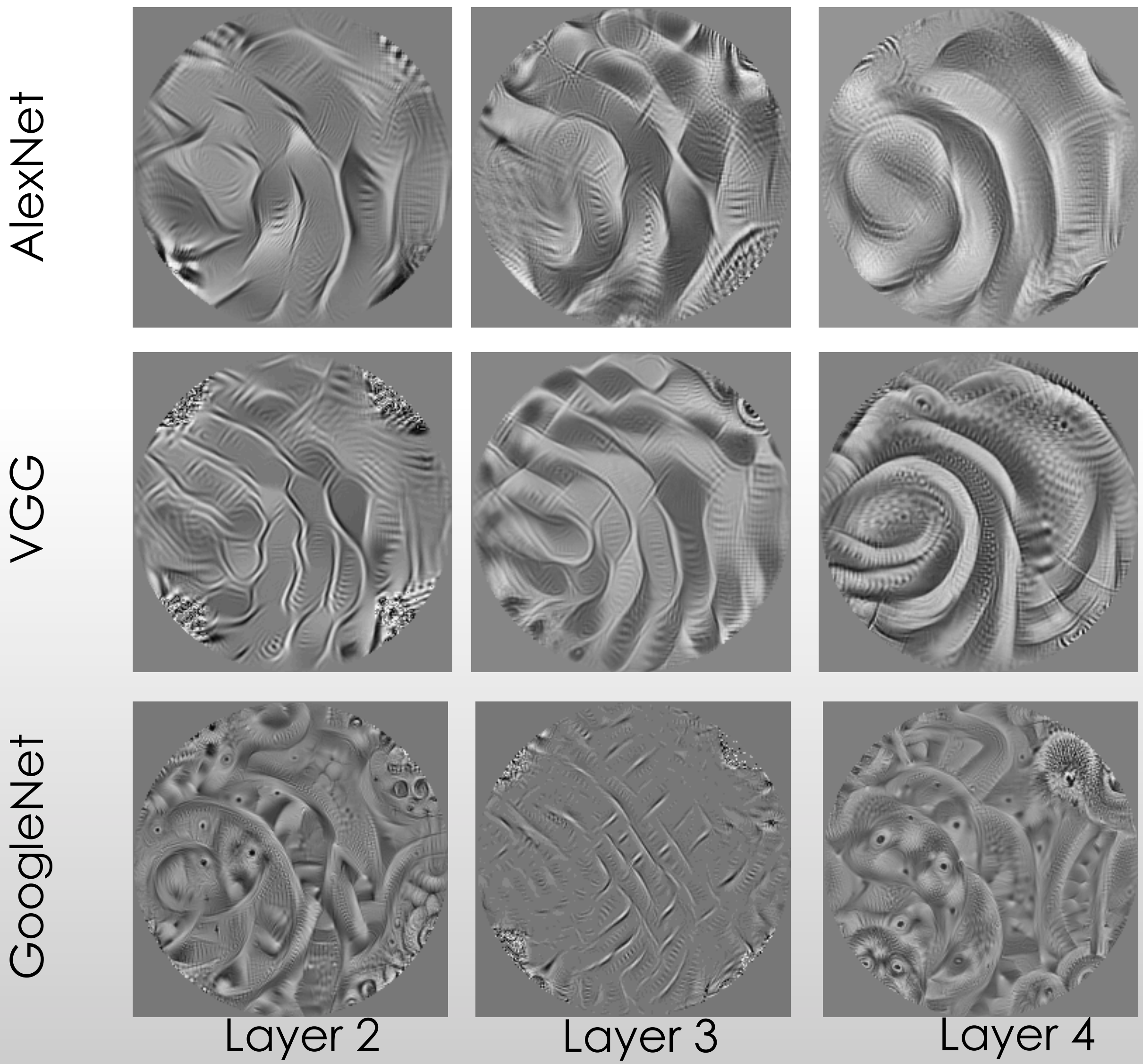
CNN feature index

Neuron response

Barter and Y. (2017)

https://github.com/rlbarter/superheat

Neuron E seems a curve neuron and DeepTune images provide intervention stimuli

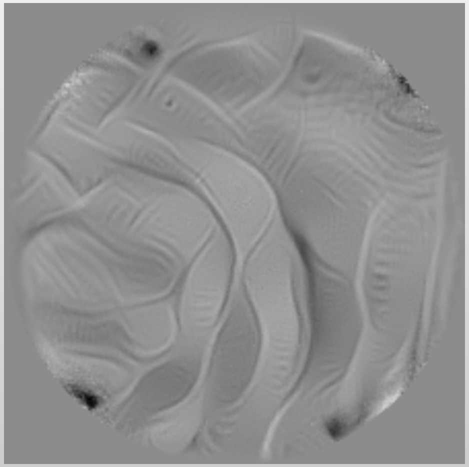


Consensus DeepTune

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

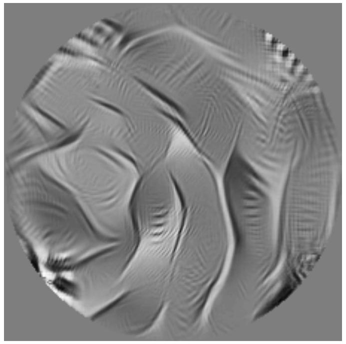
- **Consensus DeepTune:** The models have to agree with each other to create a DeepTune pattern (**Stability**)

$$|\nabla f(x)| = \underset{i=1...\#\text{models}}{\text{element-wise min}} |\nabla f_i(x)|$$

