Interpreting Deep Neural Networks towards Trustworthiness

Bin Yu
Statistics, EECS, CCB at UC Berkeley

International Congress of Mathematicians (ICM)
July 8, 2022

AI is part of modern life

Virtual assistants
(Siri, Alexa, Cortana)

Wearable health devices
(FitBit, Apple watch)

Recommendation systems
(YouTube, Facebook)

Online news

Election campaigns

Self-driving cars

Online gaming

Precision medicine

Biology

Chemistry

Neuroscience

Materials Science

Law

Sociology

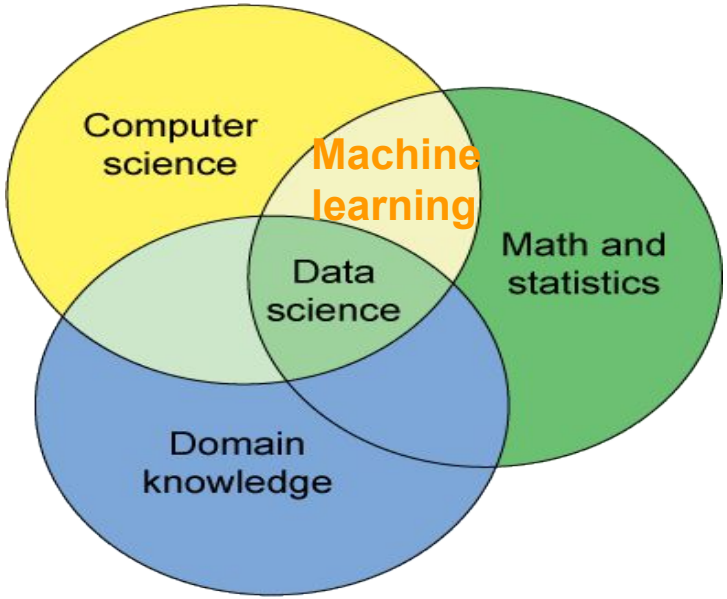
Cosmology

Economics

Political Science

… and beyond

Data science (DS) is a key element of AI

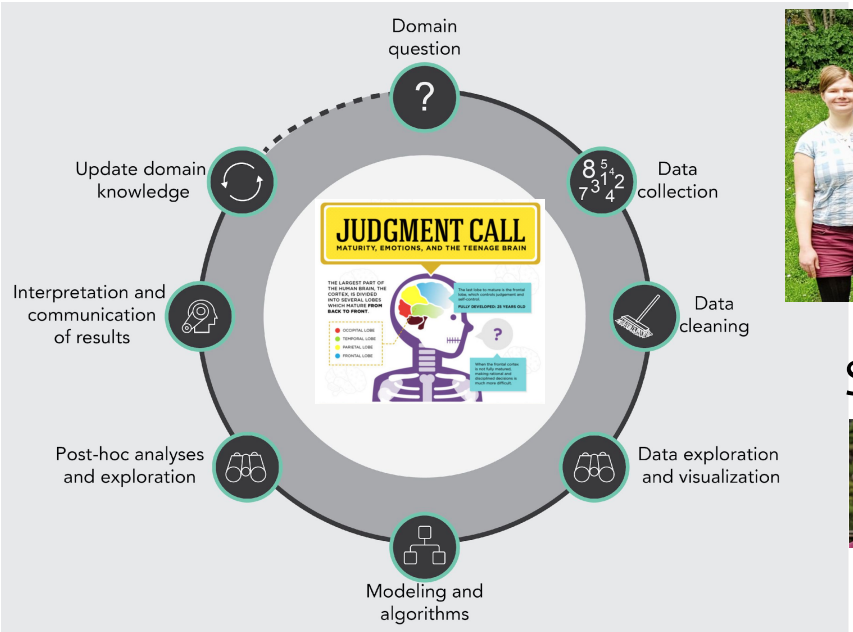


Conway's Venn Diagram

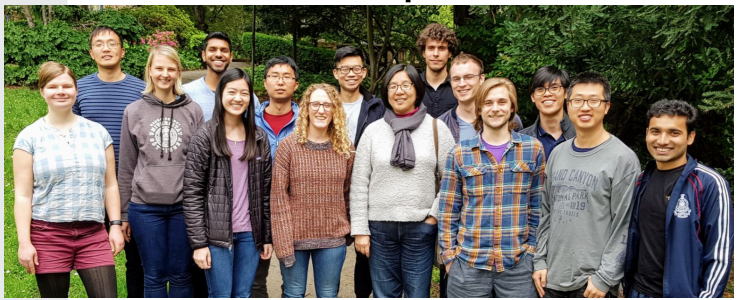
Goal:

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

Data Science Life Cycle (DSLC): A holistic view



Yu Group



Scientist collaborators

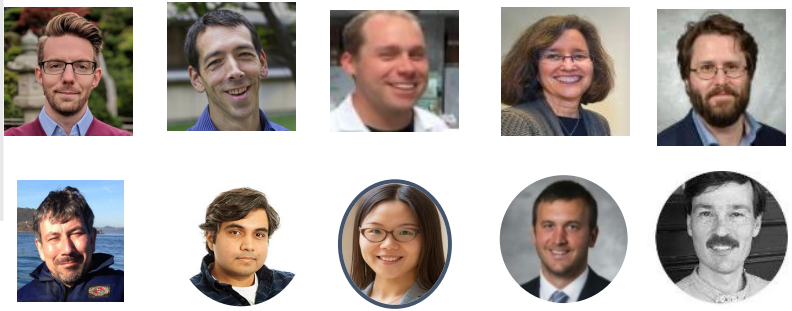


Image credits: R. Barter and toronto4kids.com

ML/Stats Frontier: interpretation

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

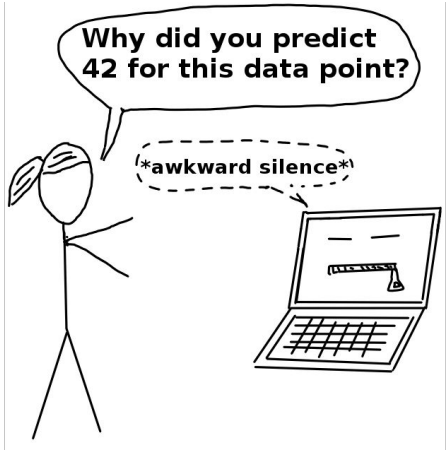
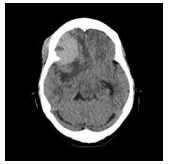


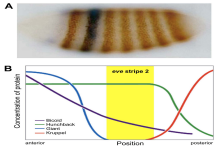
Image credit: https://christophm.github.io/interpretable-ml-book/

Where interpretation is needed in reality

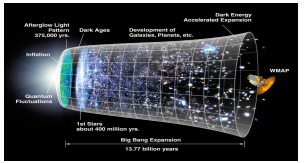
- FDA wants interpretation of DL algorithms for radiology



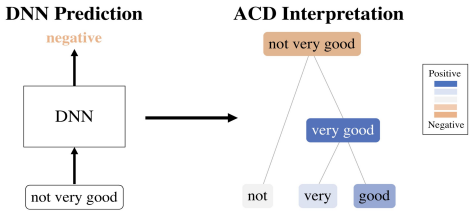
- **Biologists want to know which genes drive a phenotype by fitting iterative random forests (iRFs)**



- **Cosmologists want to know what info is most useful to figure out origin of the university using DNNs**



- Machine learners want to know how a prediction is made by a DNN



Interpretation is necessary in scientific ML

What is scientific ML?

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

What is interpretable ML (iML)?

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



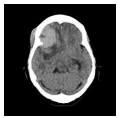

Definitions, methods, and applications in interpretable machine learning

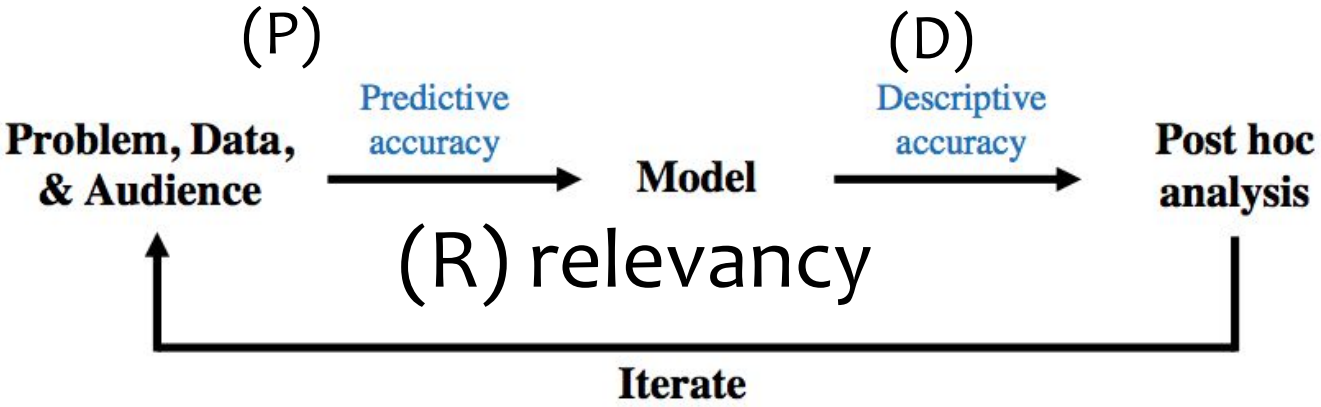
W. James Murdoch[a,1], Chandan Singh[b,1], Karl Kumbier[a,2], Reza Abbasi-Asl[b,c,d,2], and Bin Yu[a,b,3]

[a]Statistics Department, University of California, Berkeley, CA 94720; [b]Electrical Engineering and Computer Science Department, University of California, Berkeley, CA 94720; [c]Department of Neurology, University of California, San Francisco, CA 94158; and [d]Allen Institute for Brain Science, Seattle, WA 98109

Contributed by Bin Yu, July 1, 2019 (sent for review January 16, 2019; reviewed by Rich Caruana and Giles Hooker)

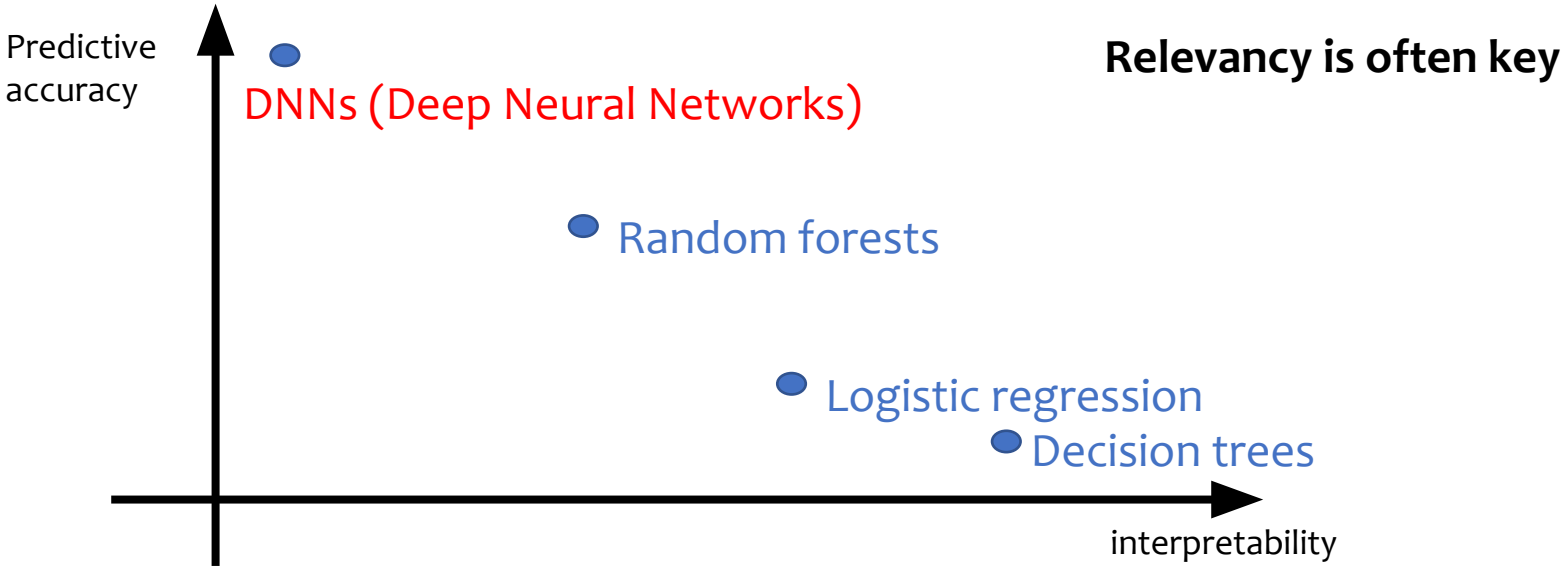
iML-PDR in one figure





R is key in the trade-off of P and D

D vs P for model-based interpretability



Relevancy is often key

Watershed Moment for computer vision: AlexNet

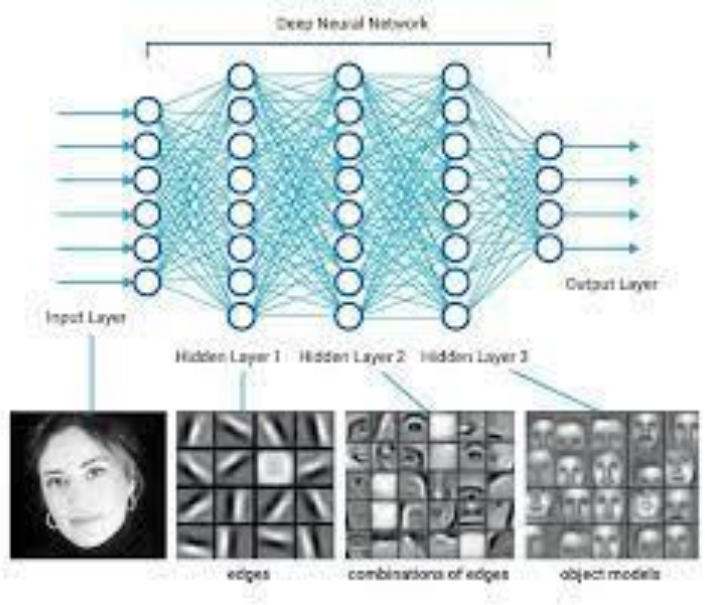
AlexNet (Krizhevsky, Sutskever, and Hinton, ‘12) sparked resurgence in neural networks research

