Interpretability
… the who, what, why, and how

Sanmi Koyejo

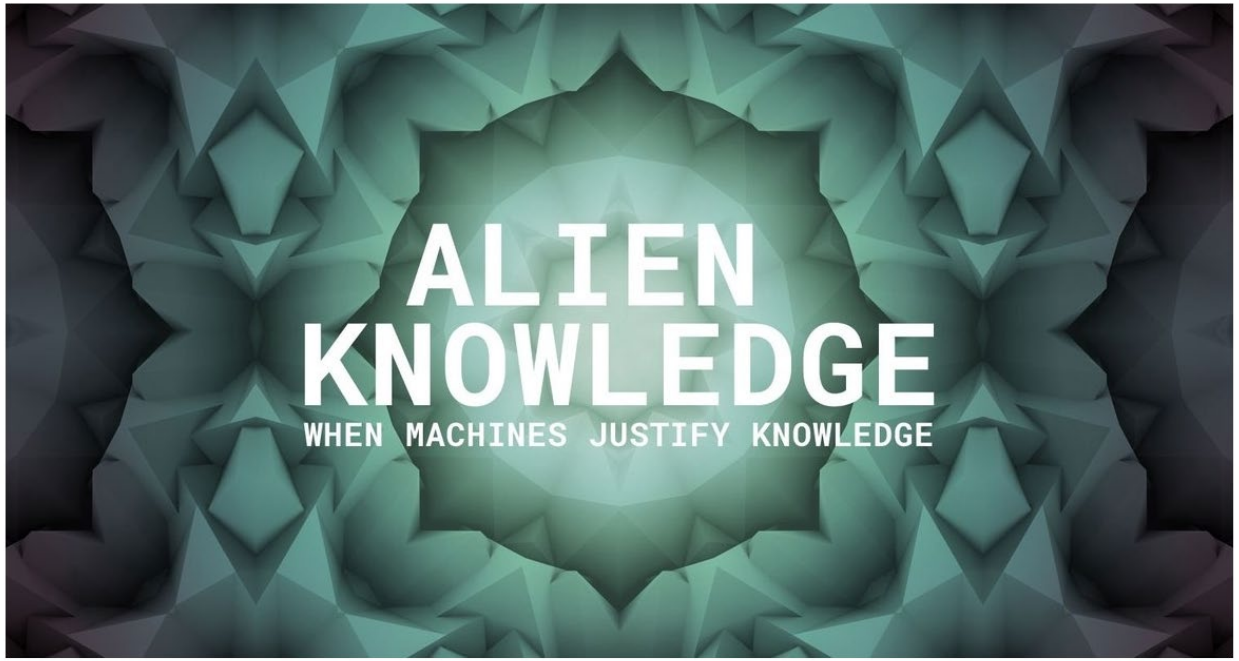
Goals of this presentation

- Learn why interpretability is important, and to whom
- Become familiar with the most common kinds of interpretable models
- Learn some of the techniques for post-hoc interpretability

We routinely construct complex models from simple building blocks

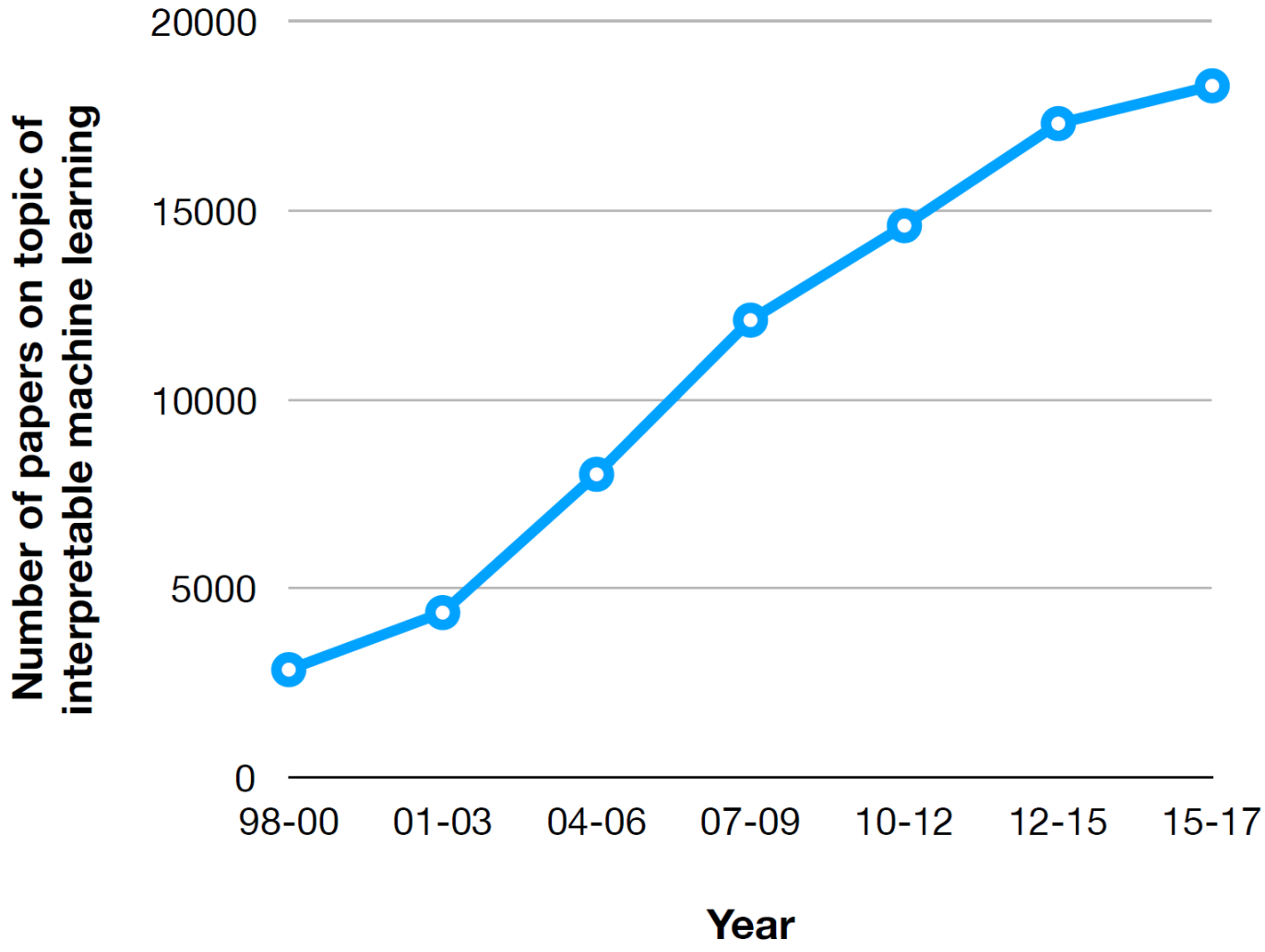
DAVID WEINBERGER BACKCHANNEL 04.18.17 08:22 PM

OUR MACHINES NOW HAVE KNOWLEDGE WE'LL NEVER UNDERSTAND



- Source: https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/

Interpretable ML an active research area with lots of open questions!



- Source: Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Overview

1

Who cares about interpretability? (should you care?)

2

What is interpretable machine learning?

3

Why do we care about interpretability?

4

How does one build interpretable models?

Who cares about interpretability?

ML experts, i.e., algorithm designers

Debugging

ML Safety & Robustness

Domain experts, i.e., expert consumers of ML

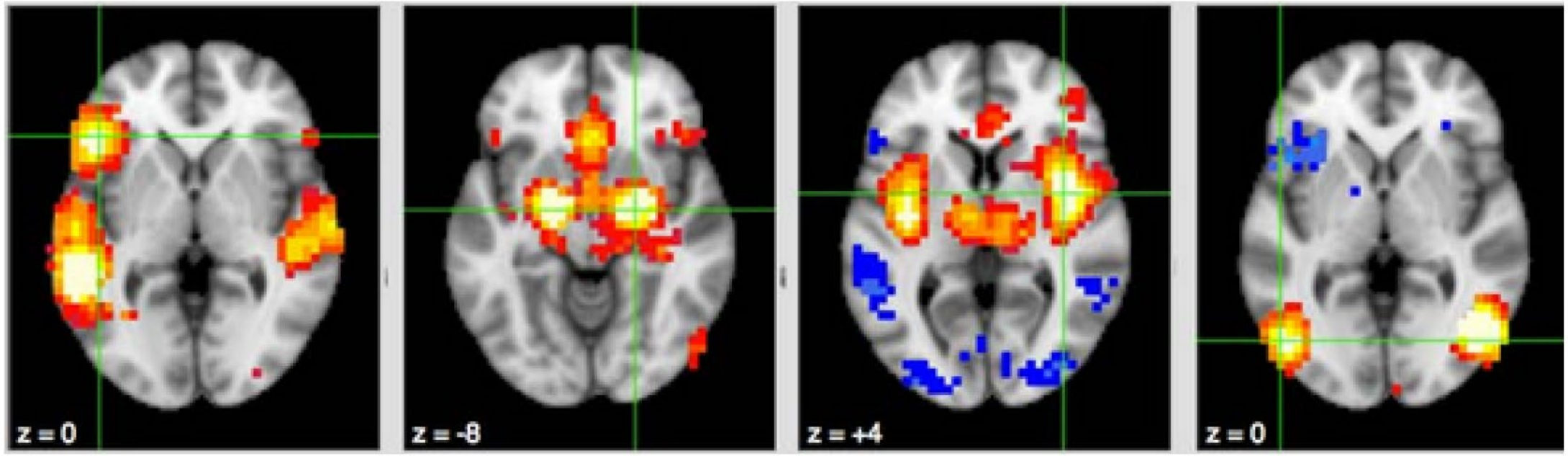
Causal discovery
e.g. scientists

Causal predictions
e.g. healthcare

Detecting model failures and safety concerns e.g. robotics

Decision-support
e.g. “data science” in business applications

Example, sparse models in neuroimaging



Wu, Park, **K.**, Pillow (2014), Dependent relevance determination for smooth and structured sparse regression