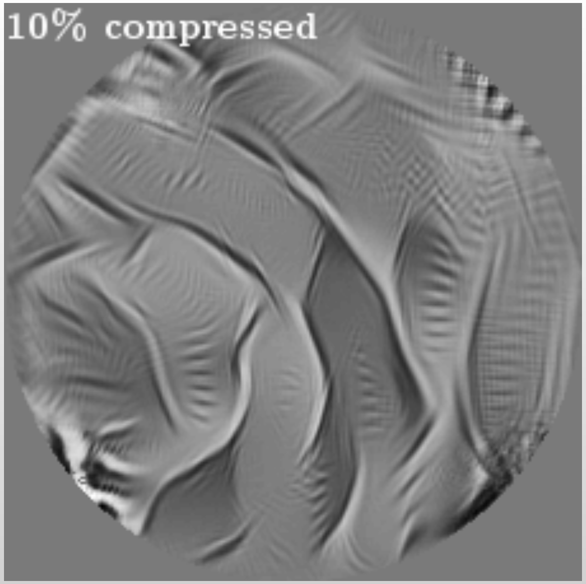
Neuron E 

Stable **curve** patterns across structurally compressed models

DeeTune image from full network

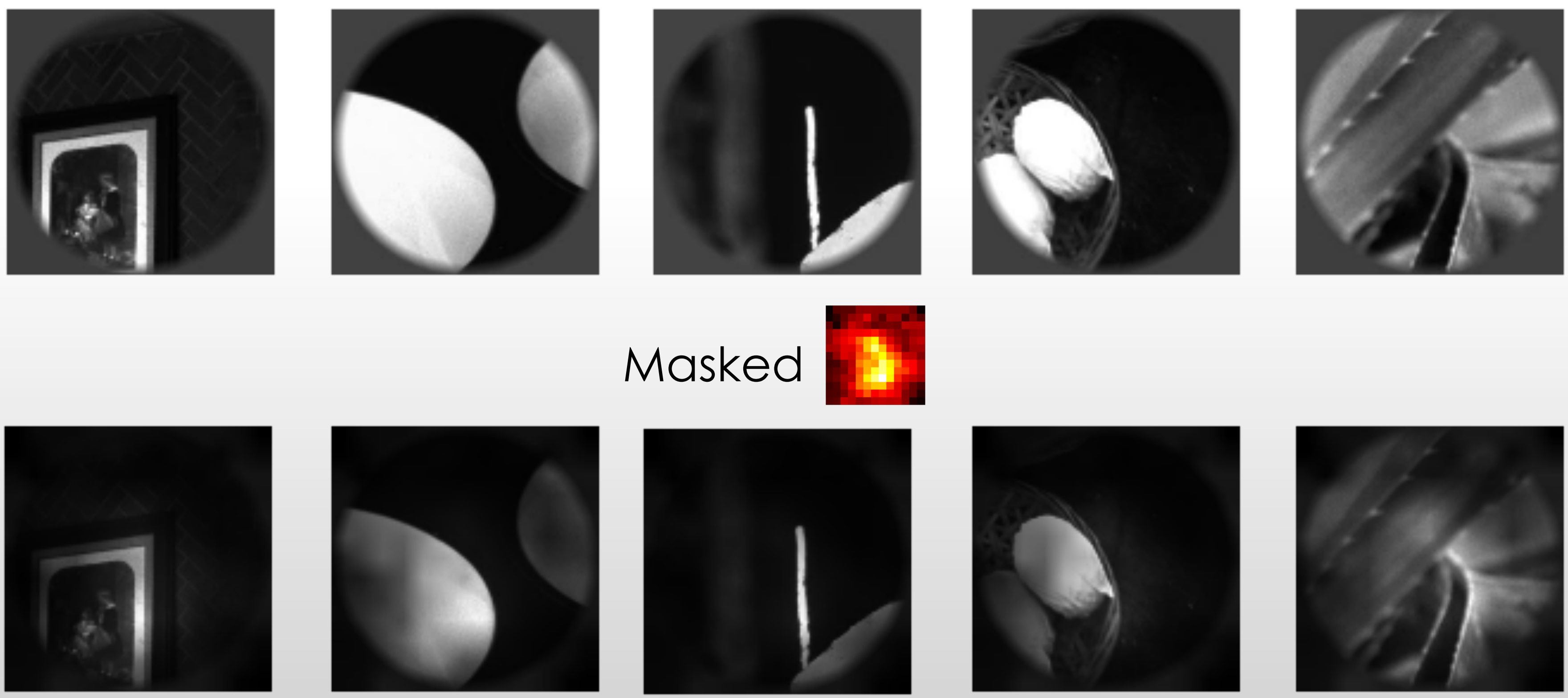


DeepTune images from compressed networks

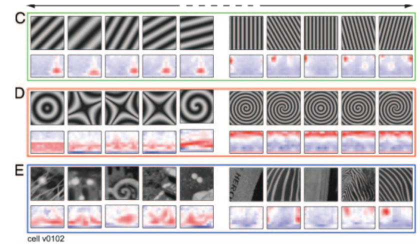


Abbasi-Als and Y. (2017)