ImageNet

Now: 90.9%

PapersWithCode benchmark for ImageNet

Dense DNN



AlexNet: CNN or Convolutional NN

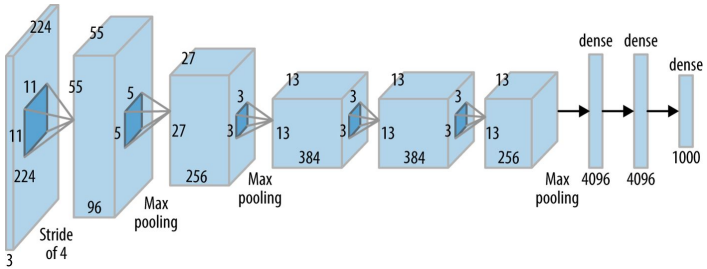


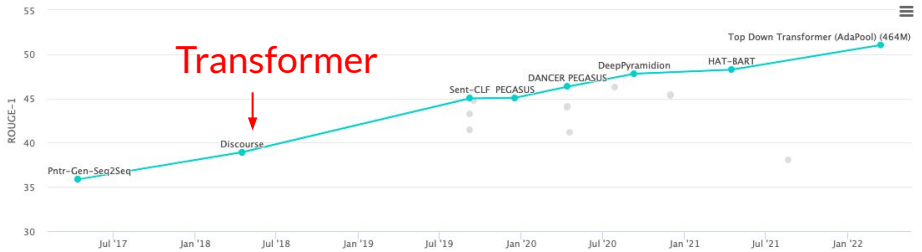
Image credits: medium.com
https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96

Watershed Moment for NLP: Transformers

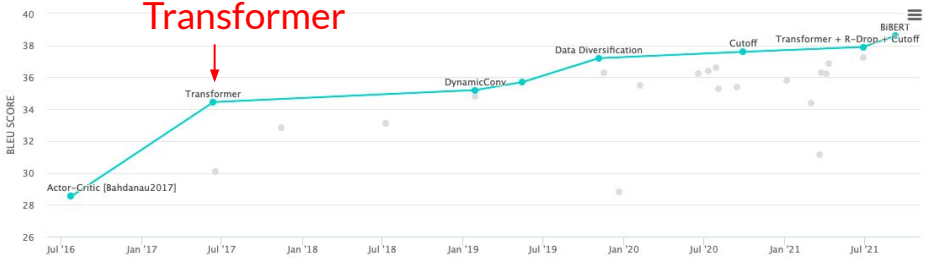
Transformers (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin, 2017)
Improved:

- Summarization
- QA
- Info Retrieval
- Translation
- and many more…
-

Now ubiquitous



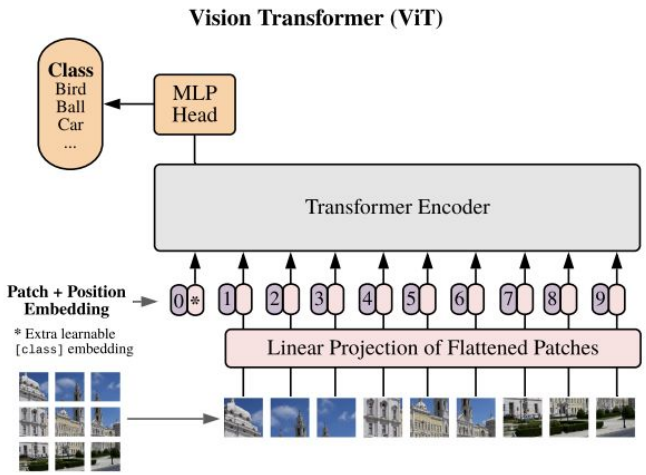
Summarizing Medical Documents



Translating German to English

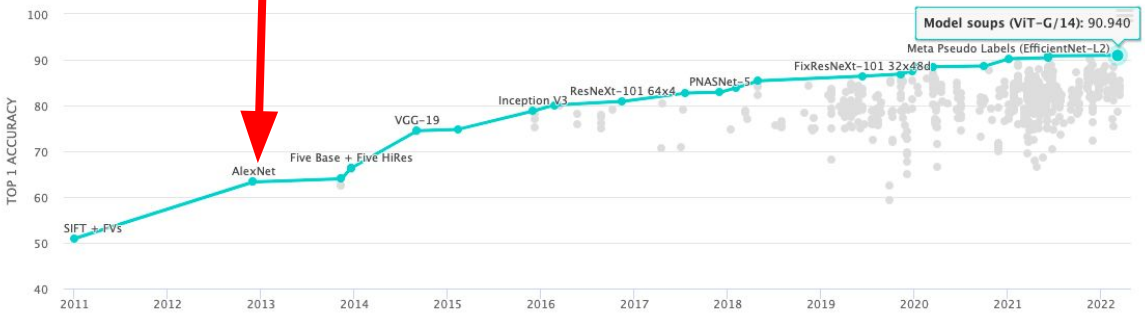
PapersWithCode benchmarks for German-English Translation, PubMed Text Summarization

Not Just NLP…



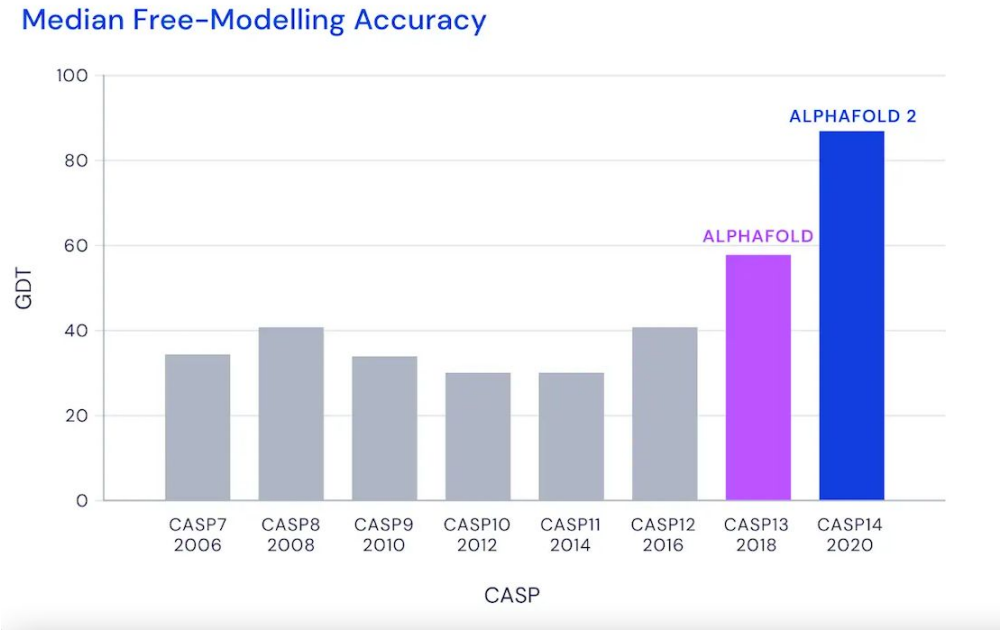
AlexNet: 63% accuracy

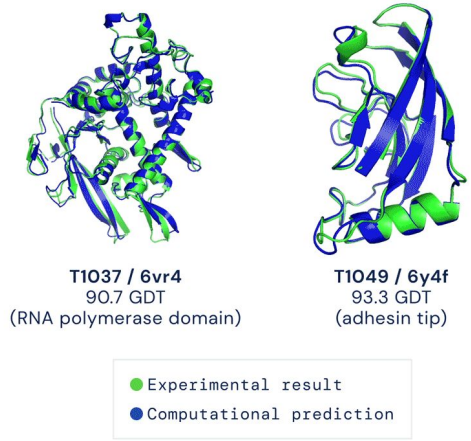
ViT: 90.9% accuracy on ImageNet



Dosovitskiy et al. An Image is Worth 16x16 Words. 2020

Not Just NLP…





AlphaFold 2 cracks Protein Folding

https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Interpretable DL in scientific ML: R is the key

Expanding R in scientific ML:

interpretation method is “relevant” if it provides insight for a particular scientific audience into a chosen scientific problem

Scientific insights have to correspond to facts about the real world. Hence interpretations of ML models need to be vetted by and calibrated against trustworthy or established factual information.

Where does trust in interpretable DL come from?

- **Relevant scientists** in established scientific fields to provide confirming scientific knowledge to the interpretation that is an accumulation of factual or reliable empirical evidence in the field

- Veridical (truthful) data science process (e.g. our Predictability, computability and stability (PCS) framework) to vet **every step of the data science life cycle**

Rest of the talk

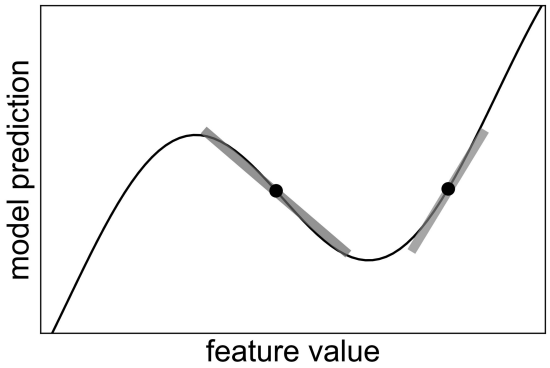
- ACD (Agglomerative Contextual Decomposition) for interpreting DNNs in ML problems and cosmology
- AWD (Adaptive Wavelet Distillation) for DNNs (cosmology and biology)
- PCS framework for veridical data science or building trustworthy DNNs and models in general

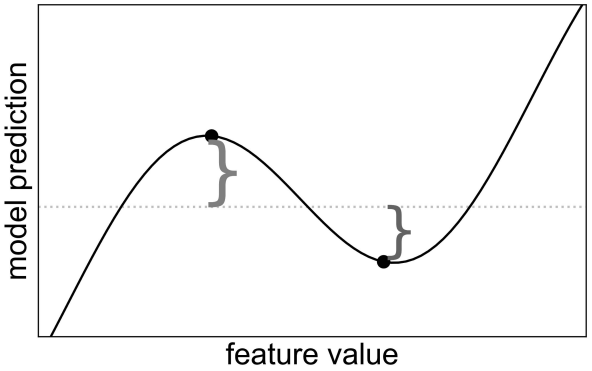
Part I:
Agglomerative Contextual Decomposition (ACD)

(1) How can we get feature-interaction importance for a DNN model prediction in general?

(2) How can we visualize these feature-interactions in an understandable way?