Example: dangers of using a predictive model for causal predictions

- Cost-effective Health Care (CEHC) built models to predict probability of death for patients [Cooper et al. 97]
- Model assigns lower risk to asthma patients, why?
- **Model:** Asthma => lower risk for pneumonia
- **Ground truth:** Asthma => Aggressive treatment => lower risk for pneumonia

Caruana, Lou, Gehrke, Koch, Sturm. Elhadad (2015), Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission

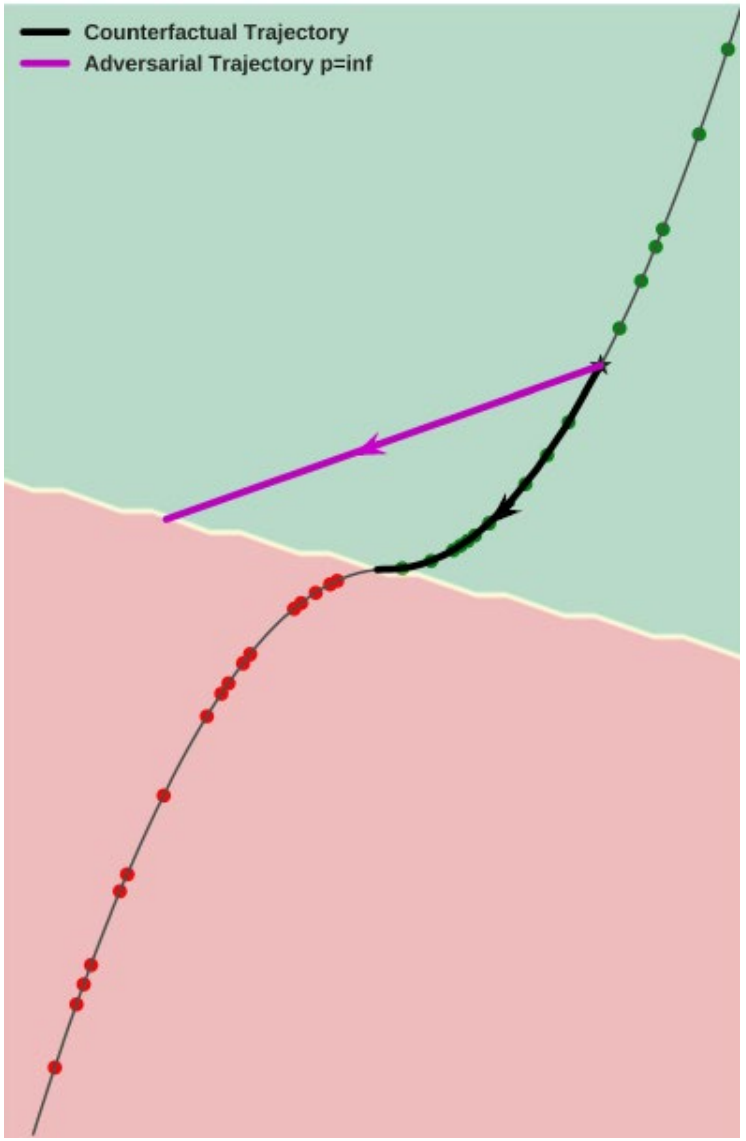
ML for decision support

- “I prefer an *inaccurate interpretable* model to an *accurate un-interpretable* model”
  - finance executive at “AI in business workshop”, Illinois, 2019

- Most “data-science” business applications use ML for pattern discovery, not for predictions
- Discovered patterns must be intelligible to the decision maker, help them make a case, or justify a decision
- Some companies have ML decision support built into their business model e.g. personalized clothing recommendations

ML consumers, i.e., (increasingly) everyone

- Fairness and accountability through transparency
- Right to explanation, e.g. , GDPR (Goodman & Flaxman, 2016), US FICO scores
- Individual recourse, i.e., “how do I change the prediction outcome?” e.g., credit rating, loan decisions, hiring decisions



What is interpretable machine learning?

We want models to be interpretable, explainable, intelligible, transparent, understandable, …

- Predictability and simulatability
- Decomposability and modularity
- Trustworthiness

Lipton (2016), The Mythos of Model Interpretability

Predictability and Simulatability

- Can a human predict the model behavior? model mistakes? Can a human “simulate” the model behavior?

- e.g. complex physics models, where one can often abstract-out simple rules that roughly predict outcomes

$$E = mc^2$$

- e.g. low-dimensional / sparse linear models, decision trees

$$y = 3x_1 - .2x_5 + 10$$

Decomposability and modularity

- Can a human decompose the model into “interpretable” components?
- e.g. block diagrams in most engineering disciplines.
- e.g. low-dimensional / sparse linear models, decision trees

Lage, Chen, He, Narayanan, Kim, Gershman and Doshi-Velez, (2019), An Evaluation of the Human-Interpretability of Explanation

Some other desirable properties

Trust

- Are we comfortable with the model for the prediction setting?
- Well-calibrated uncertainty, i.e., can tell you when its “not sure”
- Automation bias: people will sometimes follow model predictions when they are obviously wrong

Informativeness

- Does the model reveal auxiliary information useful for decision making?

Stability

- Statistical and optimization properties

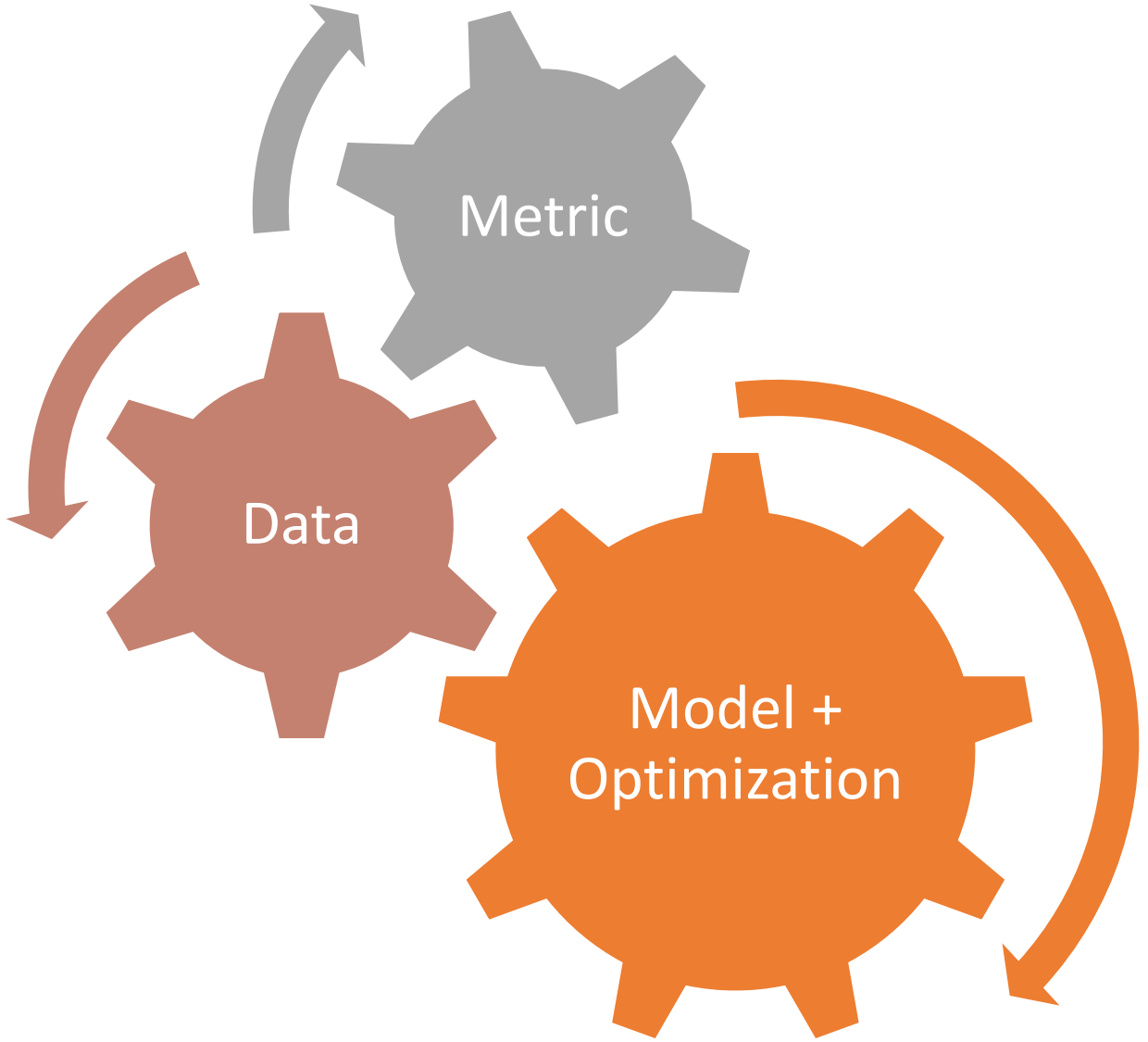
Evaluation sometimes reveals unexpected human behavior

- Participants shown a clear model with a small number of features were better able to simulate the model's predictions.
- no significant difference in multiple measures of trust across conditions.
- increased transparency hampered people's ability to detect when a model has made a sizeable mistake.
- Conclusion: important to study how models are presented to people and empirically verify that interpretability achieves its stated goals

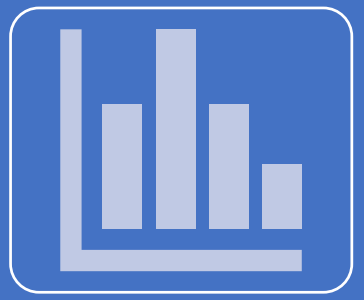
Poursabzi-Sangdeh, Goldstein, Hofman, Wortman-Vaughan, Wallach (2018) Manipulating and Measuring Model Interpretability

Why do we care about interpretability?