Top **curve** images from training set based on a model for neuron E

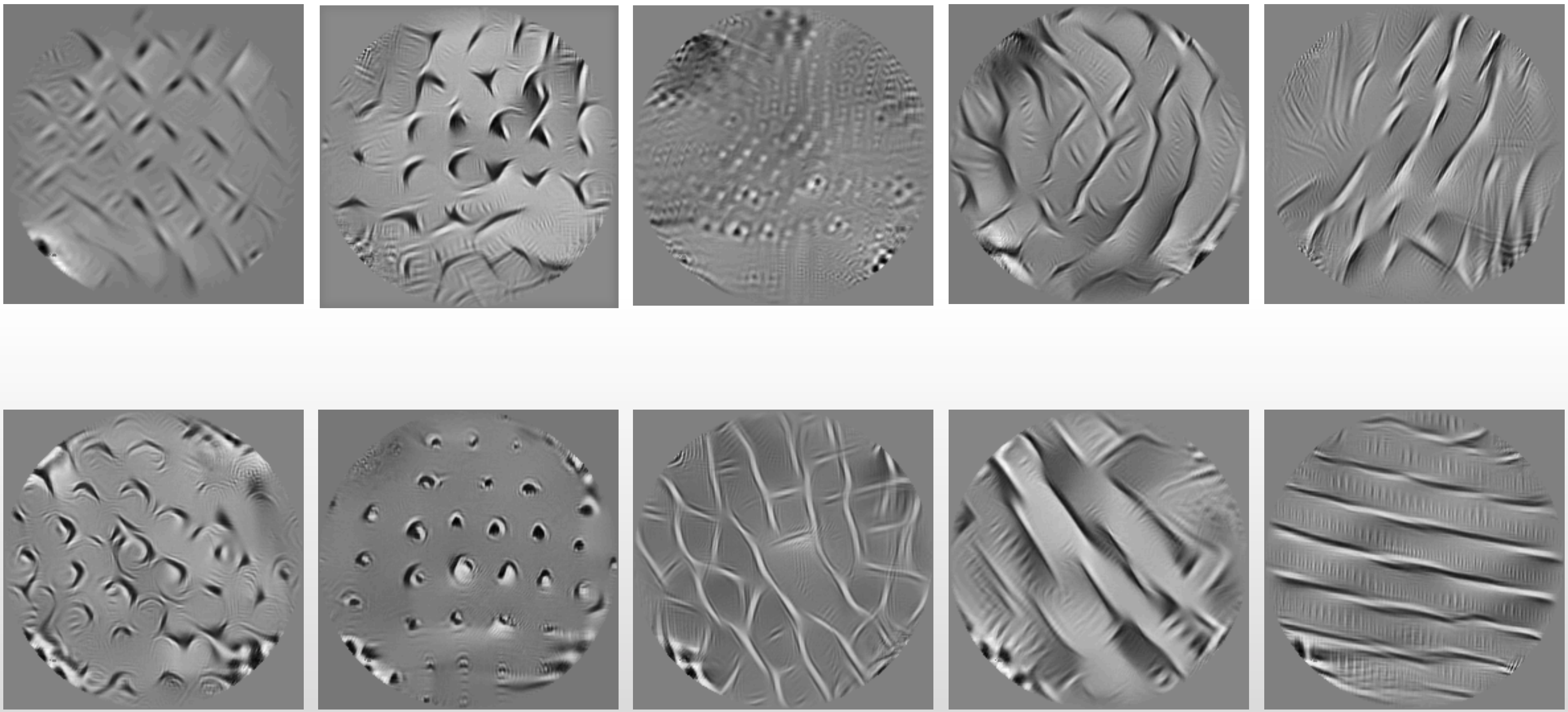


Diversity of V4 neurons via stable DeepTune images



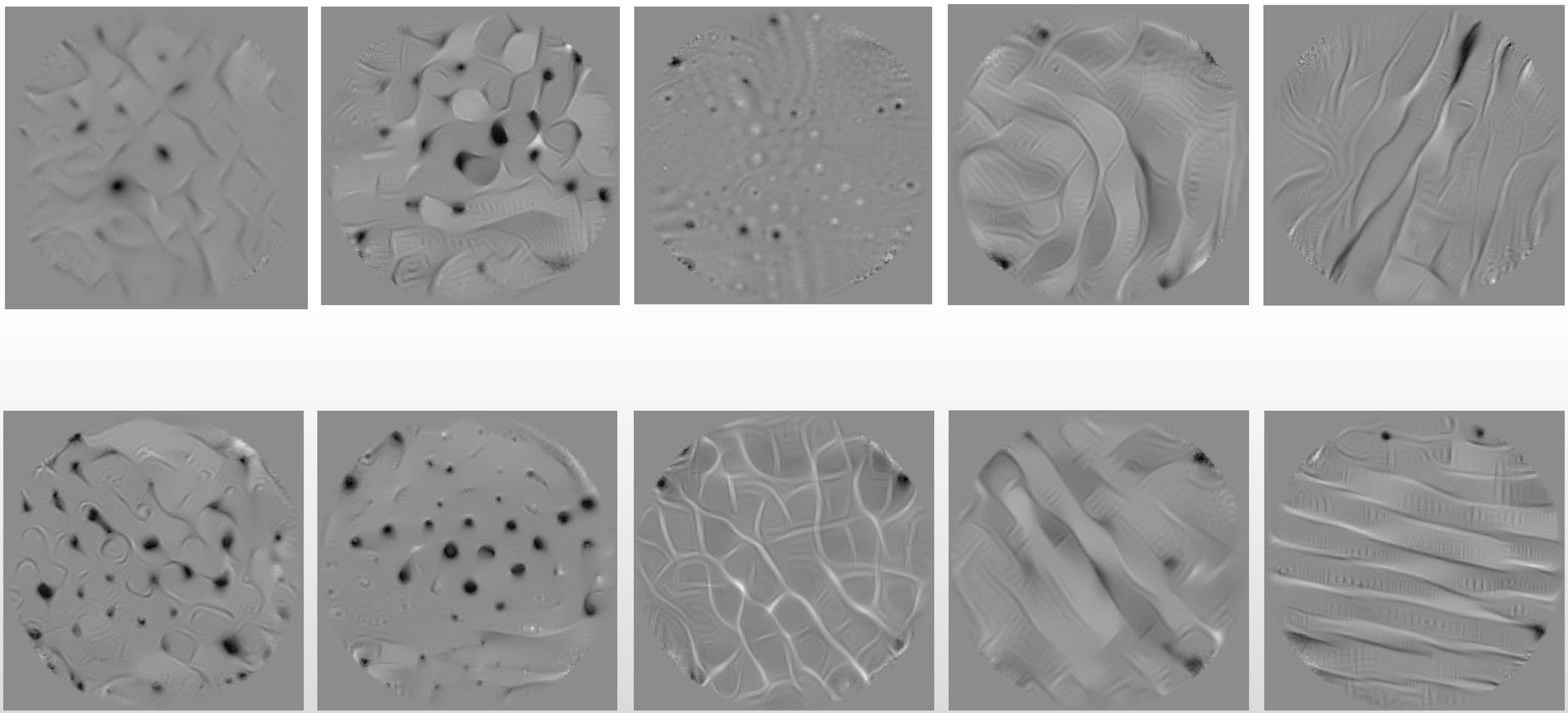
Neuron D

Neuron E



Diversity of V4 pattern selectivity via stable consensus DeepTune images

Neuron D Neuron E



Second example of PCS

iterative Random Forests (iRF) -- integrated PCS

iterative Random Forests to discover predictive and stable high-order interactions

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

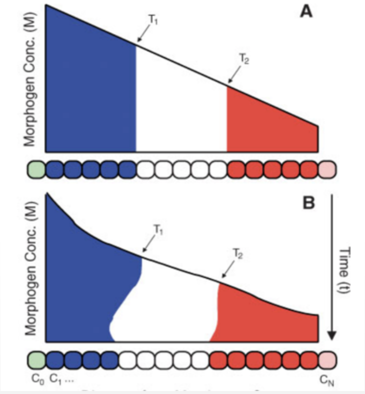
To appear in PNAS (2018)

Culmination of 3+ years of work

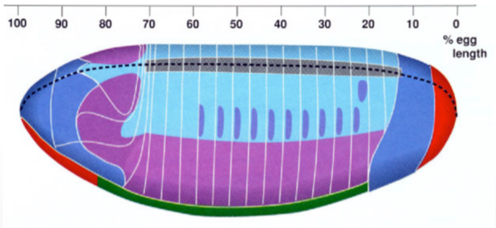