Previous work (post-hoc interpretation)





gradient-based

- LIME (Ribeiro et al 2016)
- Integrated Gradients (IG) (Sundarajan et al 2017)

contribution-based

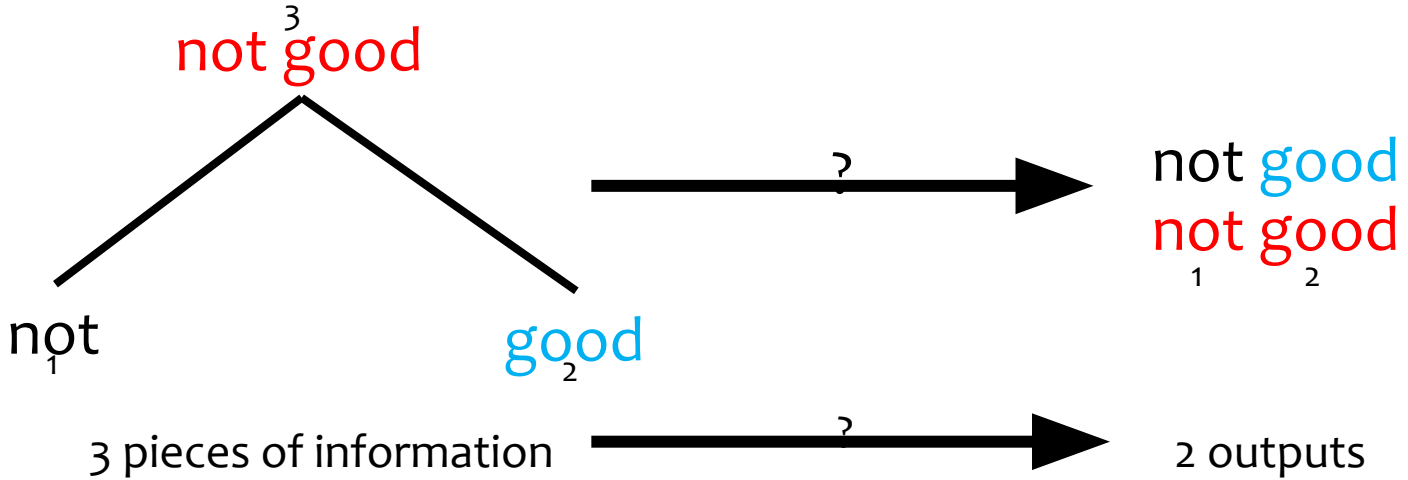
- Occlusion / saliency maps (Dabkowi & Gal 2017)
- SHAP (Lundberg & Lee 2017)

An example from sentiment analysis

- Binary sentiment analysis with standard LSTM

The movie was good

Word importance scores can't capture compositionality



CD: Contextual Decomposition

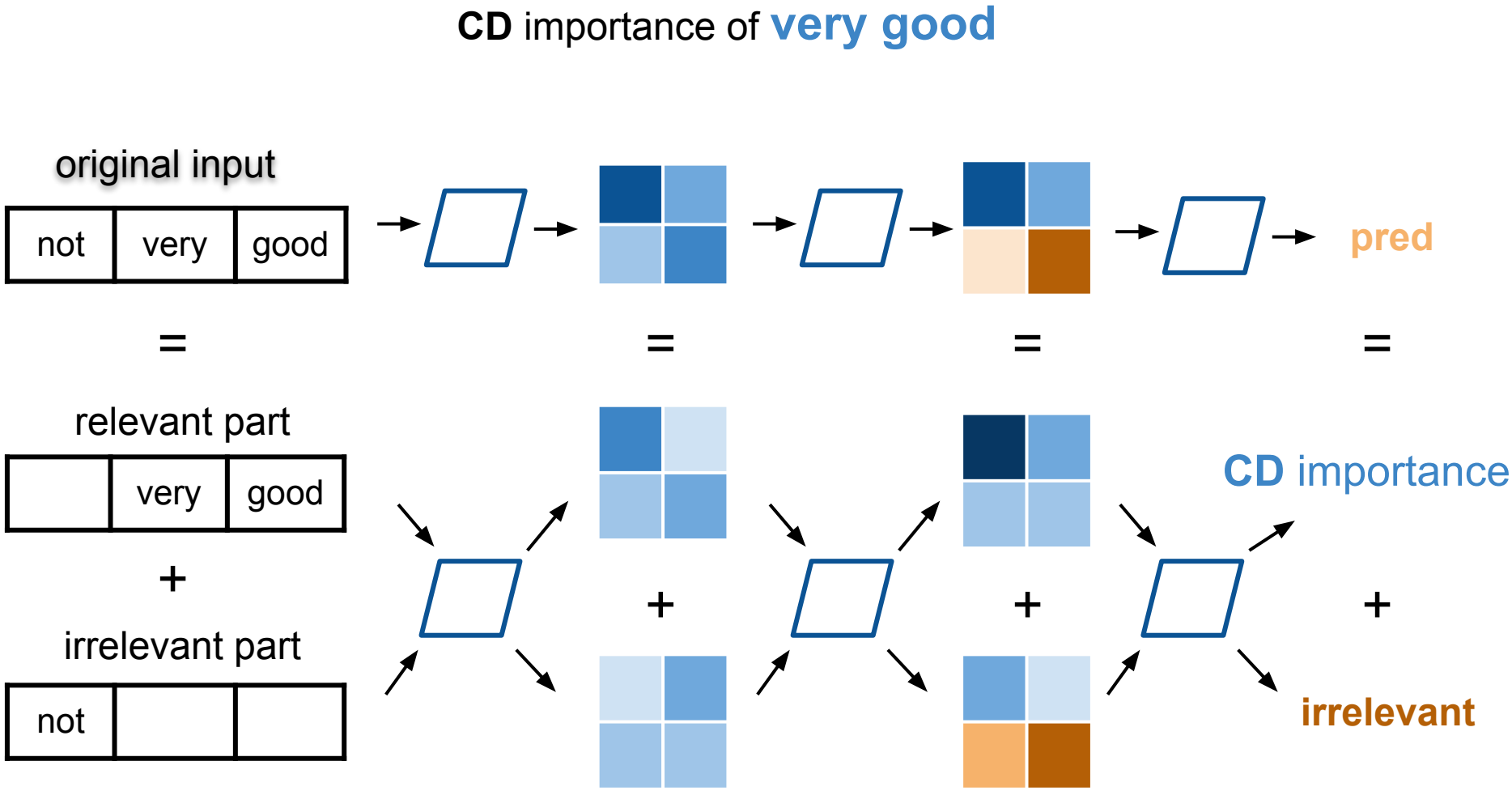
(Murdoch, Liu, Y. , 2018, ICLR)

- Given a LSTM with weights, CD gives a prediction-level score for each phrase to ``explain'' the prediction

The movie was **not good**

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

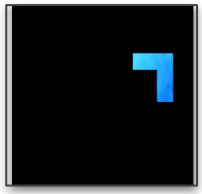
- γ_T corresponds to contributions solely from the phrase, α_T other factors

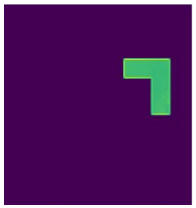

CD importance of very good
original input
not
very
good
pred
=
=
=
=
relevant part
very
good
+
irrelevant part
not
+
+
+
CD importance
irrelevant

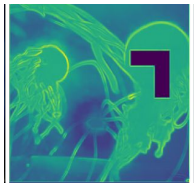
importance of
this region?

relevant
part

irrelevant
part







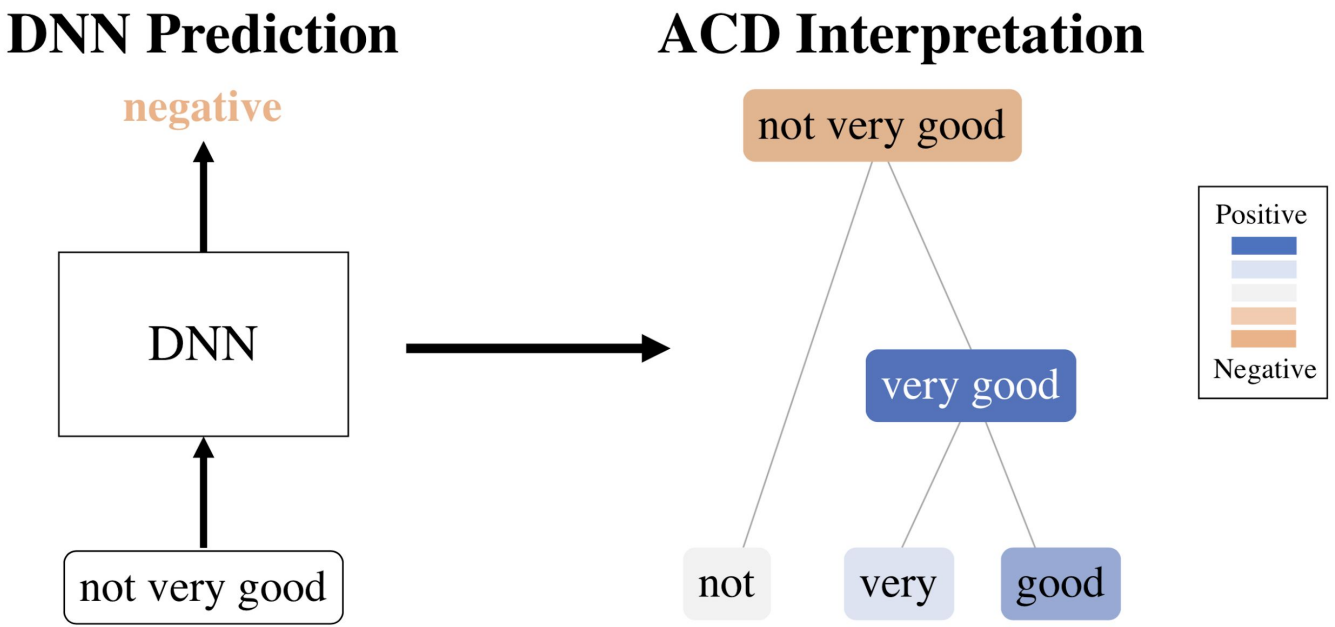
CD importance

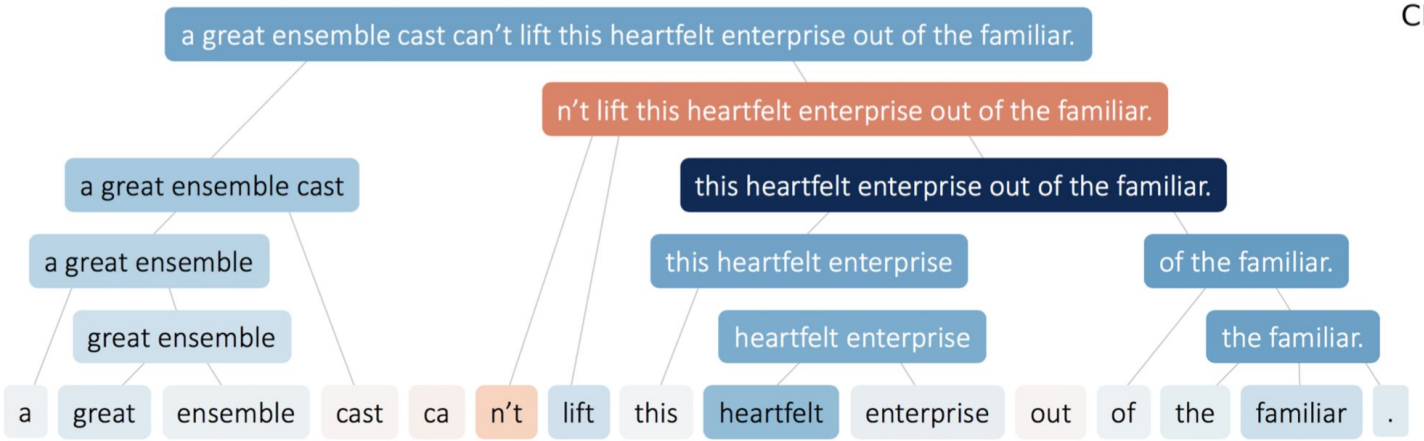
irrelevant

Agglomerative Contextual Decomposition (ACD)


*Singh, *Murdoch, Y. (2019).
Hierarchical interpretations for neural network predictions
Proc. ICLR

ACD is a hierarchical clustering algorithm with visualization, where the joining metric is CD scores

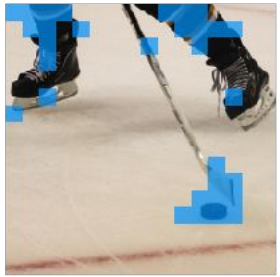

DNN Prediction
negative
DNN
not very good
ACD Interpretation
not very good
very good
not
very
good
Positive
Negative