The machine learning toolbox



ML in your textbook



Data are from the distribution of interest

- Train and test samples are “similar”
- Usually, assume data samples are IID

We know the “right” model to use

- Model class is a good approximation of “truth”
- Optimization approximates the right answer

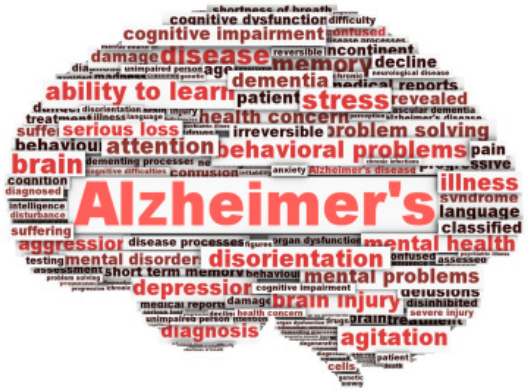
We care that the model is accurate on average

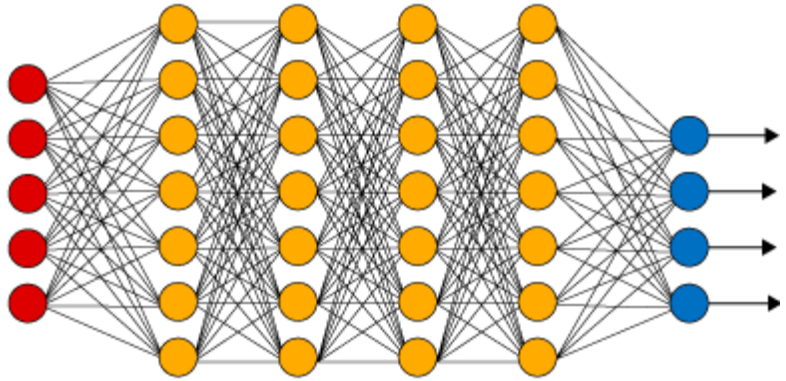
- Model usage is fully defined by the metric
- Only care about prediction accuracy

The real-world ML is a bit messier!

https://xkcd.com/

Quick aside on building a model



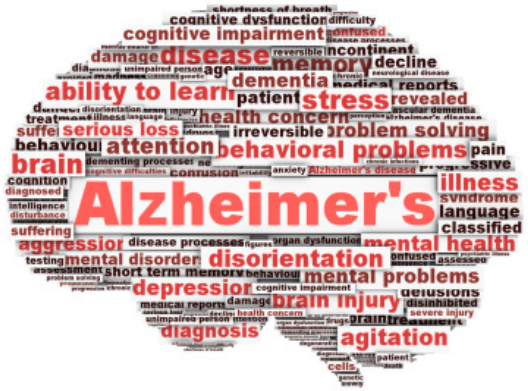


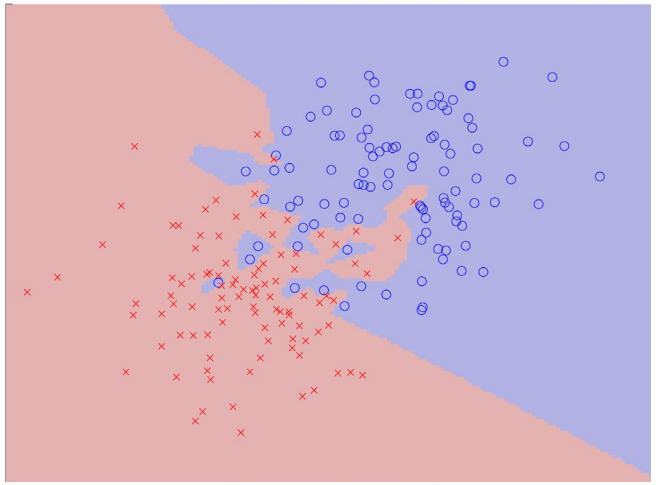
- Accuracy = 94.1%

Positive = Healthy; Negative = Alzheimer's

- False positive rate = Predict healthy when patient has Alzheimer's = 90%
- False negative rate = Predict Alzheimer's when patient is healthy = 5%

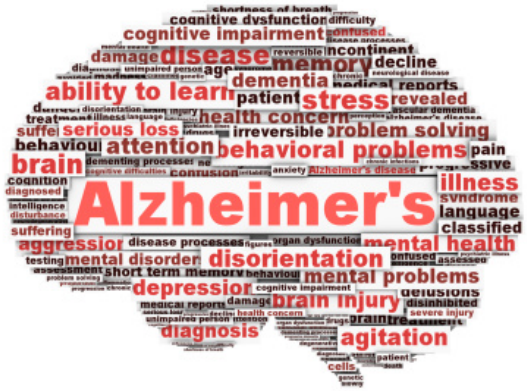
Image Source: https://becominghuman.ai/deep-learning-made-easy-with-deep-cognition-403fbe445351

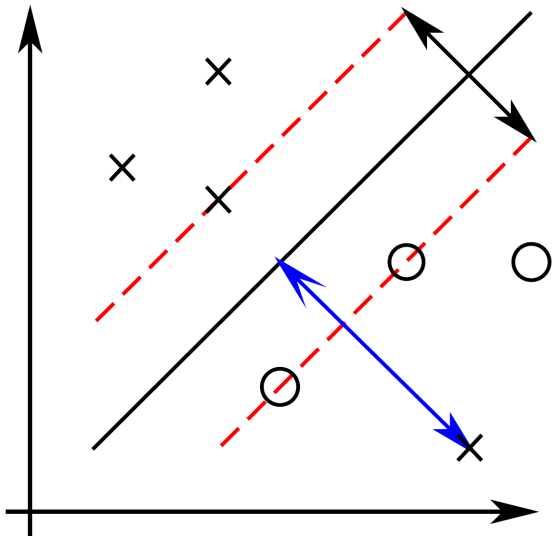




- Accuracy = 89.6%

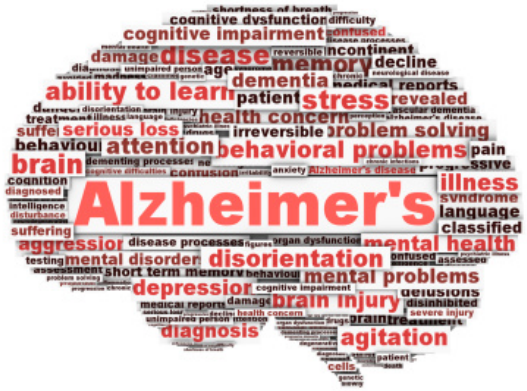
- False positive rate = Predict healthy when patient has Alzheimer's = 50%
- False negative rate = Predict Alzheimer's when patient is healthy = 1%





- Accuracy = 80.1%

- False positive rate = Predict healthy when patient has Alzheimer's = 10%
- False negative rate = Predict Alzheimer's when patient is healthy = 20%

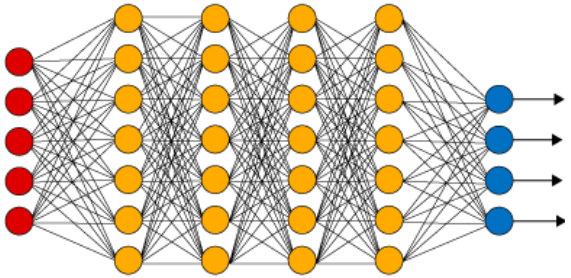


Always Predict Healthy

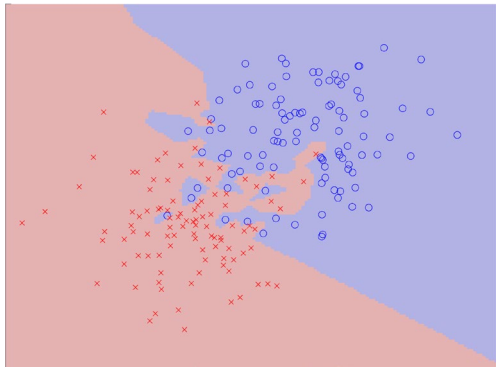
- Accuracy = 99%

- Prevalence of Alzheimer's disease is <1% of the population*
- False positive rate = Predict healthy when patient has Alzheimer’s = 100%
- False negative rate = Predict Alzheimer’s when patient is healthy = 0%

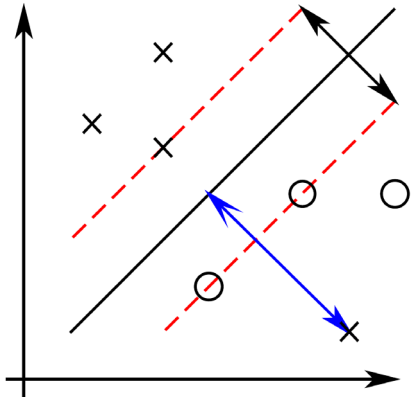
* https://www.alz.org/facts/



- 94.1% Accuracy
- 90% false positives
- 5% false negatives



- 89.6% Accuracy
- 50% false positives
- 1% false negatives

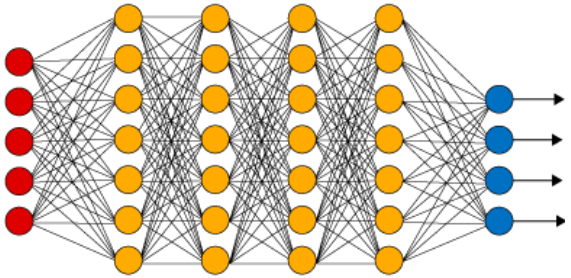


- 80.1% Accuracy
- 10% false positives
- 20% false negatives

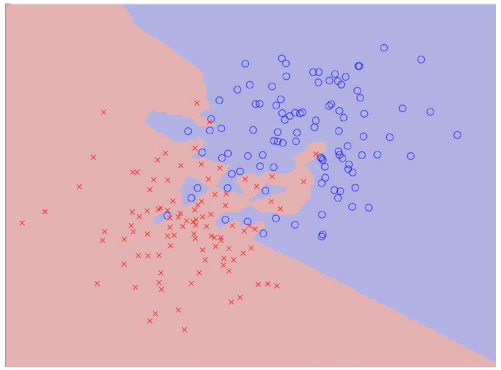
Always Predict Healthy

- 99% Accuracy
- 100% false positives
- 0% false negatives

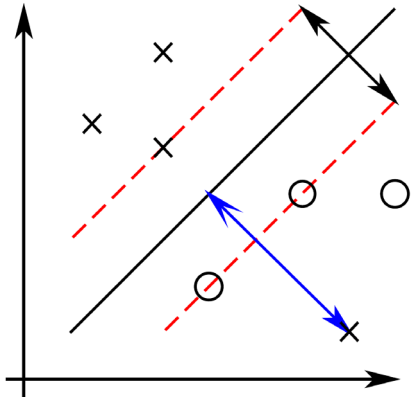
Which ML model is best?