Open source R implementation: https://cran.r-project.org/web/packages/iRF/

Capturing the form of genomic interactions

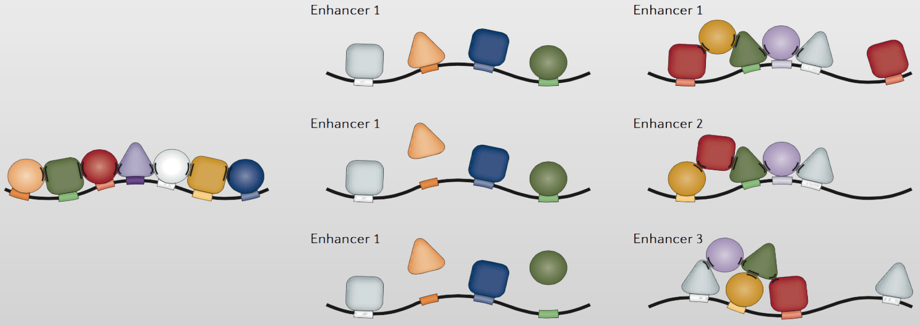
- Interactions are high-order and combinatorial in nature
- Interactions can vary across space and time as biomolecules carry out different roles in varied contexts
- **Interactions** exhibit **thresholding behavior**, requiring sufficient levels of constitutive elements before activating



(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

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

Order-s interaction,

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

iterative Random Forests (iRF)

Basu, Kumbier, Brown and Y. (2018) PNAS.