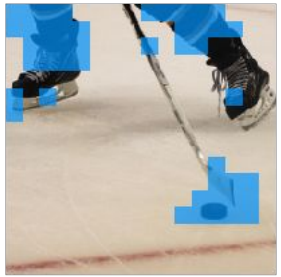


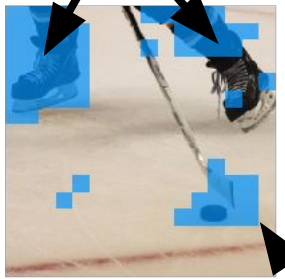
Positive

Negative

prediction: puck



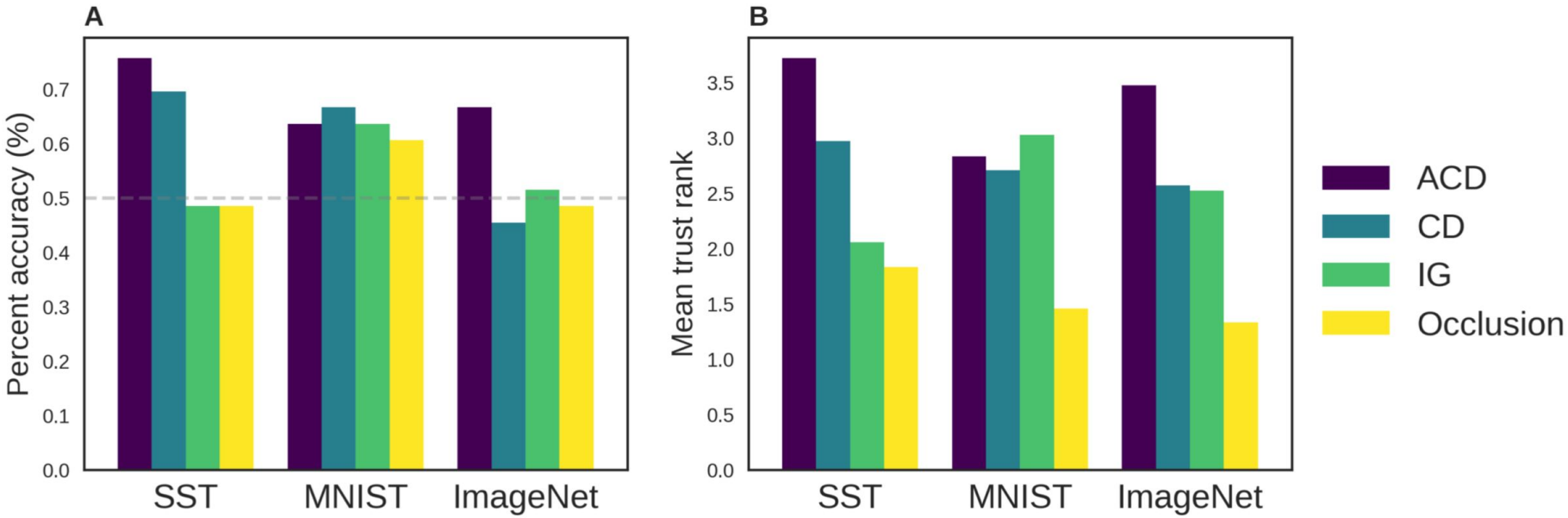




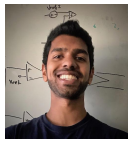
skates are important

puck is important

Human experiments to compare iML methods



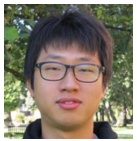
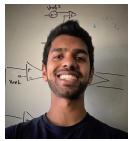
Improving models by regularizing ACD explanations

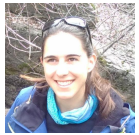

Rieger, Singh, Murdoch, Y. (2020).
ICML

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

Using CD to identify fundamental cosmological parameters of the universe

Yu group

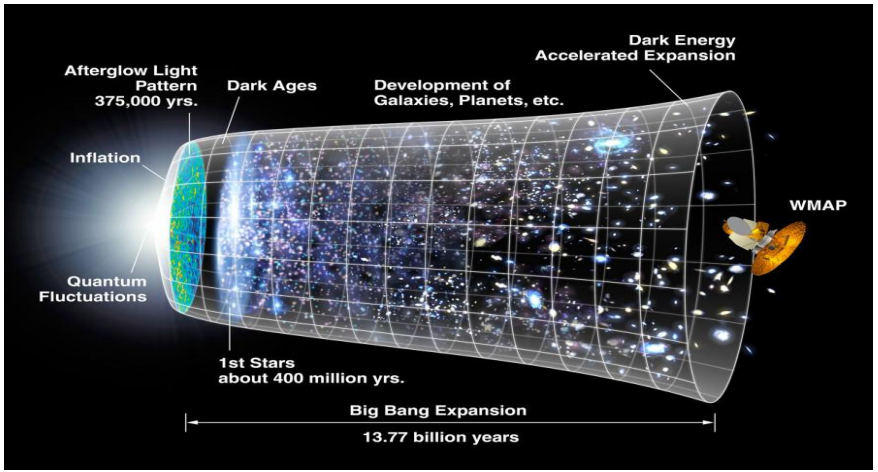




@ Berkeley Center for Cosmological Physics

W. Ha, C. Singh,
F. Lanusse, V. Boehm, J. Liu, Y. (2020), ICLR Workshop Paper

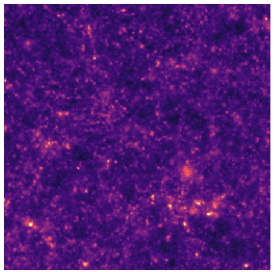
Cosmological parameters such as Ω_m, determine evolution of universe

$\Omega_m \rightarrow$



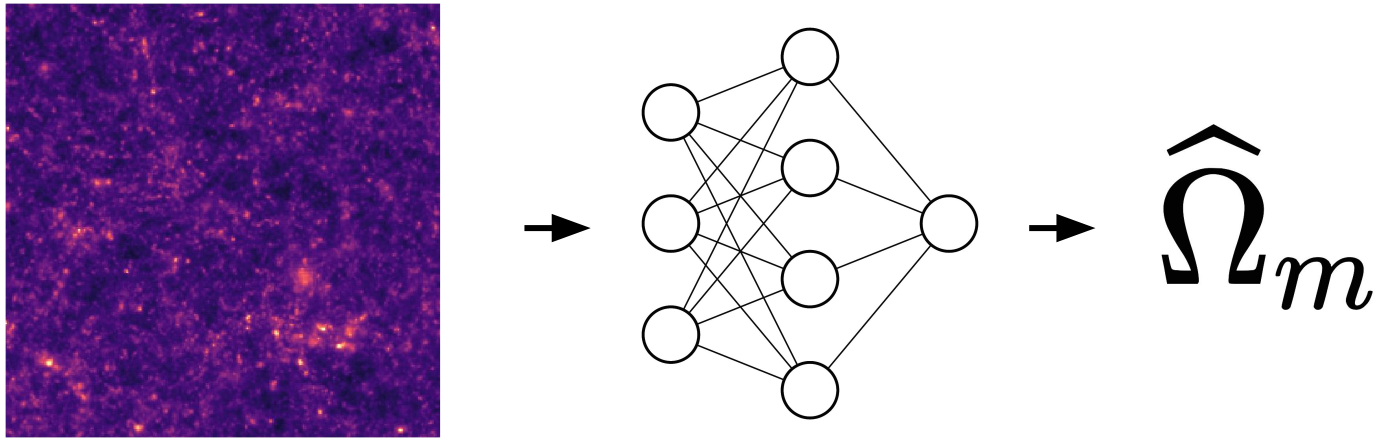
$\rightarrow$

Map of mass in the universe



Adaptation of NASA WMAP Science Team Image

CNN predicts well, but what does it learn?



Need to go beyond just identifying important pixels to frequency domain

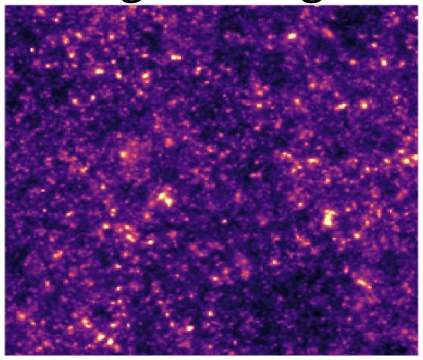
iML in scientific ML: R is the key

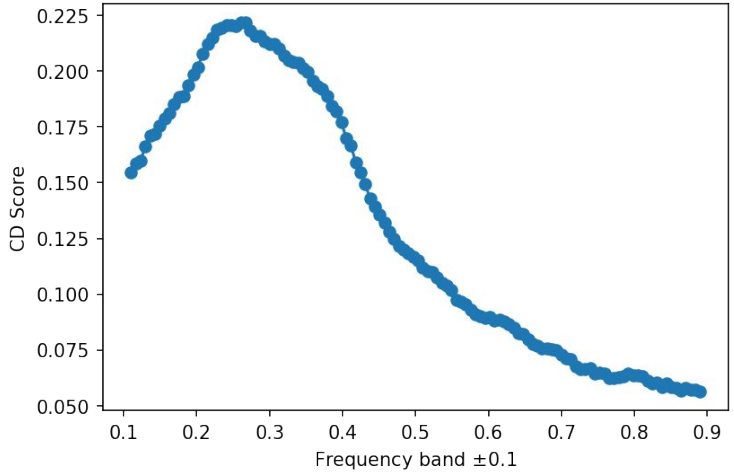
An interpretation method is “relevant” if it provides insight for a particular scientific audience into a chosen scientific problem

Scientific insights have to correspond to facts about the real world. Hence interpretations of ML models need to be vetted by and calibrated against trustworthy or established factual information.

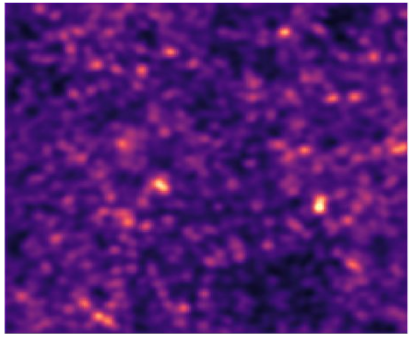
CD can measure the importance of different **frequencies** in the image to the model's prediction

Original image

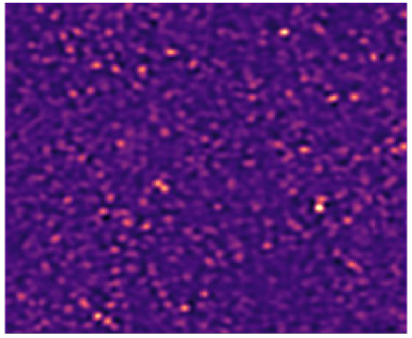




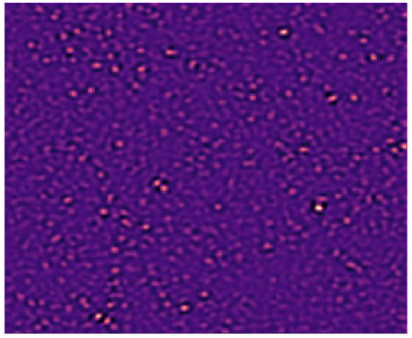
0.1



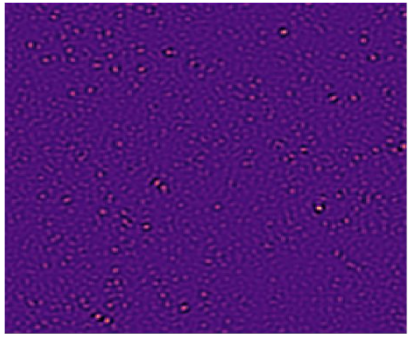
0.2



0.3



0.4



Part II: Adaptive wavelet distillation (AWD) from neural networks through interpretations




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

(NeurIPS, 2021)

Distillation

Model distillation transfers knowledge in a complex model into a simpler model (Hinton et al 2015, Bucila et al 2006)

Recent works distill DNNs into “interpretable” models, such as decision tree (Frosst and Hinton 2017) or additive models (Tan et al 2018)

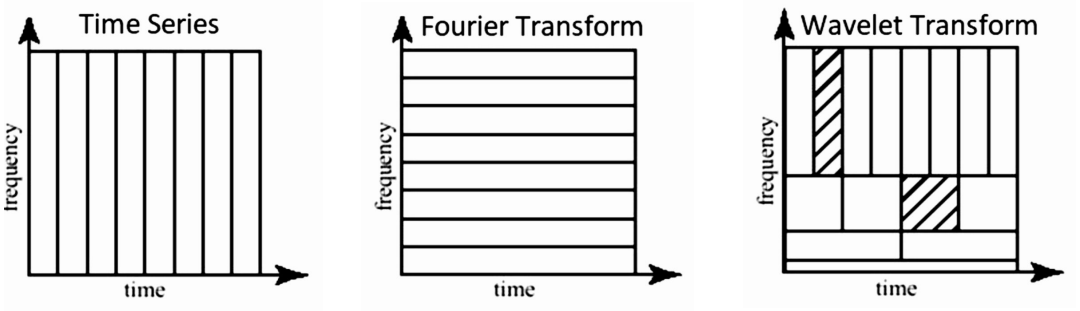
Goals

Given a domain where a DNN predicts well, distill it into a simple **learned** wavelet transform

Improves interpretability, compression, and efficiency

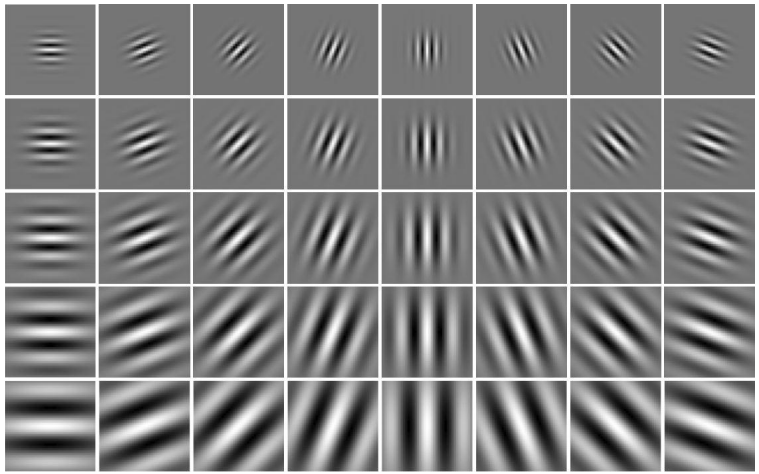
Discrete wavelet transform

The analysis of signals through wavelet coefficients allows for identifying features of different scales at different locations