- 94.1% Accuracy
- 90% false positives
- 5% false negatives



- 89.6% Accuracy
- 50% false positives
- 1% false negatives

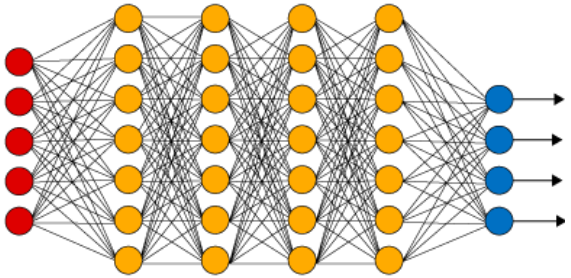


- 80.1% Accuracy
- 10% false positives
- 20% false negatives

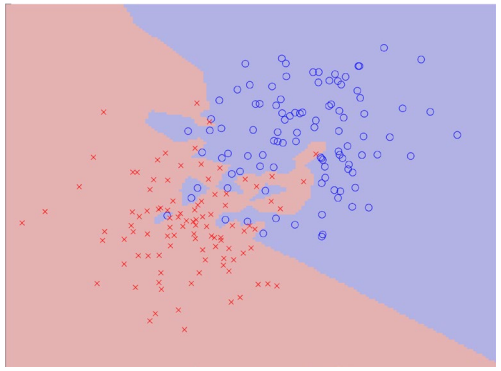
Always Predict Healthy

- 99% Accuracy
- 100% false positives
- 0% false negatives

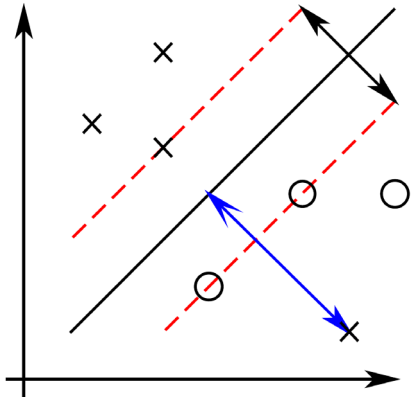
What data are the model(s) trained on? Does the training data reflect the population?



- 94.1% Accuracy
- 90% false positives
- 5% false negatives



- 89.6% Accuracy
- 50% false positives
- 1% false negatives



- 80.1% Accuracy
- 10% false positives
- 20% false negatives

Always Predict Healthy

- 99% Accuracy
- 100% false positives
- 0% false negatives

Should you trust these predictions?

Data are not from the distribution of interest

- Distribution shift between train and deployment
  - This is the norm, not the exception!
- Data are rarely independent and identically distributed (IID)
  - Most of ML uses convenience samples
  - Collection process often introduces correlation
  - Possibility of adversaries distorting data or models
- Can interpretability help detect data and training issues?

We do not know the right model to use

- Model class may not be a good approximation of “truth”
  - Challenging in practice, we rarely know the right model family
  - We often try lots of models based on convenience
  - Good general-purpose inductive biases are rare (c.f. convolution)
- Optimization may not approximate the right answer
  - Most interesting model classes are non-convex (c.f. neural networks)
  - Simple optimization seems to work most of the time, but sometimes unclear why (c.f. inductive bias)
- Can interpretability help detect issues due to model quality and optimization performance?

We care that the model is accurate on average + other properties

- Metric often chosen by default, may not capture real world utility
- Usually care about a lot more than good predictions
  - Causality: does the model recover causal mechanisms?
  - Trust: can I “trust” the model predictions?
  - Decision support: does the model information help me make better decisions?
- Can interpretability help capture other model properties we care about?

Interpretability is often a proxy solution for an incomplete problem specification

How does one build interpretable models?

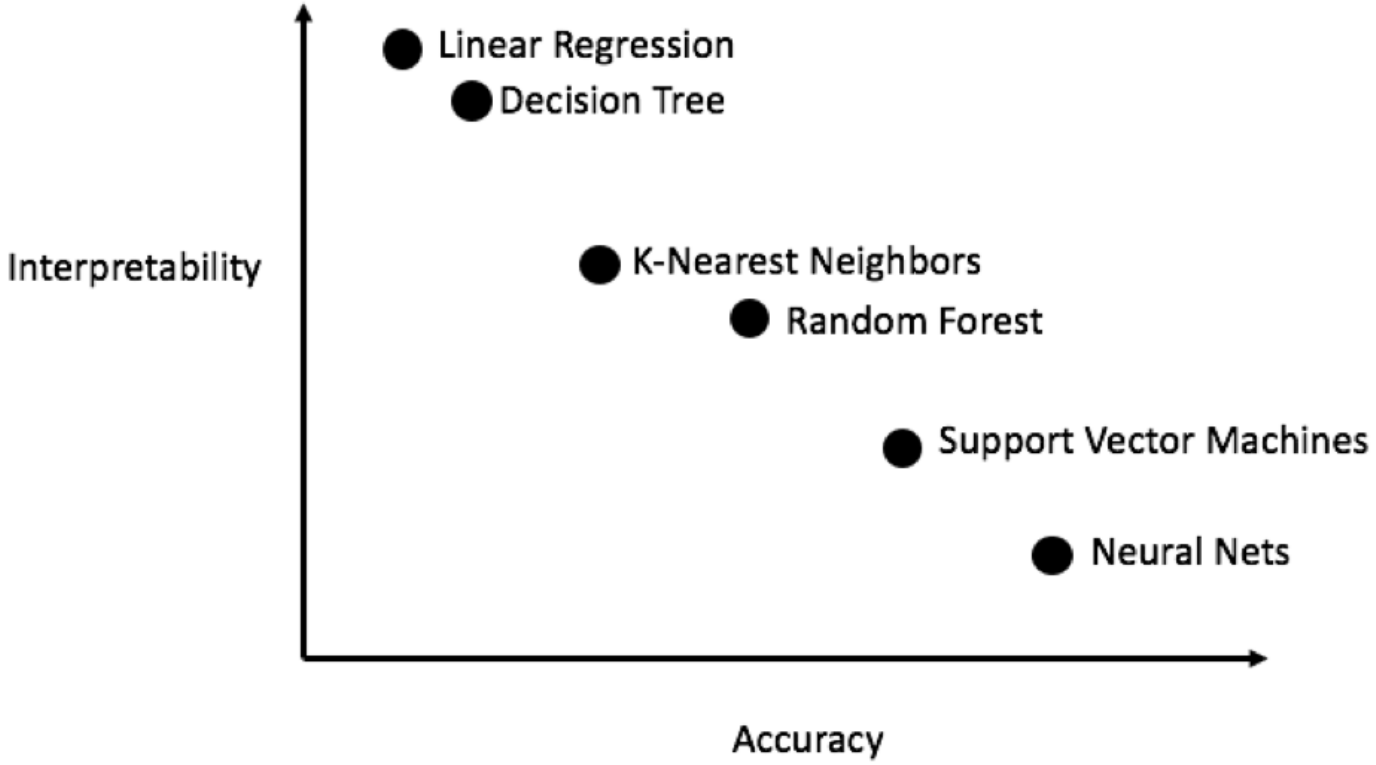
Transparent models and post-hoc explanations

Transparent Models

- Linear models
- Decision trees, Rule lists, Rule sets
- Case-based models, i.e., retrieve similar examples. “Explainable” by analogy
- Regularizers & constraints:
  - sparsity
  - monotonicity

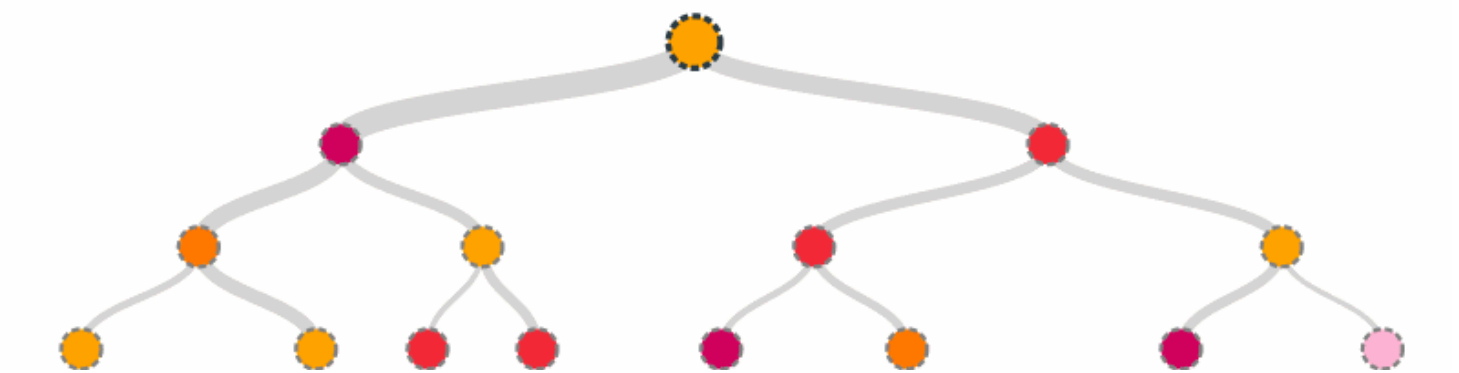
Post-hoc explanations

- Feature-wise:
  - Feature ablation
  - Saliency maps
  - Trained explanations
  - LIME
  - TCAV
- Sample-wise:
  - Prototypes and critics
  - Influential examples



Is there a fundamental tradeoff between interpretability and model complexity?