Project started from Brown's 10+ years of empirical work in genomics using RF and took 3+ years

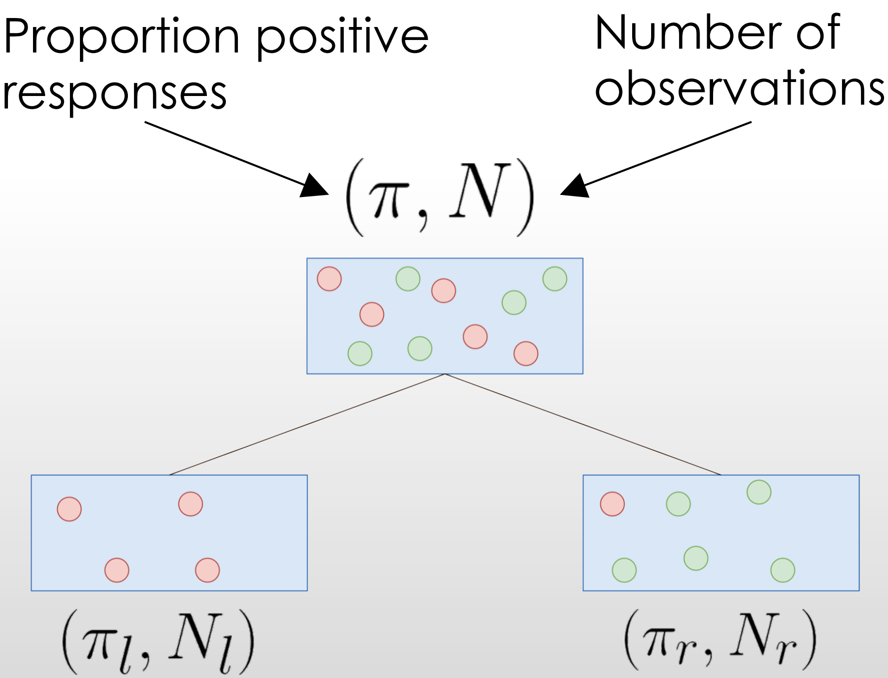
Developed and tested using extensive simulation studies based on synthetic and real data with biologically inspired generative models

iRF output: feature interaction sets with stability scores

iterative Random Forests (iRF) core ideas

1. Interpret RF decision paths
2. Stabilize RF decision paths
3. Assess interaction stability

Interpreting RF: decrease in Gini Impurity as importance measure of a feature

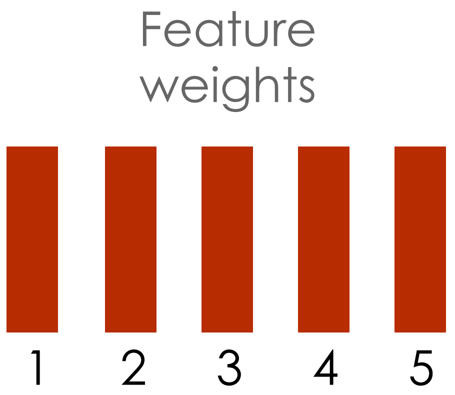


Decrease in Gini Impurity:

$$I_G(\pi) - \frac{N_l}{N} \cdot I_G(\pi_l) - \frac{N_r}{N} \cdot I_G(\pi_r)$$

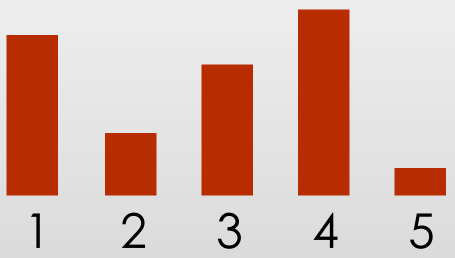
Mean Decrease in Impurity:
On average, how much does splitting on a feature decrease the Gini Impurity?

Feature-weighted RF
Amaratunga et al., 2014



Random Forest:

At each node of the decision tree, uniformly sample `mtry` features to evaluate splitting criteria.



Feature-weighted Random Forest:

At each node of the decision tree, sample `mtry` features with probability proportional to $w \in \mathbb{R}_{+}^{p}$

Generalized RIT: fast computation uses sparsity

Random Intersection Trees (RIT) or 0-1 feature vectors

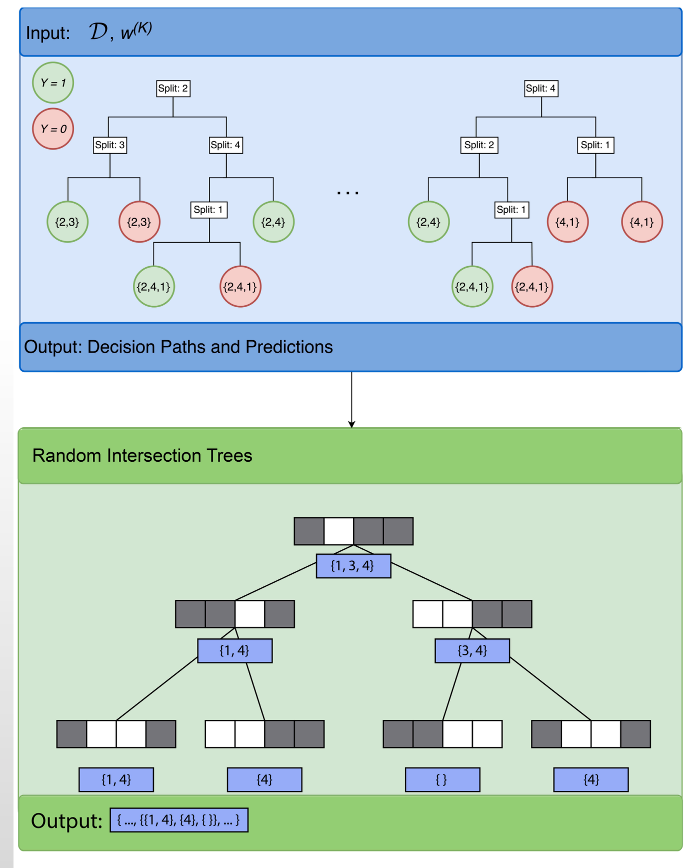
Shah and Meinshausen (2014)