Multiresolution Analysis (Mallat 1989, Meyer 1992) provides a way to build orthogonal wavelet basis such that the wavelet transform can be done fast

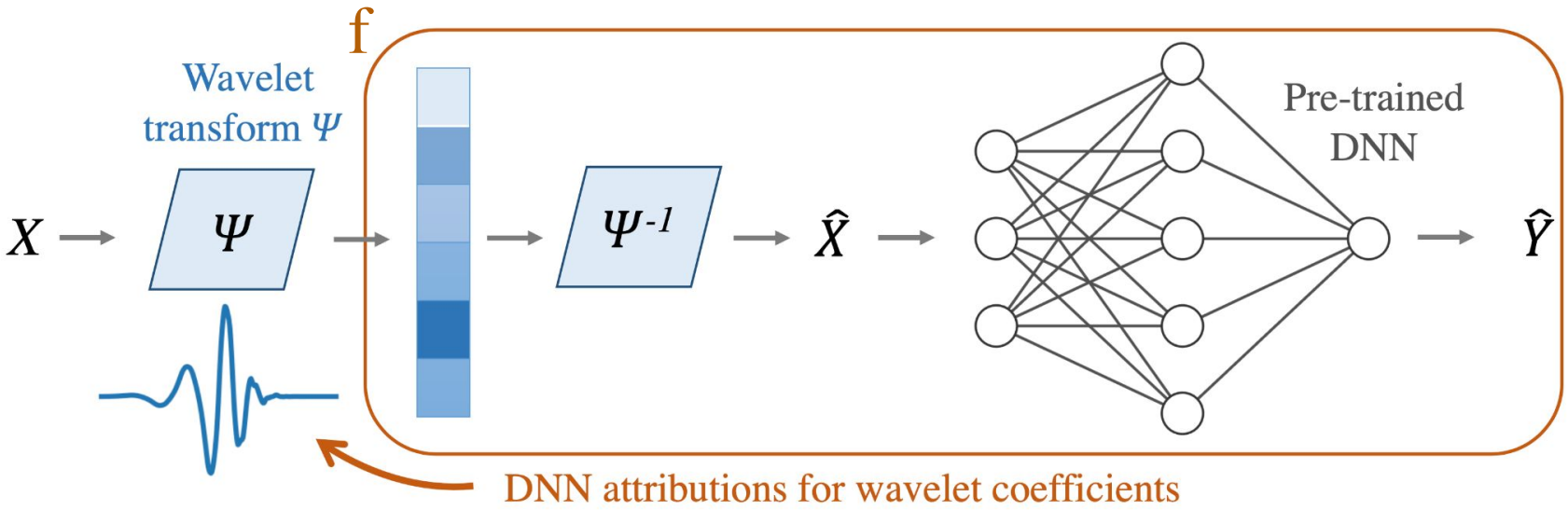
2-dim “Wavelet” Transforms



Gabor wavelet filters

first layer of AlexNet

Adaptive Wavelet Distillation (AWD)



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

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

Sufficient conditions on the low-pass filter

Theorem 7.2: *Mallat, Meyer.* Let $\phi \in \mathbf{L}^2(\mathbb{R})$ be an integrable scaling function. The Fourier series of $h[n] = \langle 2^{-1/2}\phi(t/2), \phi(t-n)\rangle$ satisfies

$$\forall \omega \in \mathbb{R}, \quad |\hat{h}(\omega)|^2 + |\hat{h}(\omega+\pi)|^2 = 2, \tag{7.29}$$

and

$$\hat{h}(0) = \sqrt{2}. \tag{7.30}$$

Conversely, if $\hat{h}(\omega)$ is 2π periodic and continuously differentiable in a neighborhood of $\omega = 0$, if it satisfies (7.29) and (7.30) and if

$$\inf_{\omega \in [-\pi/2, \pi/2]} |\hat{h}(\omega)| > 0, \tag{7.31}$$

then

$$\hat{\phi}(\omega) = \prod_{p=1}^{+\infty} \frac{\hat{h}(2^{-p}\omega)}{\sqrt{2}} \tag{7.32}$$

is the Fourier transform of a scaling function $\phi \in \mathbf{L}^2(\mathbb{R})$.

penalize

$$\begin{cases} \sum_{w \in \{\frac{2\pi k}{N}, k=1,\ldots,N\}} (|\widehat{h}(w)|^2 + |\widehat{h}(w+\pi)|^2 - 2)^2 \\ (\sum_n h[n] - \sqrt{2})^2 \end{cases}$$

Mallat (2008) "A wavelet tour of signal processing: The sparse way"

Necessarily conditions on the low-pass filter

Theorem 8 *For any* $h(n) \in \ell^1$,

$$\sum_n h(n)\, h(n-2k) = \delta(k) \quad \textit{if and only if} \quad |H(\omega)|^2 + |H(\omega+\pi)|^2 = 2$$

penalize $\longrightarrow$ $\sum_k (\sum_n h[n]h[n-k] - \mathbb{I}_{k=0})^2$

Burrus et al 1997, Sec 6.3

Sufficient and necessary conditions on the high-pass filter

Theorem 7.3: *Mallat, Meyer.* Let ϕ be a scaling function and h the corresponding conjugate mirror filter. Let ψ be the function having a Fourier transform

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{g}\left(\frac{\omega}{2}\right) \hat{\phi}\left(\frac{\omega}{2}\right), \tag{7.52}$$

with

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi). \tag{7.53}$$

In the time domain,

$$g[n] = (-1)^n h[N-1-n]$$

(7.53) implies

$$\underbrace{\widehat{g}(0)}_{=\sum_n g[n]} = \widehat{h}^*(\pi) = 0$$

→ { parameterize $g[n] = (-1)^n h[N-1-n]$; penalize $(\sum_n g[n])^2$

Mallat (2008) "A wavelet tour of signal processing: The sparse way"

Understanding the wavelet loss

Sparsity of wavelet coefficients

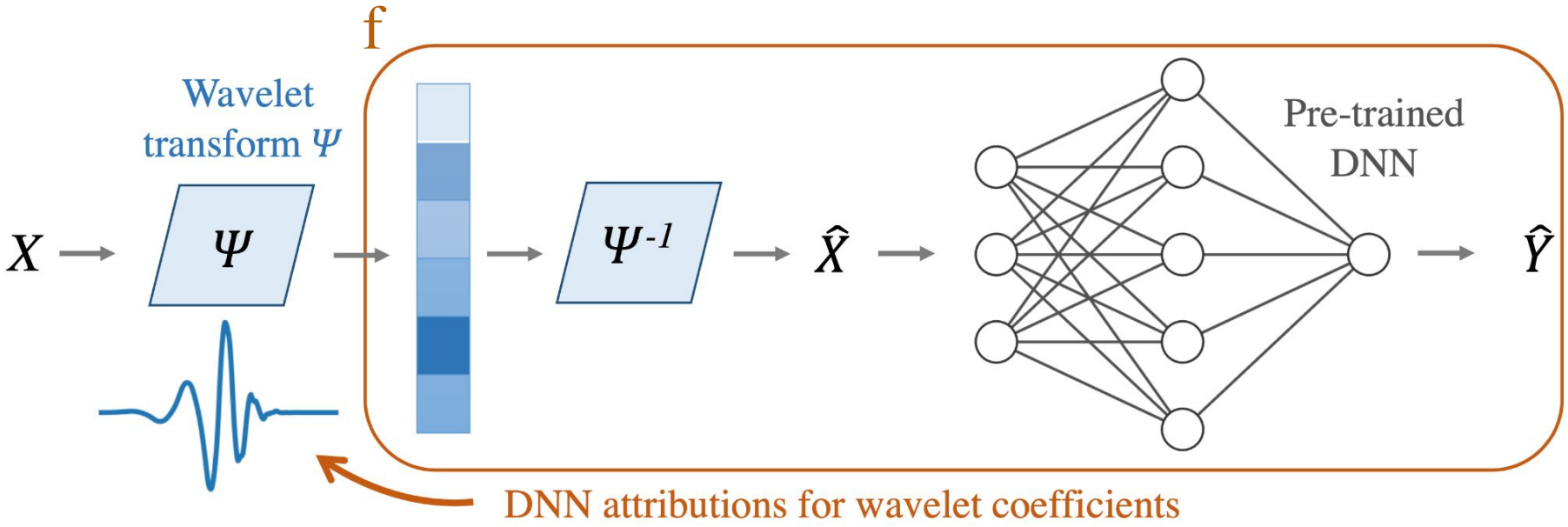
$$W(h,g,x_i;\lambda) = \lambda\|\Psi x_i\|_1 + (\sum_n h[n] - \sqrt{2})^2 + (\sum_n g[n])^2 + (\|h\|_2 - 1)^2$$
$$+ \sum_w (|\widehat{h}(w)|^2 + |\widehat{h}(w+\pi)|^2 - 2)^2 + \sum_k (\sum_n h[n]h[n-k] - \mathbf{1}_{k=0})^2$$

$$g[n] = (-1)^n h[N-1-n]$$

Number of scales is fixed

h, *g* are low-pass and high-pass filters

Understanding the TRIM loss

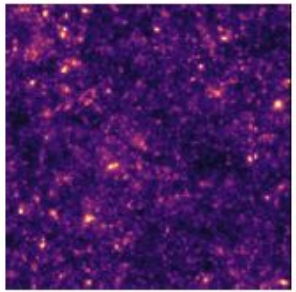


$$\sum_i \|\mathrm{TRIM}_f(\Psi x_i)\|_1$$

In the simple case,

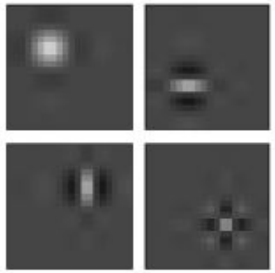
$$TRIM_f(\Psi x_i) = \left|\frac{\partial f(\Psi x_i)}{\partial(\Psi x_i)}\right|$$

Can identify key wavelet filters for cosmological parameter prediction

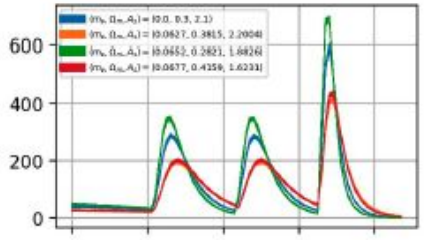


Cosmological mass maps

Resnet



Distilled wavelet



Peak counting predicts $\widehat{\Omega}_m$

Prediction error for Ω_m (RMSE)

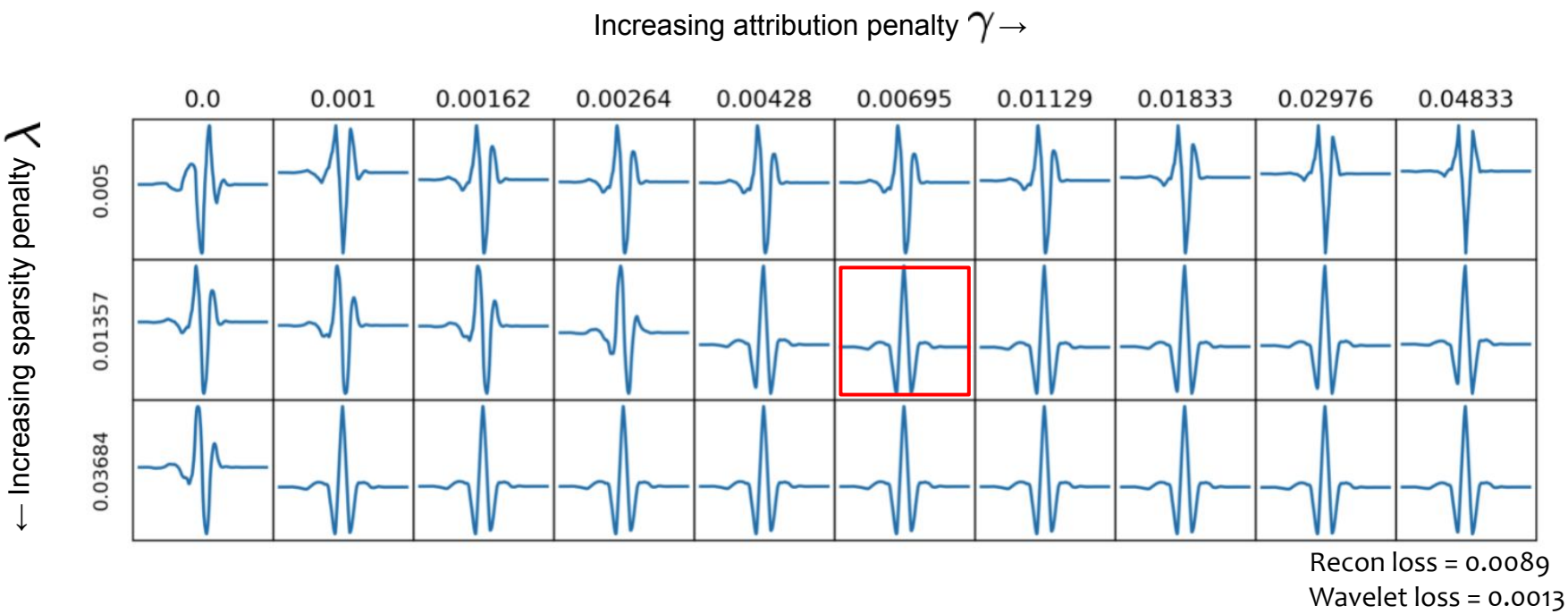
x 10^{-2}

AWD	Peak Height	Laplace	Roberts-Cross	DB5 Wavelet	Resnet	AWD w/o interp. loss
1.029	1.609	1.369	1.259	1.569	1.156	1.354

Note AWD outperforms other methods including Resnet with 4 filters (about 10^3 coefficients) instead of 10^7 parameters in Resnet18.