Source: https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9



Are all transparent models interpretable?

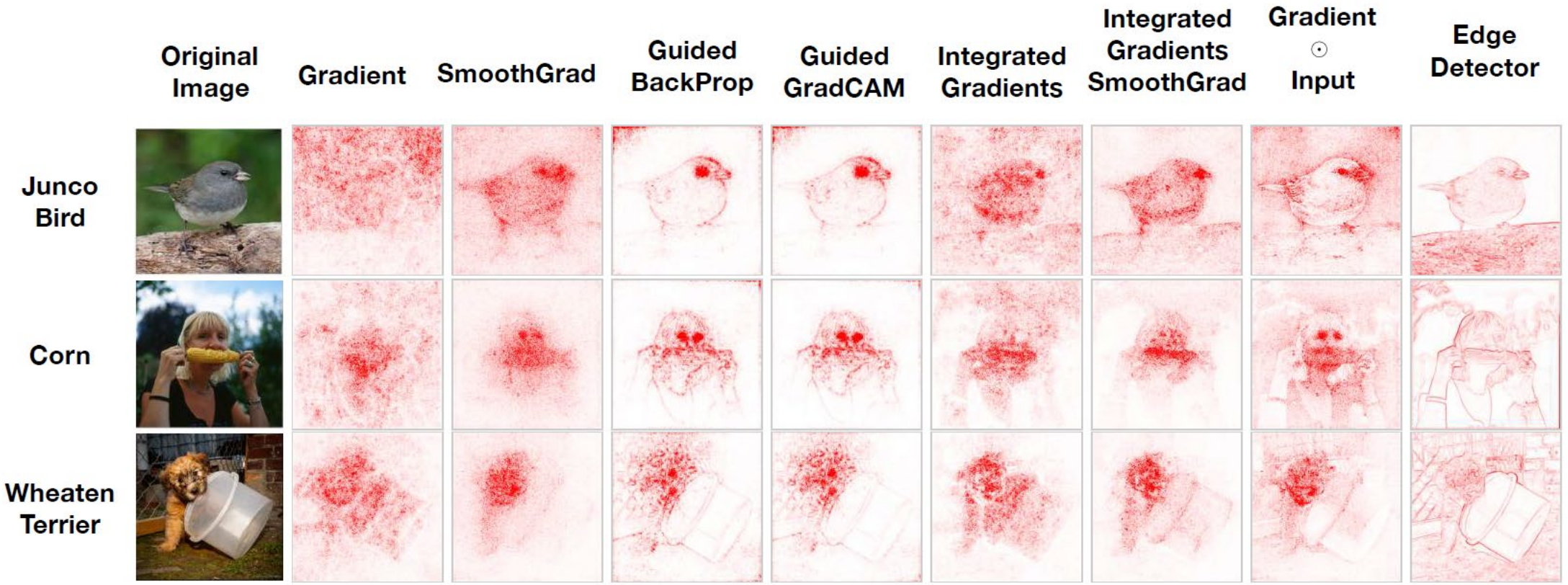
Source: https://blog.bigml.com/2013/05/07/a-sunburst-of-insight/

Post-hoc Explanations

Ablation test

- Train with/without each feature.
- Return the change in model without feature vs. general case
- Pros:
  - Conceptually straightforward
- Cons:
  - Does not account for feature correlations

Saliency maps; local model approximation



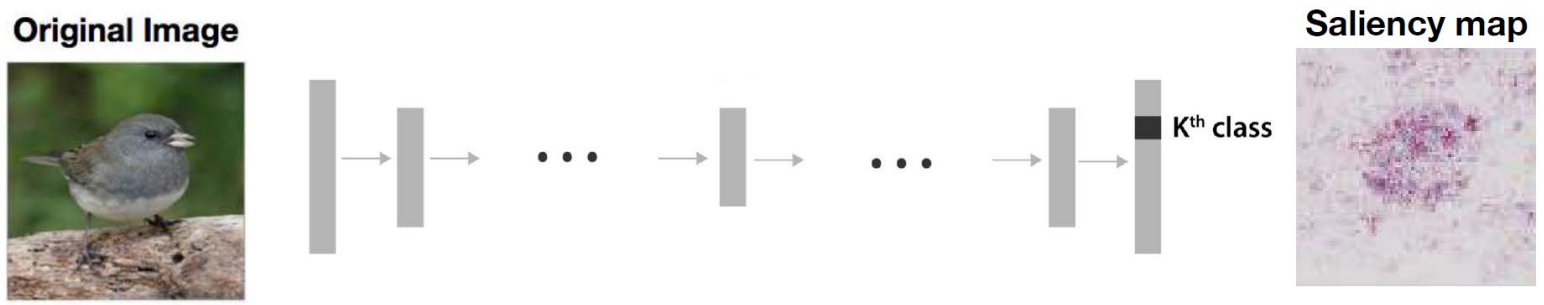
Adebayo, Gilmer, Muelly, Goodfellow, Hardt, Kim, (2018), Sanity checks for Saliency Maps

Saliency Map

- Gradient of the prediction wrt. input for an image
- Generalizes the weight vectors of a linear model
- Easy to compute -- a single backward pass through the model.

Simonyan, Vedaldi, Zisserman, (2013), Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Can be unreliable. Here we compute similar results when applied to random network

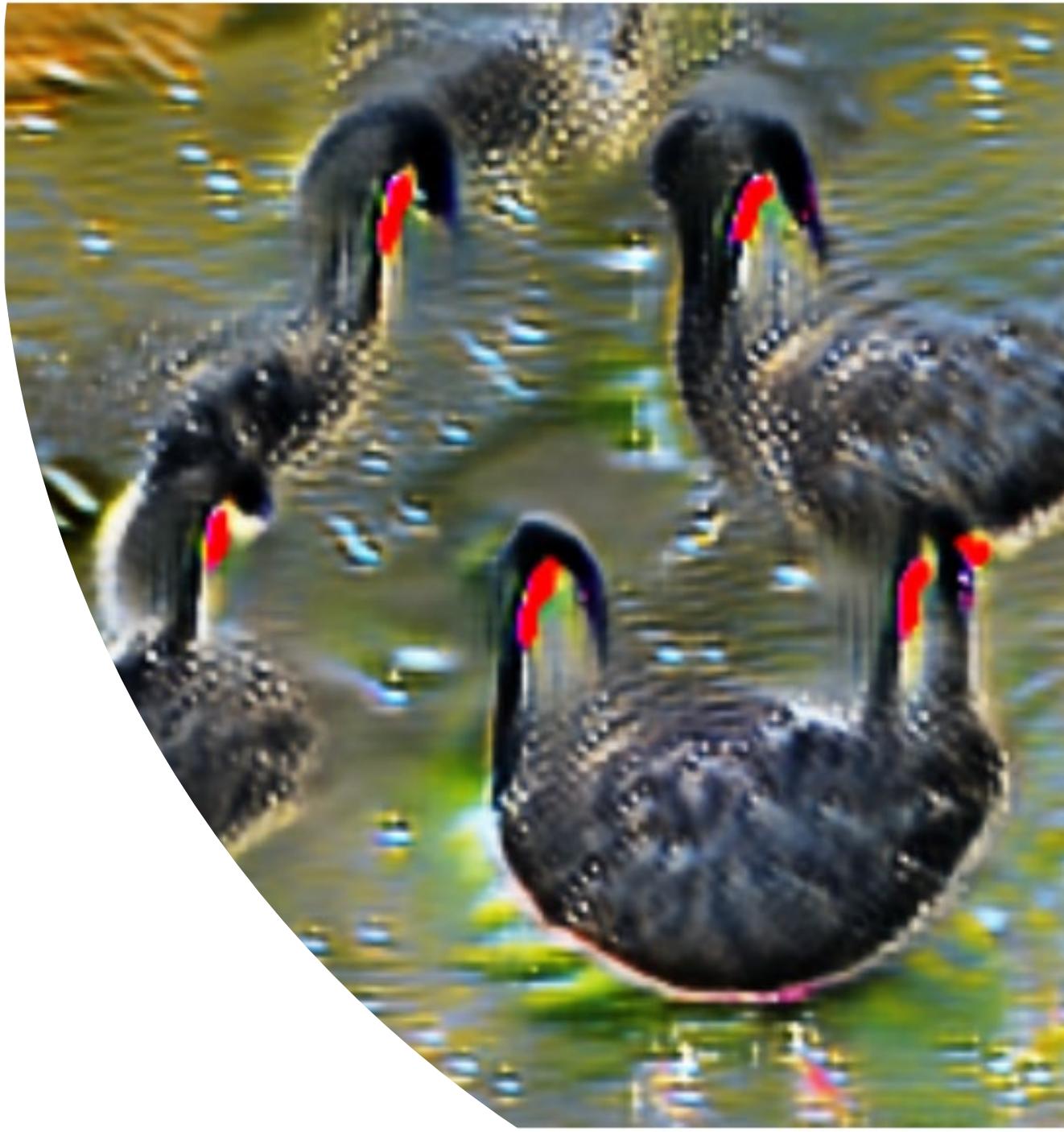


Randomized weights!
Network now makes garbage prediction.