Combining RF and RIT allows us to evaluate prevalent feature combinations on decision paths of RF

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

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

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



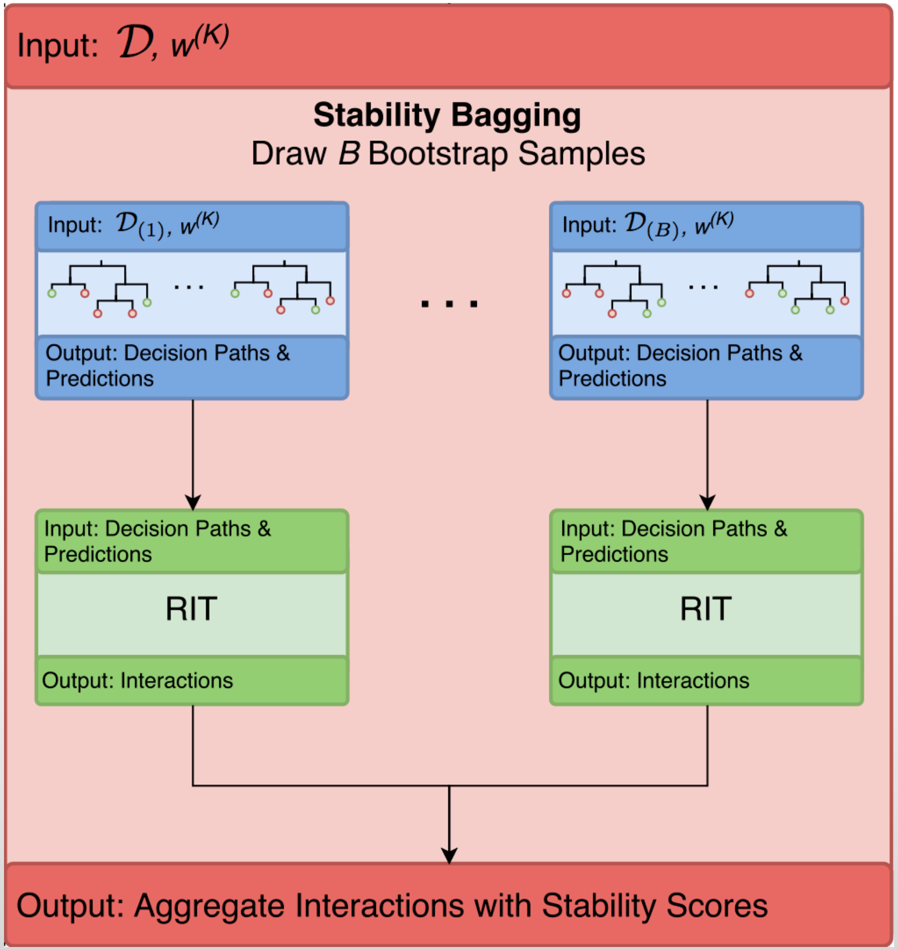
Stability bagging

Output feature interaction sets with stability scores:

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

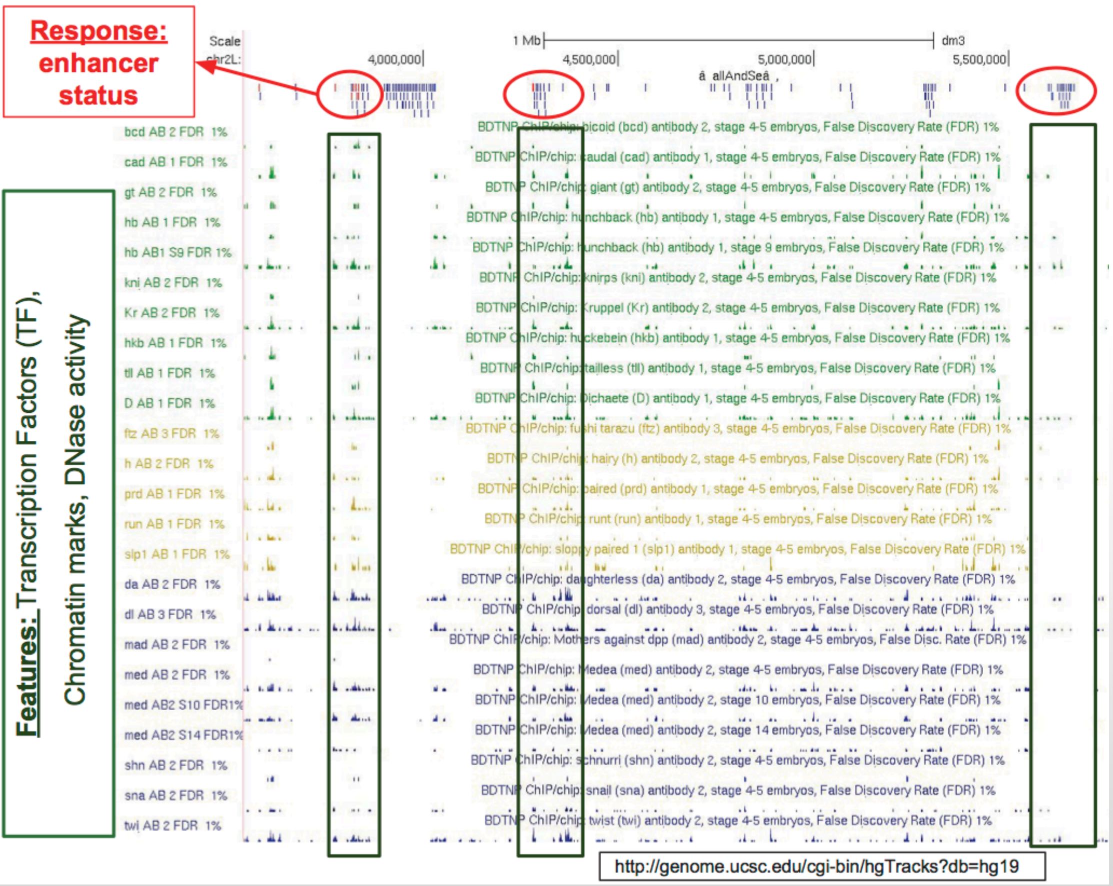
$$S \subseteq \{1, \ldots, p\}$$

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



Reference: (Breiman, 1996)

Case study: Enhancer activity in *Drosophila*

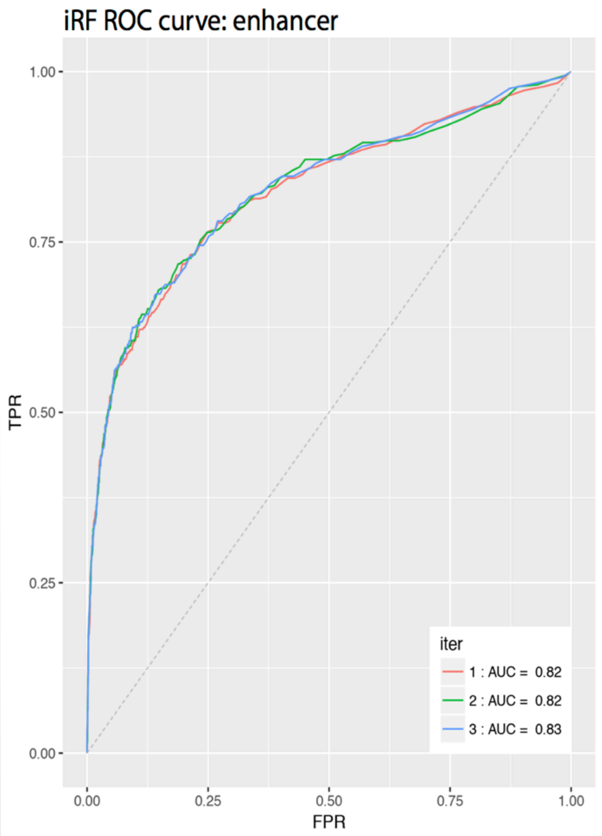


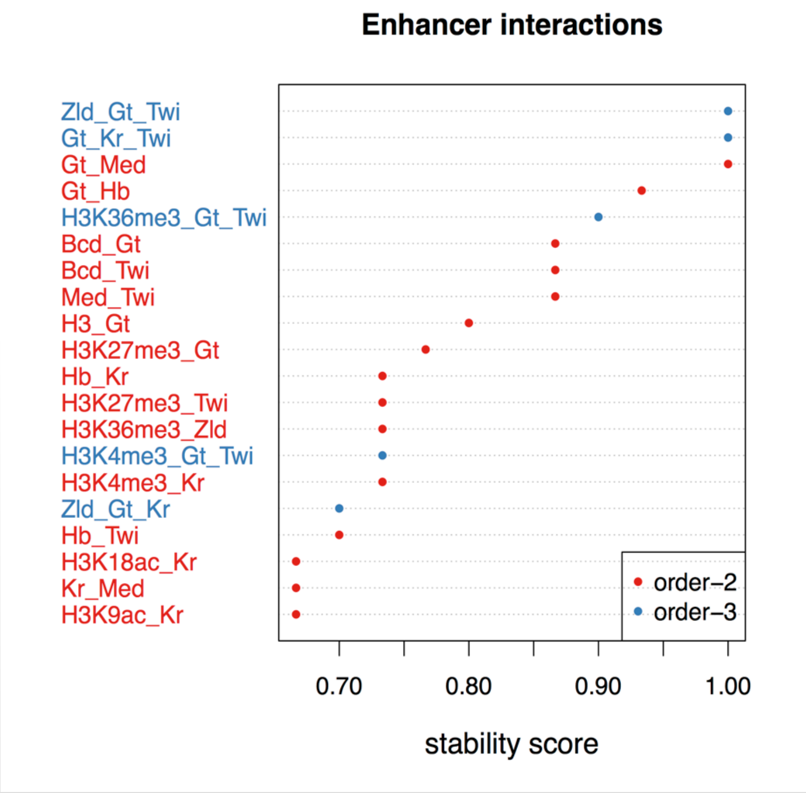
Drosophila blastoderm embryos:

- 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 increases stability hence interpretability while maintaining predictive accuracy