Ribli et al (2019) Nature Atron.

Inspecting the learned wavelet

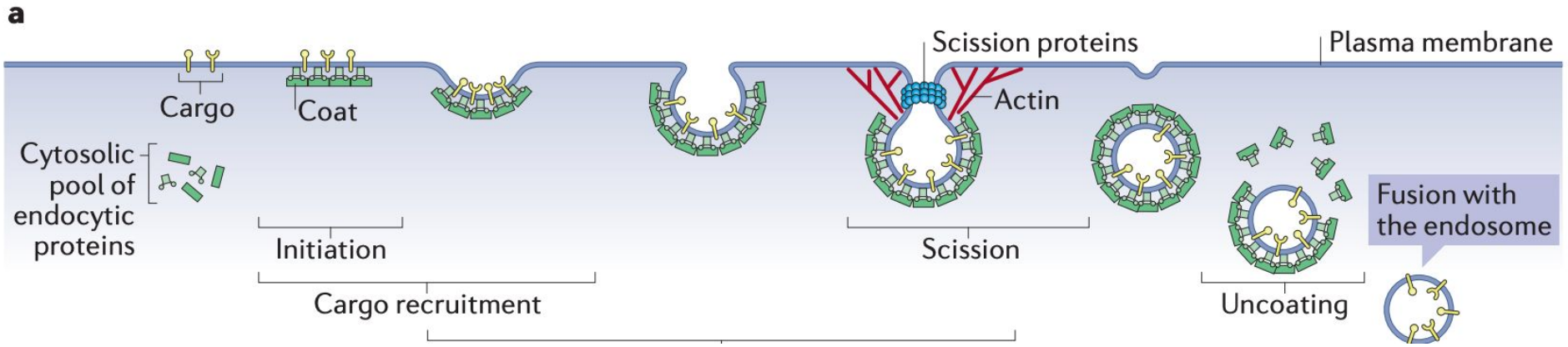


Recon loss = 0.0089
Wavelet loss = 0.0013

The difference of mass in the regions of space with large densities and their surroundings contains information about predicting Ω_m

Another case study of AWD

Clathrin-mediated endocytosis (CME)



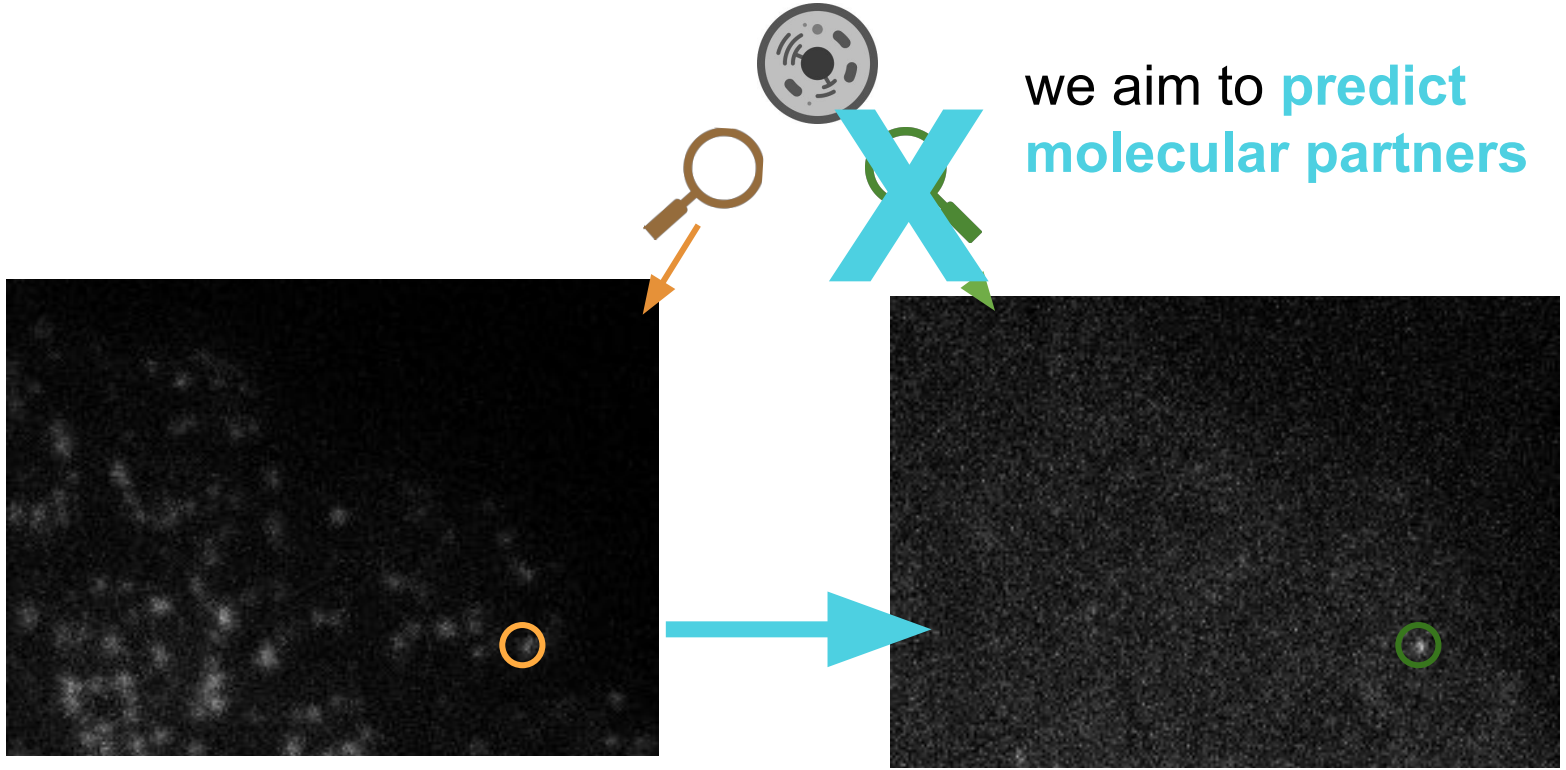
CME ”is a key process in vesicular trafficking that transports a wide range of cargo molecules from the cell surface to the interior.”

-- Kaksonen and Roux (2018) *Nature Reviews.*

https://www.nature.com/articles/nrm.2017.132

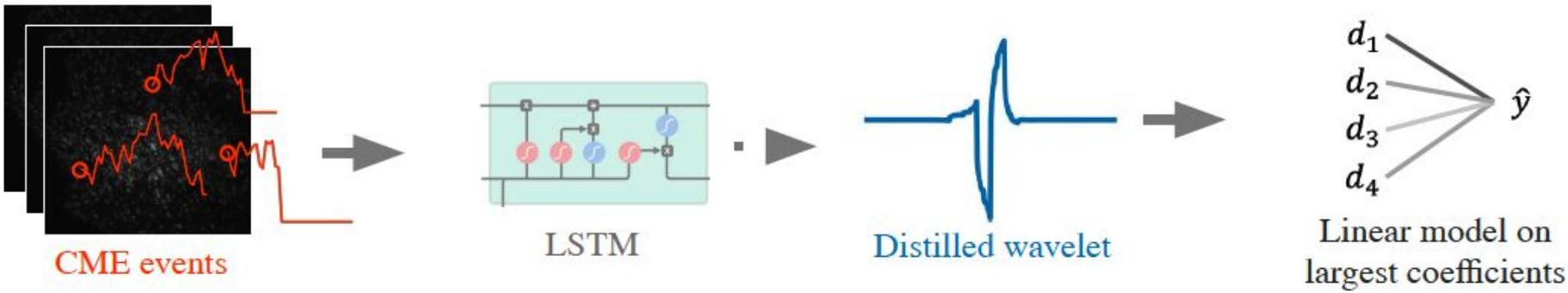
Tacking molecular partners is a central problem in cell biology

...but is experimentally difficult



we aim to **predict molecular partners**

Different experimental setups



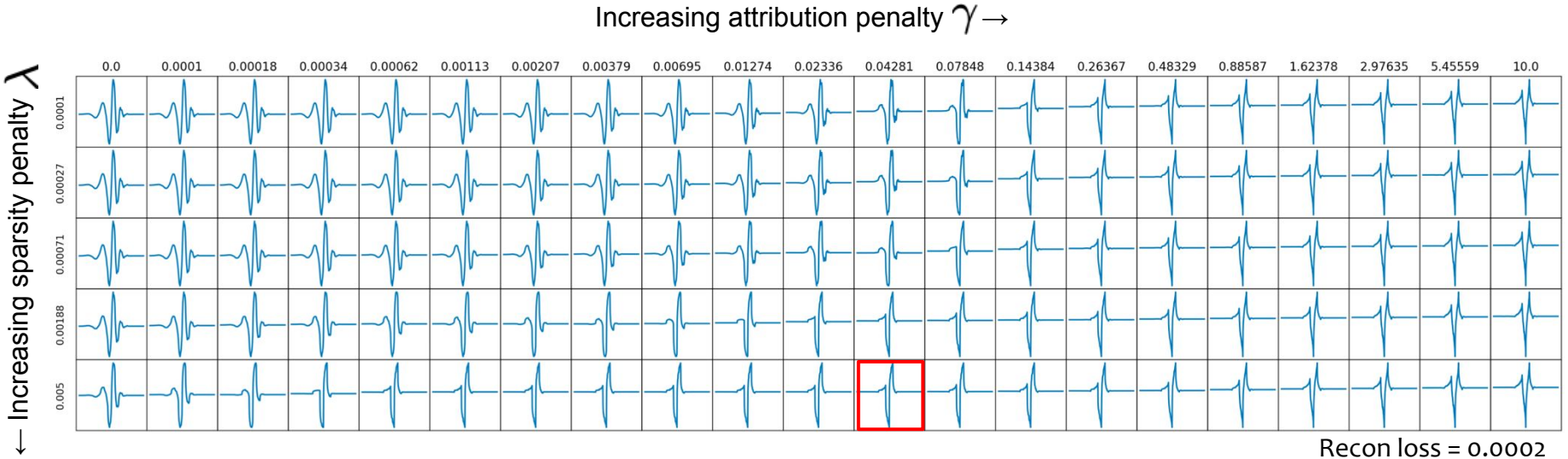
For each scale among 5 scales, we select 6 largest coefficients so 30 total, compared to 1000 parameters in LSTM