Adebayo, Gilmer, Muelly, Goodfellow, Hardt, Kim, (2018), Sanity checks for Saliency Maps

Model inversion / activation maximization

- Activation maximisation for class neurons
  Erhan, Bengio, Courville (2009) Visualizing higher-layer features of a deep network.
- Activation maximization using empirical prior, deconvnet
  Zeiler, Fergus (2014) Visualizing and understanding convolutional networks
- Activation maximization and saliency Deep Inside Convolutional Networks
  Simonyan, Zisserman, Vedaldi (2014) Visualising Image Classification Models and Saliency Maps
- Inversion at different depths, natural image prior
  Mahendran, Vedaldi (2015) Understanding deep image representations by inverting them



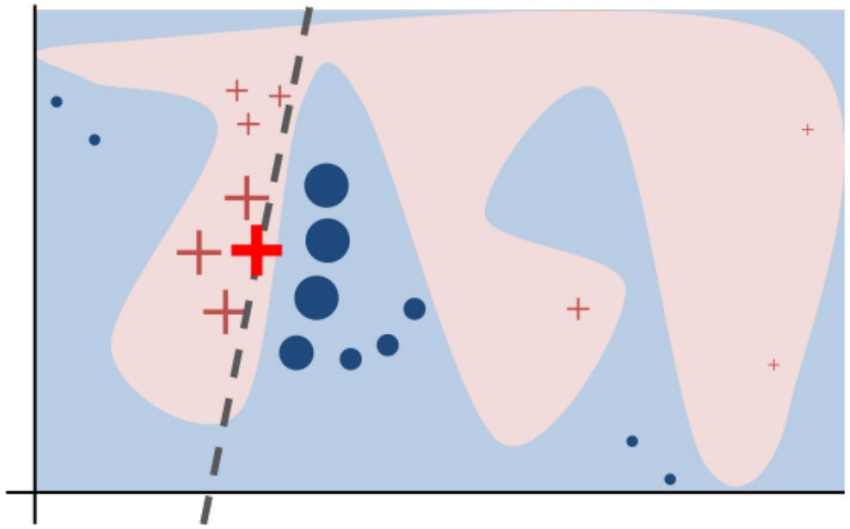
For instance, by combining feature visualization (*what is a neuron looking for?*) with attribution (*how does it affect the output?*), we can explore how the network decides between labels like **Labrador retriever** and **tiger cat**.

Olah et. Al. (2018), The Building Blocks of Interpretability

Several floppy ear detectors seem to be important when distinguishing dogs, whereas pointy ears are used to classify "tiger cat".

CHANNELS THAT MOST SUPPORT ...	LABRADOR RETRIEVER			...	TIGER CAT		
feature visualization of channel							
hover for attribution maps →							
net evidence	1.63	1.51	1.19		1.32	1.54	1.72
for "Labrador retriever"	1.22	1.24	1.32		-0.70	-1.24	-0.43
for "tiger cat"	-0.40	-0.27	0.13		0.62	0.30	1.29

Local Interpretable Model-agnostic Explanations (LIME)

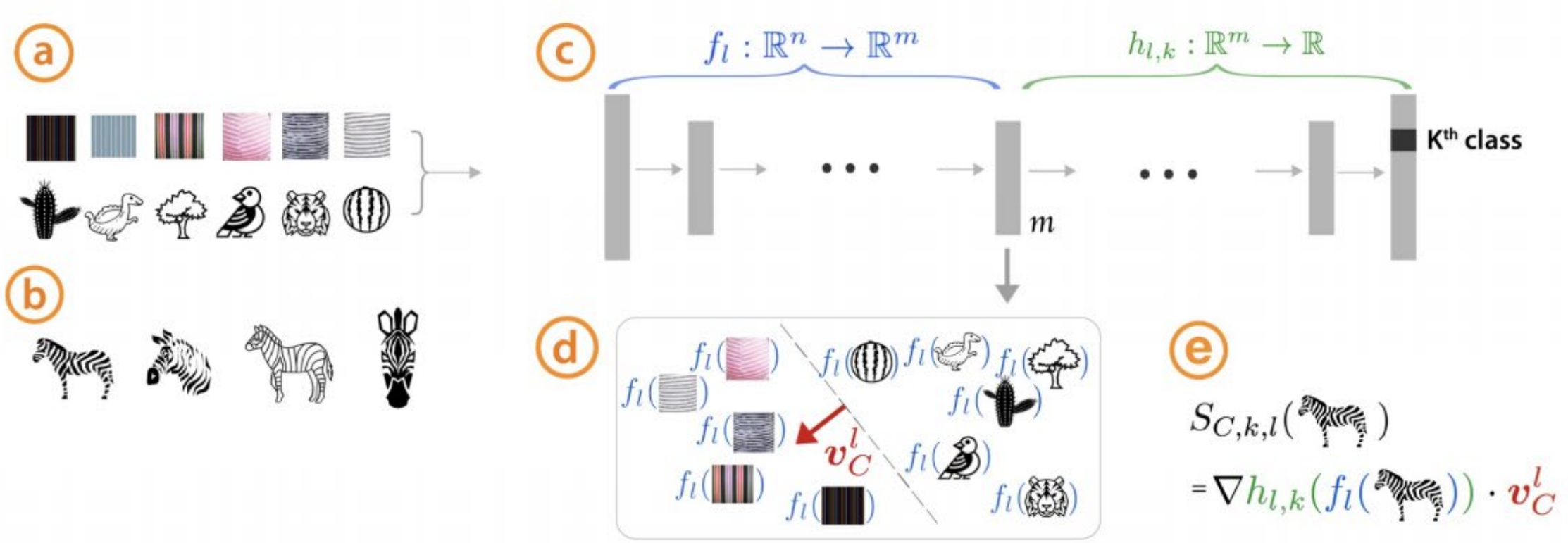


(a) Original Image

(b) Explaining *Electric guitar*

Ribero, Singh, Guestrin (2016) “Why Should I Trust You?” Explaining the Predictions of Any Classifier

Testing with Concept Activation Vectors



Kim, et al. (2018) Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

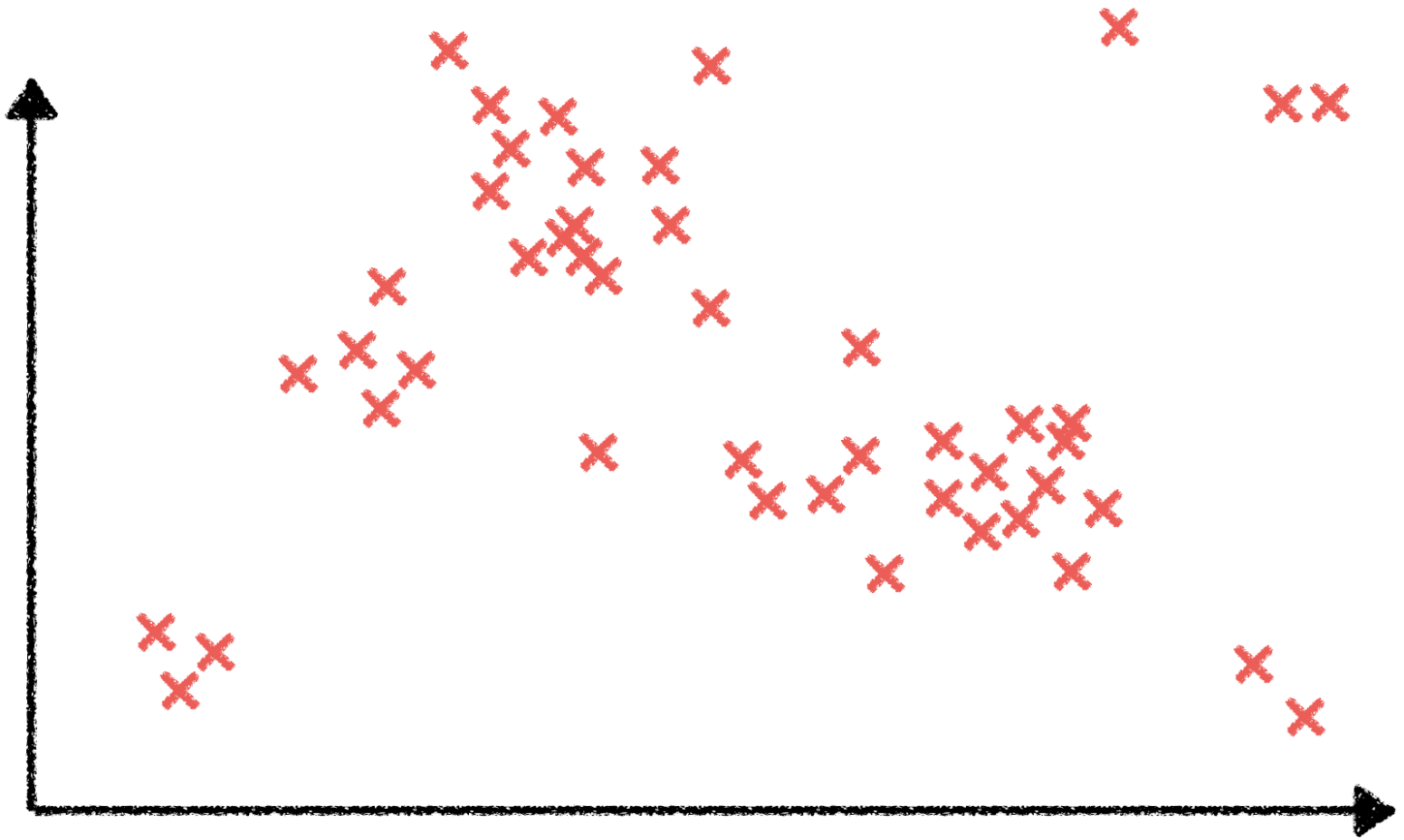
Prototypes and criticisms

Kim, Khanna, **K.** (2016), Examples are not enough, learn to criticize! criticism for Interpretability.