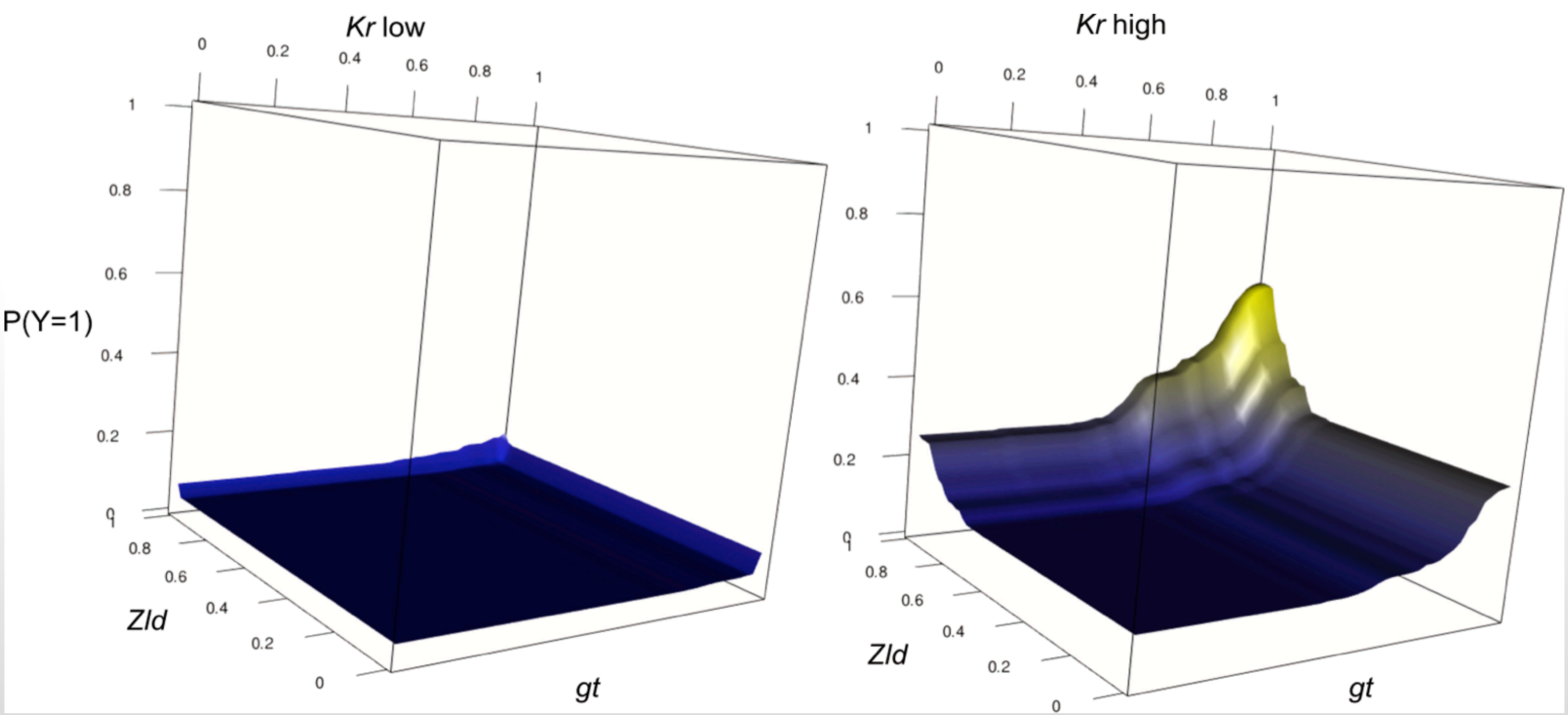




iRF identifies 20 stable pairwise interactions in *Drosophila* – ***80%*** *are proven physical interactions in the literature*

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



PCS-related theory: iterative learning algorithms

Algorithmic stability

Equivalent

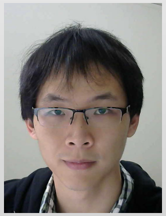
Generalization

*

Trade-off

"Stable algorithms can not converge too fast…"

Practical computability: convergence rate



* Chen, Jin and Y. (2018) https://arxiv.org/abs/1804.01619
"Stability and convergence trade-off of iterative optimization algorithms": optimization error is like computational bias in large scale problems

PCS workflow

- Stability is as fundamental as predictability (reality check) in data science life cycle
- **PCS workflow** documentation: **transparent written arguments** for prediction set-up, model/algorithm choices, “appropriate” perturbations, target, and metric
- PCS workflow leads to predictive and stable models for interpretation and scientific recommendations for intervention experiments
- Structural match of model and domain knowledge is essential since we are still in **data poor** situations relative to biological complexity, even with big data

Berkeley DS Intellectual and Organizational Vision

Summary of the 2016 Report by the Faculty Advisory Board of the Data Science Planning Initiative

Prepared: 19 August 2016
Cathryn Carson, FAB Chair

Contents

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

Interim Dean of a new div:
David Culler

New DS Major coming…

Data8 Spring18 – 1000 students

Home » Education Program
Data Science Education Program

DS100 Spring18: 600+ students

Thanks to my group members and grants

National Science Foundation
WHERE DISCOVERIES BEGIN

National Institutes of Health
Turning Discovery Into Health

Center for
Science of Information
NSF Science and Technology Center

ARO, ONR

Links and thanks

Berkeley Data Science FAB report summary
https://drive.google.com/open?id=0B8gpOw0SuKG4NTR5MVJWQjhoc2s
https://www.stat.berkeley.edu/~binyu/ps/FAB-ExecutiveSummary2016.pdf

Berkeley Data Science FAB report
https://drive.google.com/open?id=0B8gpOw0SuKG4cGR1NTZpTzBQRGM
https://www.stat.berkeley.edu/~binyu/ps/FAB2016.pdf

Center for Science of Information
NSF Science and Technology Center

NIH National Institutes of Health
Turning Discovery Into Health

ARO, ONR