R^2 score

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

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

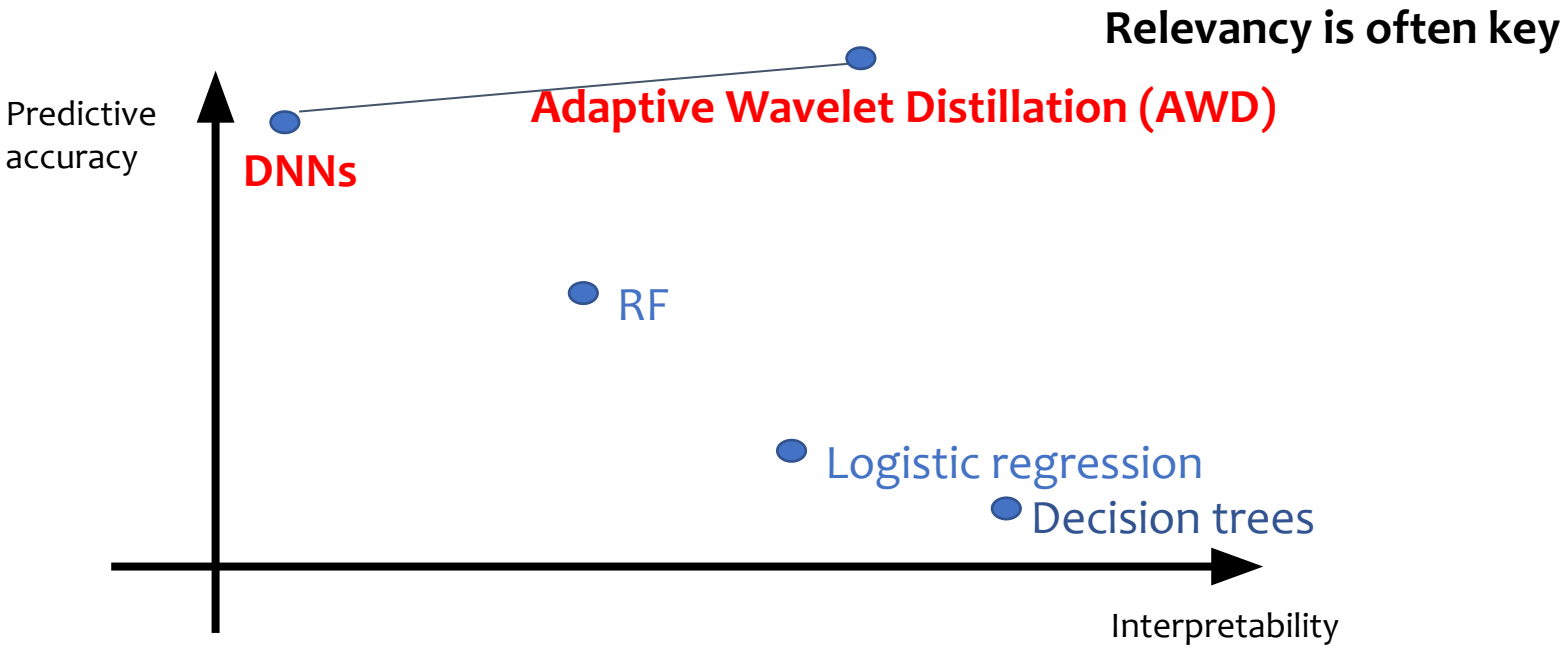
Inspecting the learned wavelet



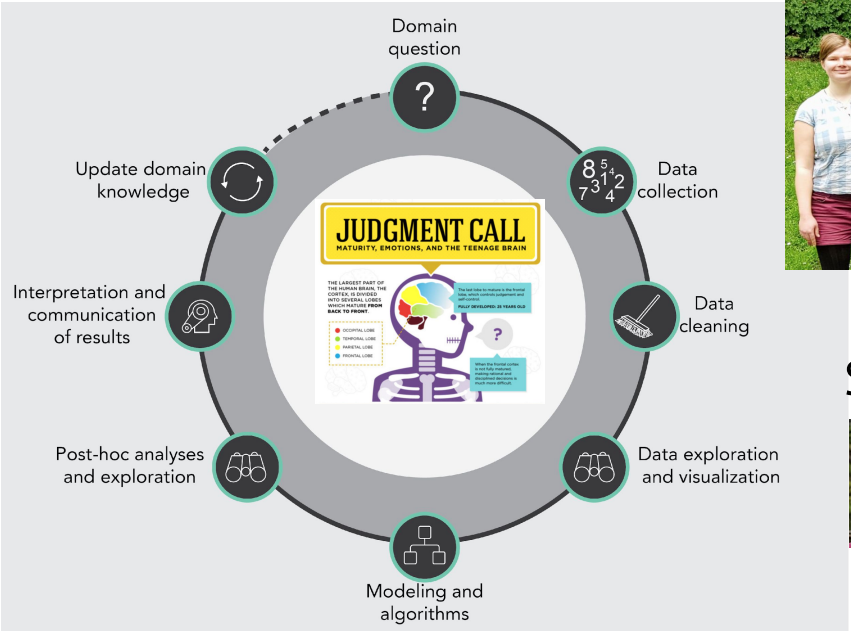
Recon loss = 0.0002
Wavelet loss < 0.00001

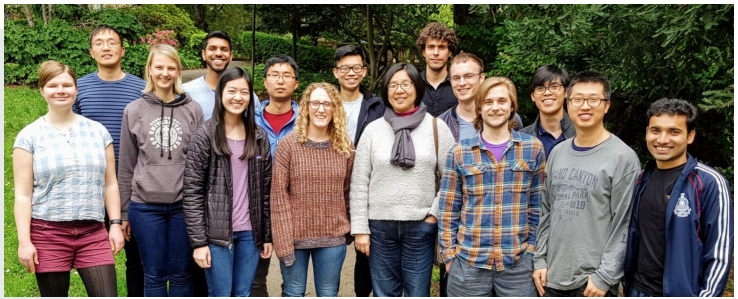
A large build up in clathrin fluorescence followed by sharp drop is a highly predictive signature of a successful CME event

D vs P for model-based interpretability



Recall Data Science Life Cycle (DSLC)





Scientist collaborators





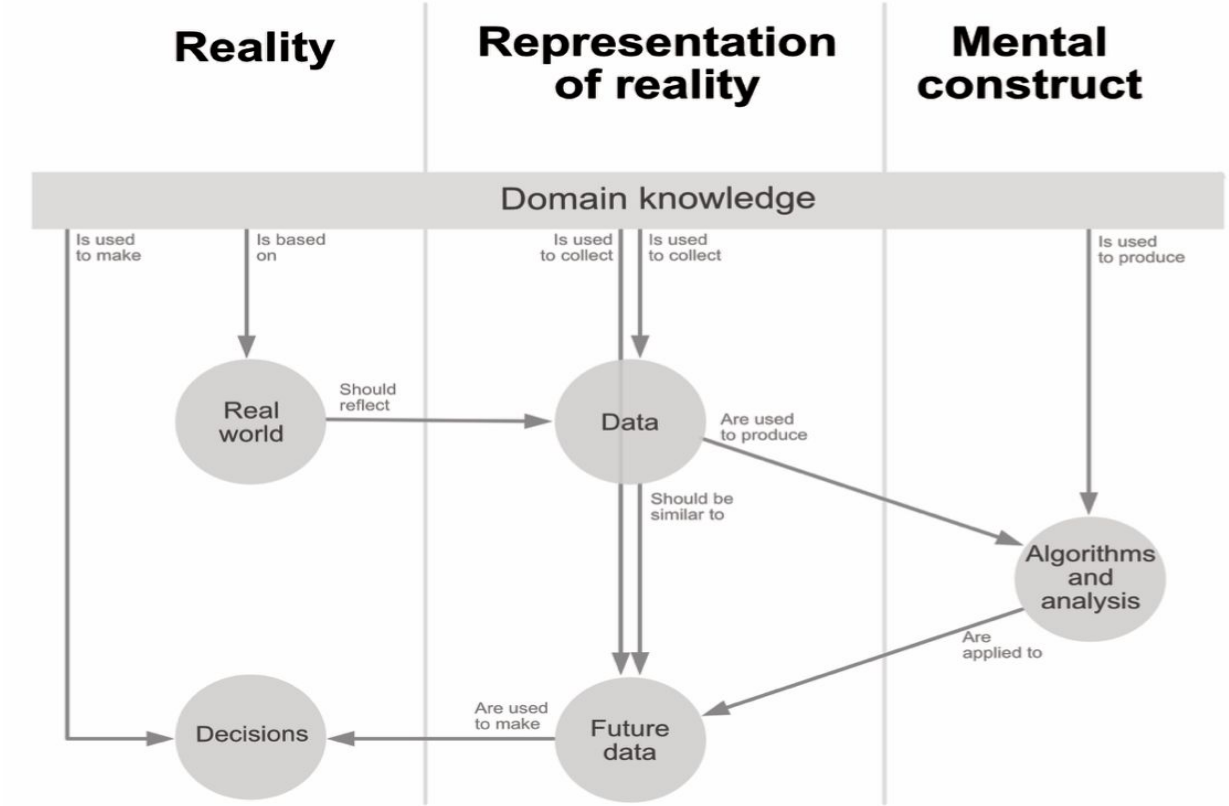



Image credits: R. Barter and toronto4kids.com

Part III: Trustworthy AI via quality-controlled DSLC

- “Best practices” to **maximize** the **promise** (prevention)
- Damage control to **reduce** the **danger** (risk management)

Any DNN model comes from a data science life cycle which starts from a three-realm conceptual process



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

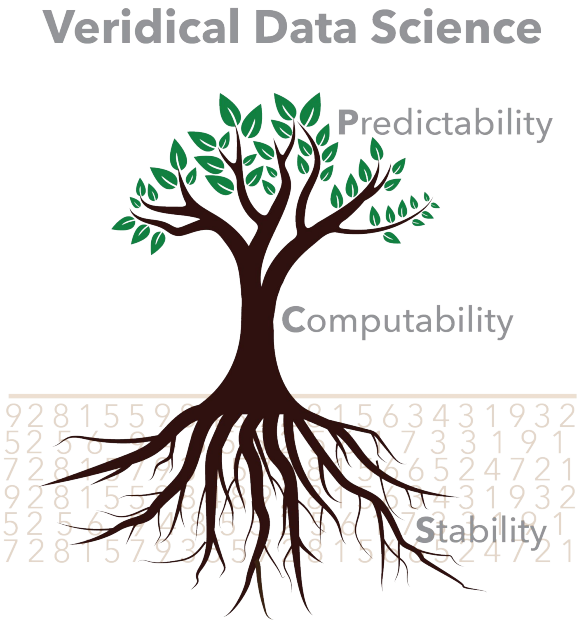
Extracts **reliable** and **reproducible information** from data,
with an enriched technical language to **communicate**
and **evaluate** empirical **evidence in context**
of human decisions and domain
knowledge

Realizes the promises and mitigates the dangers of AI.
It quality-controls DSLC.

PCS framework for **veridical data science**

It is based on
three principles of data science:

(**P**)redictability [ML and Stats]
(**C**)omputability [ML]
(**S**)tability [Stats, Control Theory, ...]

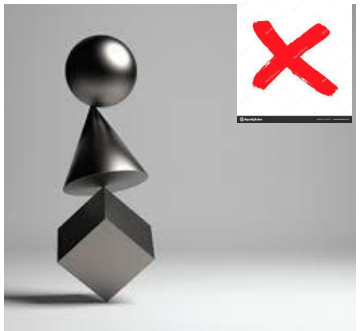


PCS unifies, streamlines, and expands ideas and best practices in **both** ML and Stats and beyond for the entire data science life cycle

The stability principle

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

- Yu (2013) [Stability]



This principle has been expanded to cover **every step of the data science life cycle** in Y. and Kumbier (2020) [Veridical data science]

Image credits:https://mobilitymgmt.com/articles/2019/05/01/core-stability.aspx
https://depositphotos.com/291132712/stock-illustration-x-red-mark-cross-sign.html

Stability principle is an ancient philosophy

"For true opinions, as long as they remain, are a fine thing and all they do is good, but they are not willing to remain long, and they escape from a man's mind, so that they are not worth much until one ties them down . . . That is why knowledge is prized higher than correct opinion, and knowledge differs from correct opinion in being tied down."

-- Plato, in the Meno

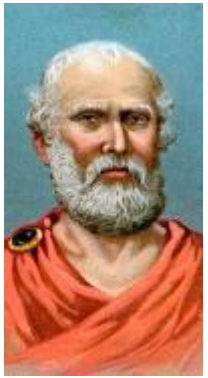
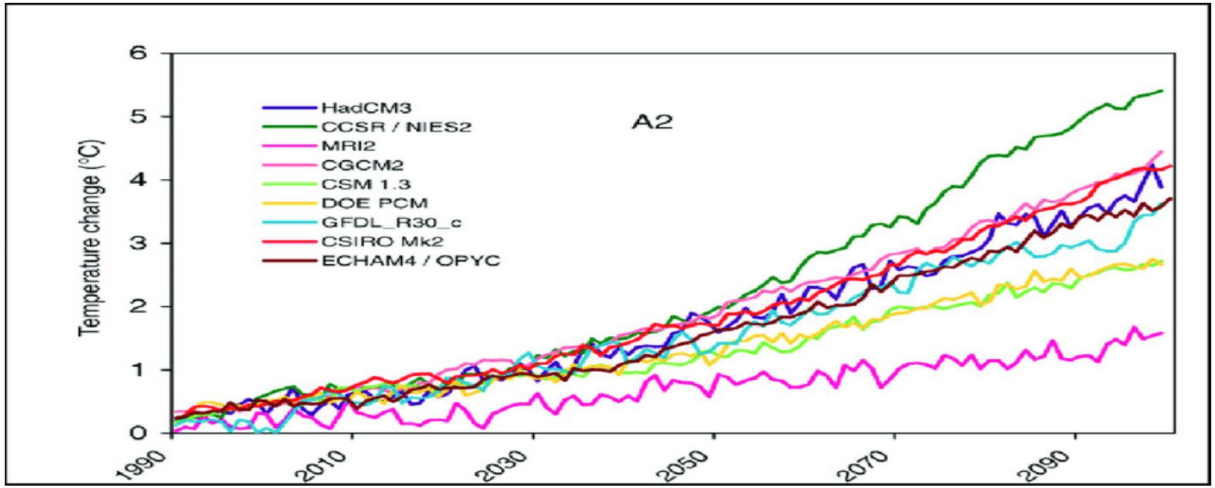