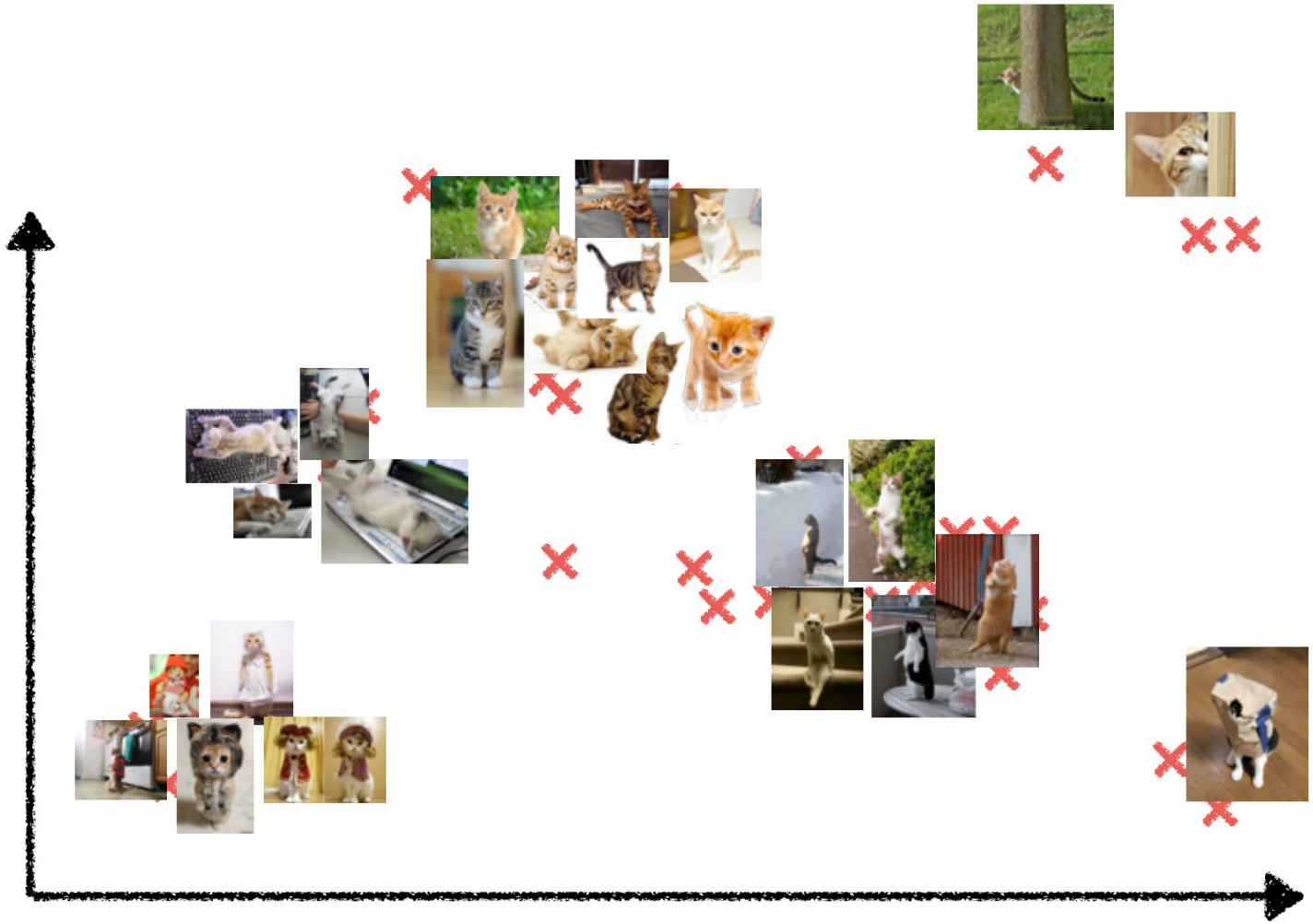
Interpretation via examples

- We often use examples to understand complex scenarios [Cohen 96, Newell 72]
- Example [Klein 1998]:
  - Fire ground commander coordinating his crew, while fighting a fire at a low-rise apartment building
  - Notices billboards on the building's roof
  - Recalls earlier incident where flames burned through the wooden billboard supports, causing them to crash to the street below
  - Orders that spectators be moved farther back to prevent injury from falling billboards

Cohen, M. S., Freeman, J. T., & Wolf, S. (1996). Metarecognition in time-stressed decision making: Recognizing, critiquing, and correcting. *Human Factors*, *38*(2), 206-219.
Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.
Klein, G. Sources of Power: How People Make Decisions. Massachusetts Institute of Technology: 1998.



Which examples are most salient?



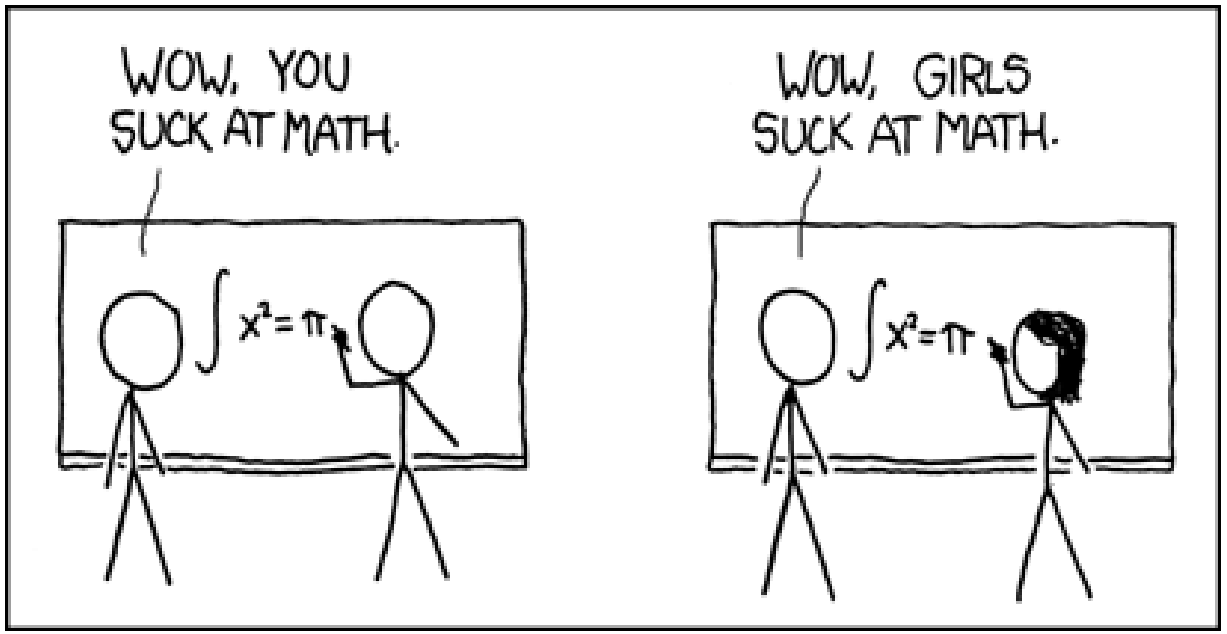
Which examples are most salient?

Related work

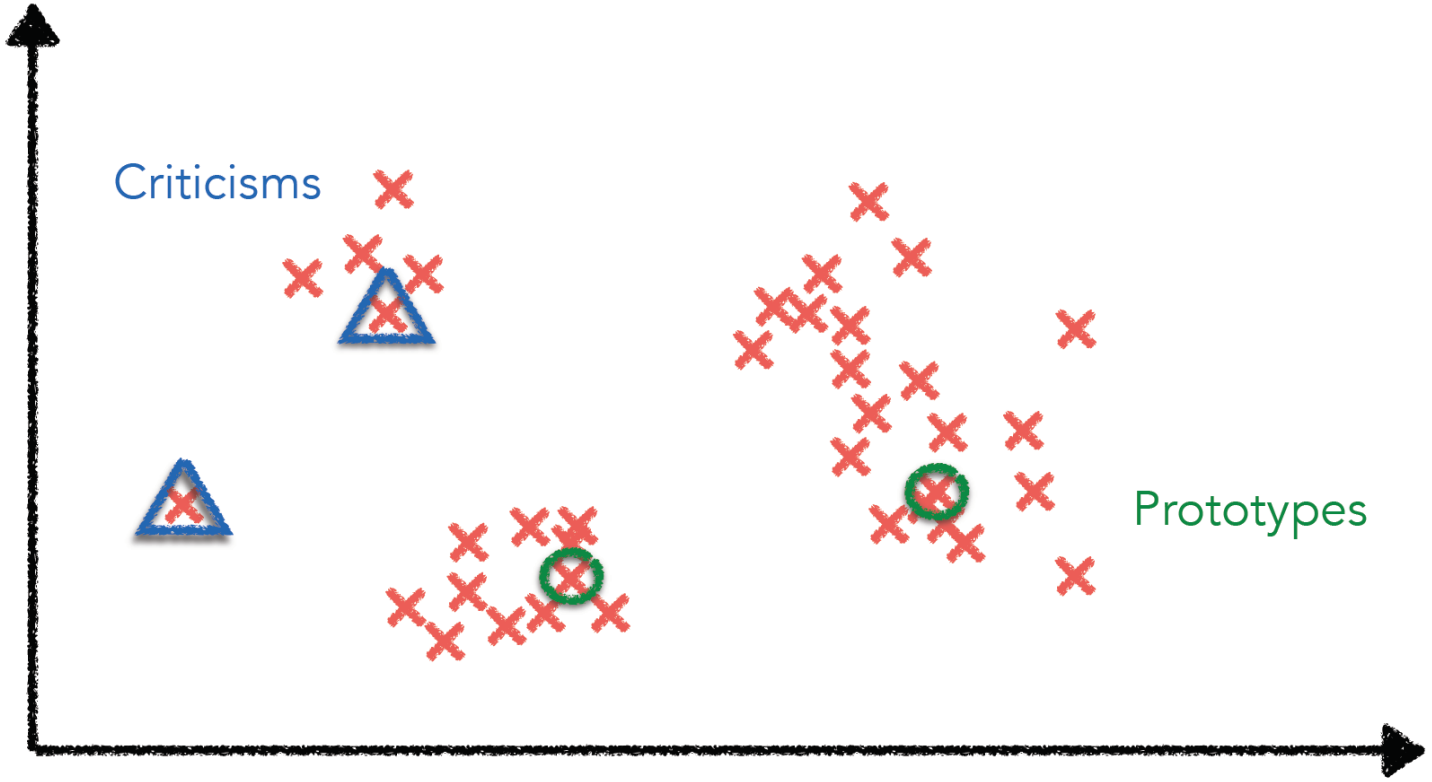
- K-medoids clustering [Kaufman & Rousseeuw, 1987]
- Case-based reasoning [Aamodt & Plaza, 1994]
- Prototype selection for interpretable classification [Bien & Tibshirani, 2011]
- Image summarization [Simon, 2007]
- and many more…