Image credit: wikipedia

Global mean temp. change: 9 leading models

Researcher to researcher (or team to team) perturbation

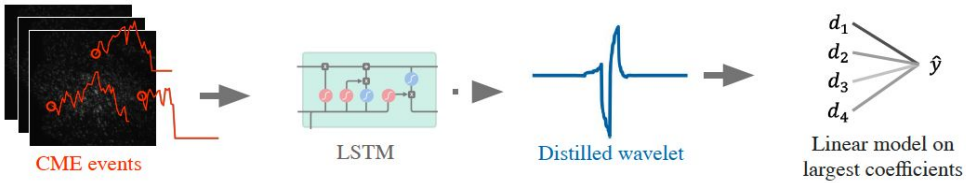
9 climate models



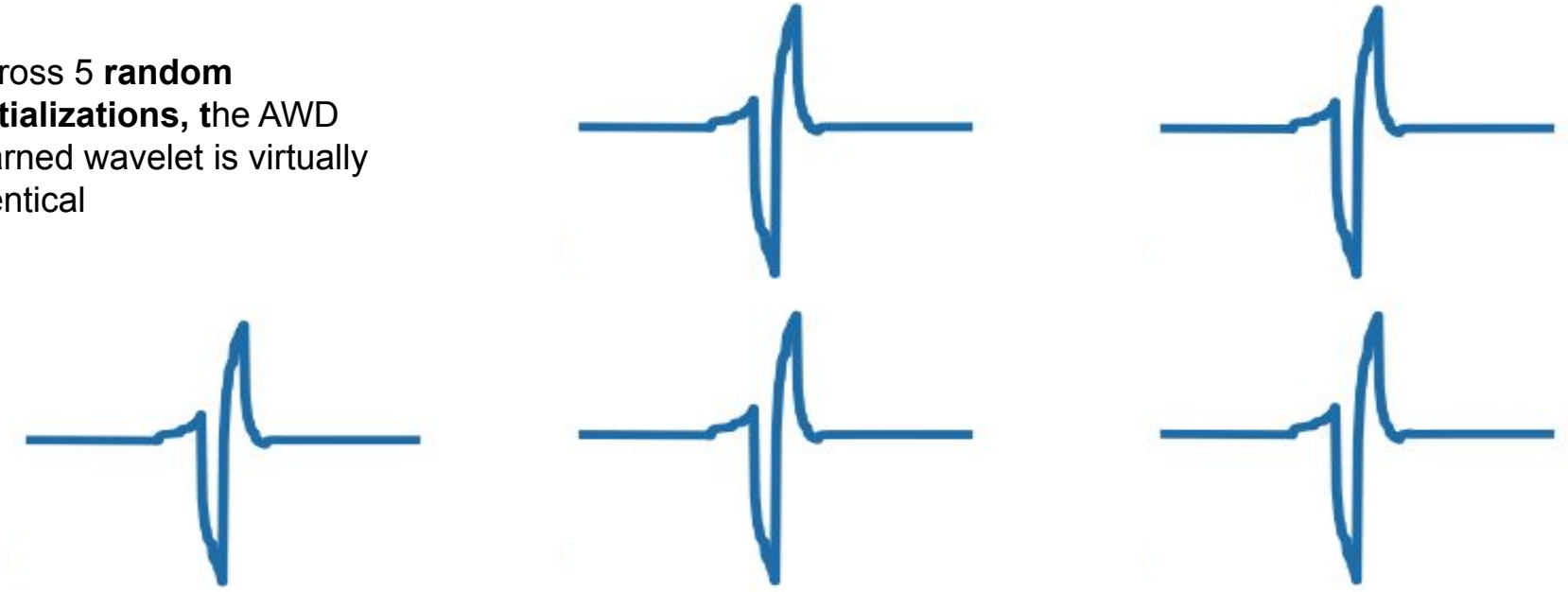
The change in global-mean temperature estimated by nine climate models forced by the SRES A2 emission scenario. (Source: IPCC TAR, Chapter 9)

Global mean-temp change

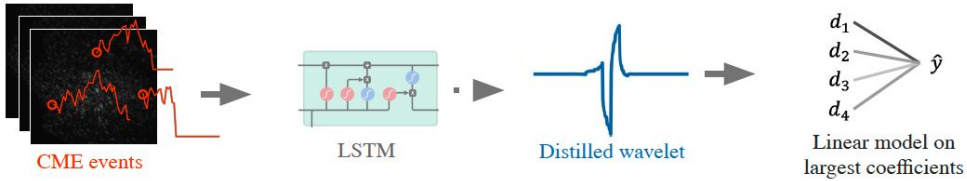
Molecule-partner prediction problem: **stability analysis** over algorithm perturbation



Across 5 **random initializations, t**he AWD learned wavelet is virtually identical



Stability analysis over data perturbation



Subsample only 80% of training data and averaged over 10 random seeds

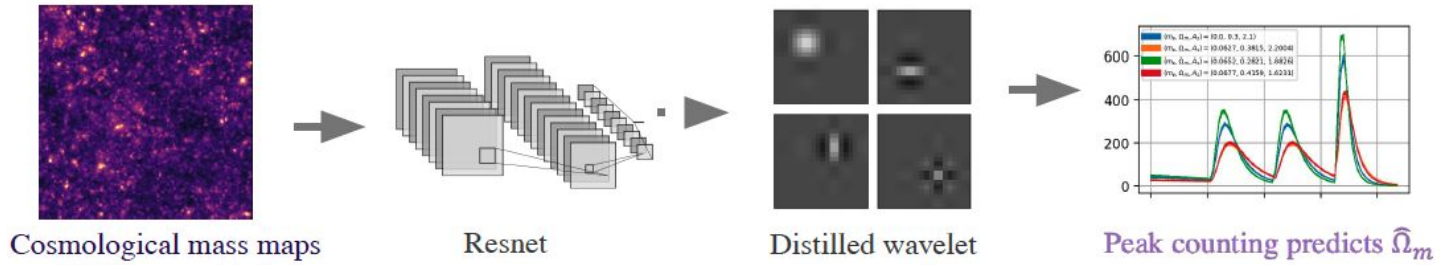
For each scale, we select 6 largest coefficients

R^2 score

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

AWD	DB5 Wavelet	LSTM
0.264 (0.01)	0.197	0.237

Cosmology problem: **stability analysis** over data perturbation



Cosmological mass maps | Resnet | Distilled wavelet | Peak counting predicts $\widehat{\Omega}_m$

Subsample only 80% of training data and averaged over 5 random seeds

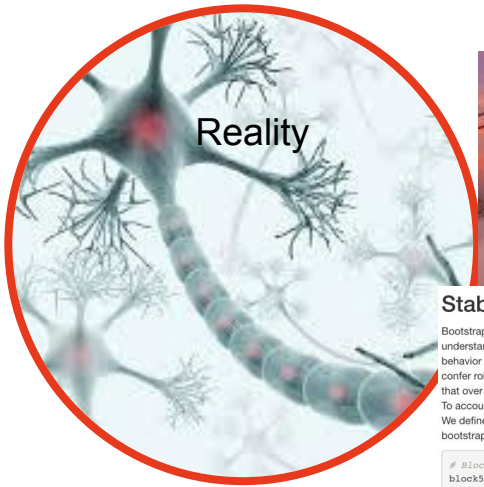
Prediction error for Ω_m (RMSE)

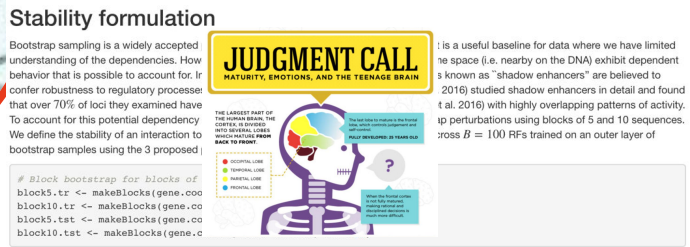
x 10^{-2}

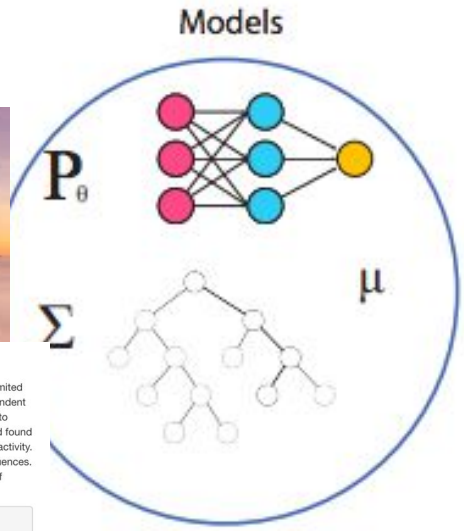
AWD	Peak Height	Laplace	Roberts-Cross	DB5 Wavelet	Resnet
1.029	1.609	1.369	1.259	1.569	1.156

PCS documentation to record all human judgment calls on

data collection and cleaning processes, domain knowledge, EDA/algorithm/interpretation and data perturbation choices, and reproducible codes in Rmarkdown or JupyterNotebook







8 PCS success stories from Yu Group

Methodology -- **adding stability to Lasso (CV), NMF, RFs**

- ESCV for predictive and stable Lasso
- staNMF for number of component selection in NMF
- Iterative random forests (iRF) for Boolean high-order predictive and stable interactions

Domain science -- **solving problems in neuroscience and medicine**

- DeepTune for characterizing V4 neurons
- staDISC for calibrated and stable subgroup discoveries in RCT
- staDTRIP for drug discovery
- PCS-stress-test for clinical decision rule development
- ...

PCS software developments

Design Principles:

Transparent (P)
Realistic (P)

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

Reproducible (S)

Veridical Flow

PCS-style data analysis made easy!

simChef

PCS-style simulations made easy!

To go with standardized PCS documentation

Future directions

Gaining novel scientific insights for other scientific and medical problems from ACD and AWD

Investigating which DNNs are well distilled by AWD

Connecting AWD with mechanistic models

Integrating PCS into data analysis protocols/pipelines at Hutchson Cancer Center and Joint Genome Institute of DoE

What to get into Data science?

Watch out for our book covering the whole DSLC

Yu and Barter
MIT Press
(plan: free on-line copy **in fall 2022; book in 2023**)

Veridical Data Science: A Book

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


Berkeley UNIVERSITY OF CALIFORNIA

What skills does the book teach?

Veridical Data Science (VDS) will teach the critical thinking, analytic, human-interaction and communication skills required to effectively formulate problems and find reliable and trustworthy solutions. VDS explains concepts using visuals and plain English, rather than math and code.
The primary skills taught are:

Critical thinking

Readers will learn to:
Formulate answerable questions using the data available
Scrutinize all analytic decisions and results
Document all analytic decisions
Appropriate common techniques to unfamiliar situations
Deal with real, messy data

Technical skills

Data processing
Data cleaning
Exploratory Data Analysis
Data merging

Algorithmic
Dimension reduction
Clustering
Least Squares & ML
Regularization

Stability-based inference
Inference
Causal Inference
Perturbation Intervals
Trustworthiness Statements

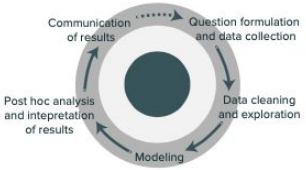
Communication

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

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

Core guiding principles for the book

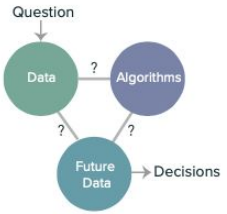
The DS Lifecycle



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

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

Three realms



Readers will learn to view every data problem through the lens of connecting the three realms:
(1) the question being asked and the data collected (and the reality the data represents)
(2) the algorithms used to represent the data
(3) future data on which these algorithms will be used to guide decision-making.
Guiding the reader to connect the three realms is a means of guiding the reader through the data science lifecycle.

PCS framework

The PCS framework provides concrete techniques for finding evidence for the connections between the three realms.
Predictability: if the patterns found in the original data also appear in withheld or new data, they are said to be predictable. If an analysis or algorithm finds predictable patterns, then these patterns are likely to be capturing real phenomena.
Computability: algorithmic and data efficiency and scalability is essential to ensuring that the results and solutions (e.g. a predictive algorithm) can be efficiently applied to new data.
Stability: minimum requirement for reproducibility. If results change in the presence of minor modifications of the data (e.g. via perturbations) or human analytic decisions, then there might not be a strong connection between the analysis/algorithms and the reality that underlies the data.

Intended Reader/Audience

Anyone who wants to learn the intuition and critical thinking skills to become a data scientist or work with data scientists.
Neither a mathematical nor a coding background is required.
VDS could form the basis of a semester- or multi-semester-long introductory data science university course, either as an upper-division undergraduate or early graduate-level course.

Interested? Get in touch!

Bin Yu
Email: binyu@stat.berkeley.edu
Website: https://www.stat.berkeley.edu/~binyu/Site/Welcome.html

Rebecca Barter
Email: rebeccabarter@berkeley.edu
Website: www.rebeccabarter.com
Twitter: @rlbarter

Thank you!

SIMONS FOUNDATION

CHAN ZUCKERBERG BIOHUB

Weill Neurohub

1. Definitions, methods and applications in interpretable Machine Learning
   Paper: https://www.pnas.org/content/116/44/22071 (PNAS, 2019)

2. Adaptive Wavelet Distillation (AWD) code
   https://github.com/Yu-Group/adaptive-wavelets
   Paper: https://arxiv.org/abs/2107.09145 (NeurIPS, 2021)

3. Interpretable ML/DL review paper (including CD, ACD and AWD) https://arxiv.org/abs/2108.06847

4. Veridical data science (PCS)
   Paper: https://www.pnas.org/content/117/8/3920 (PNAS, 2020)
   Breiman Lecture video on PCS:
   https://slideslive.com/38922599/veridical-data-science
   Updated slides:
   https://www.stat.berkeley.edu/~binyu/ps/papers2020/Breiman19-NeurIPS-yu.pdf