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications, 7(1), 39-59.
Kaufman, L., & Rousseeuw, P. (1987). Clustering by means of medoids. North-Holland.
Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. The Annals of Applied Statistics, 2403-2424.
Simon, I., Snavely, N., & Seitz, S. M. (2007, October). Scene summarization for online image collections. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8). IEEE.

Danger: we have a tendency to over-generalize from examples



https://xkcd.com/385/



Prototypes for explanation, plus *criticism* to minimize over-generalization

P is the observed data empirical distribution

Q is the empirical distribution of the examples

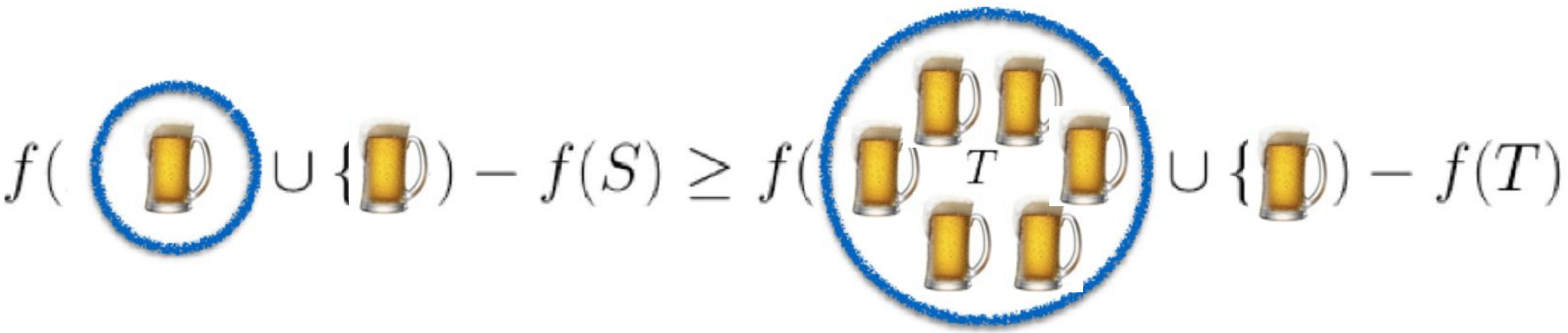
$$\mathrm{MMD}(\mathcal{F}, P, Q) = \sup_{f \in \mathcal{F}} \left(\mathrm{E}_{X \sim P}\left[f(X)\right] - \mathrm{E}_{Y \sim Q}\left[f(Y)\right] \right)$$

F is a kernel function space; depends on the similarity function $k(x_i, x_j)$

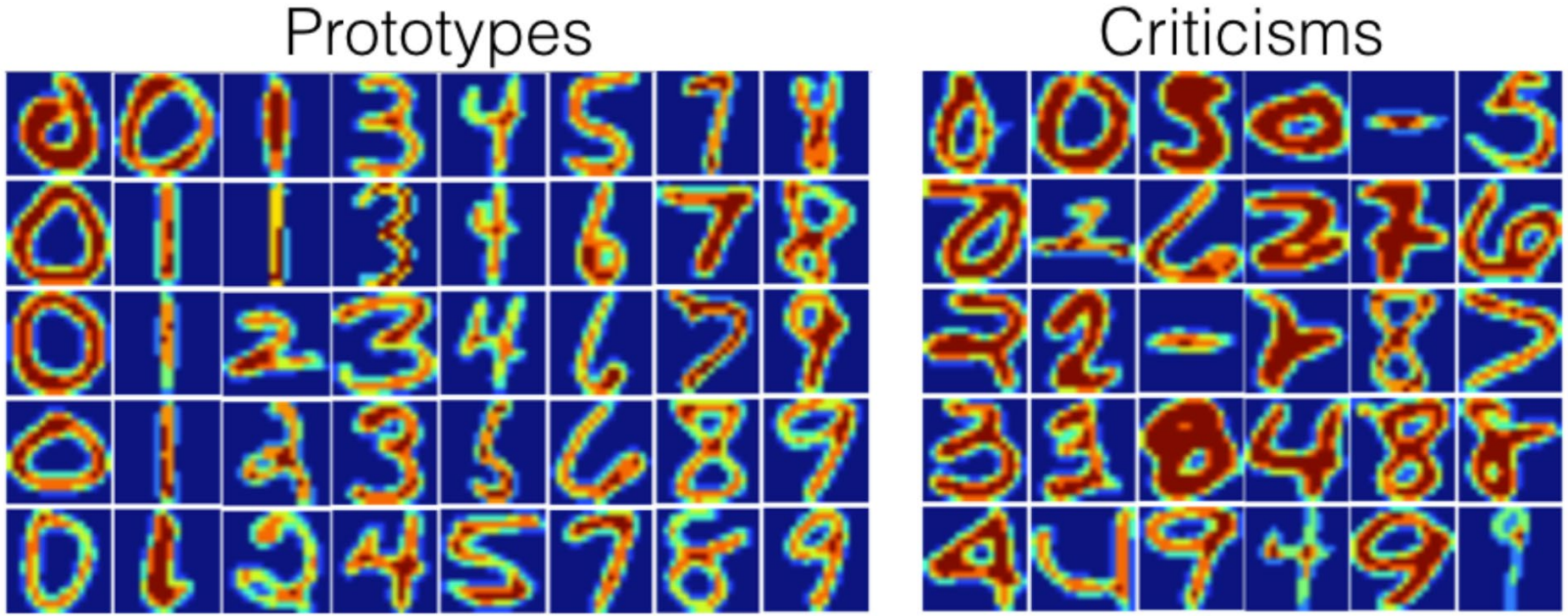
Maximum Mean Discrepancy

SIMILARITY BETWEEN THE DATA DISTRIBUTION AND THE “PROTOTYPE DISTRIBUTION”

Prototype selection problem is submodular

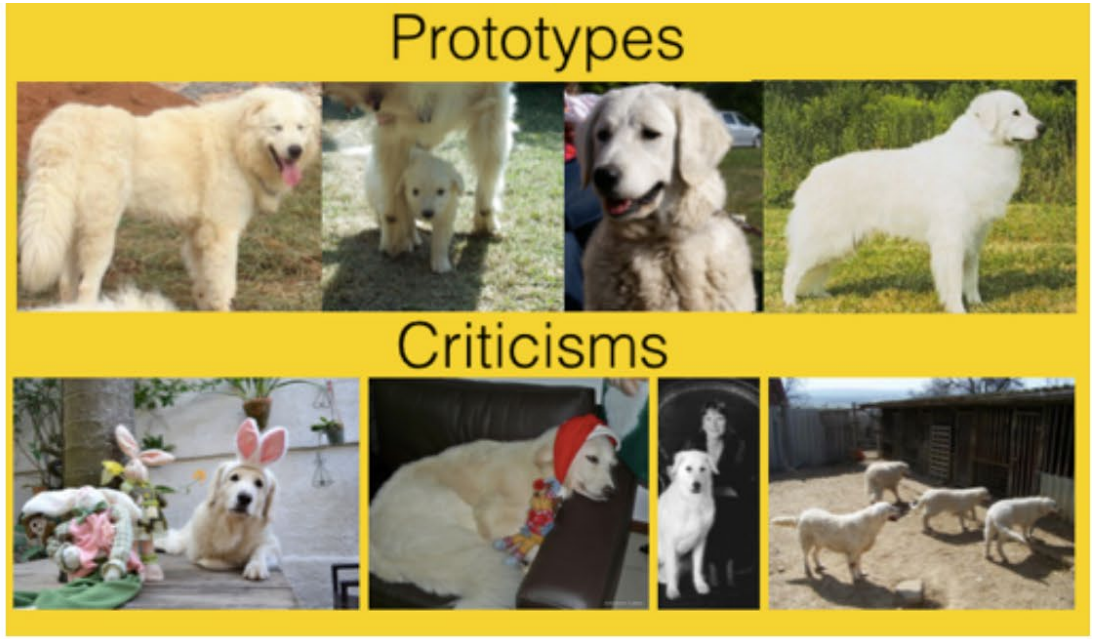


Implies that greedy selection is efficient and effective



USPS digits

Imagenet



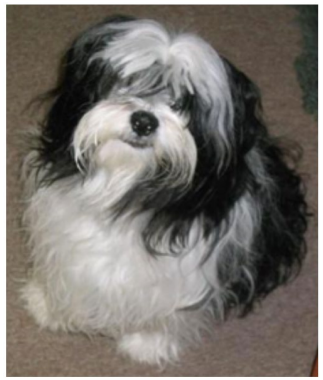
Using image embeddings from [He ’15]

Study with Human Judges

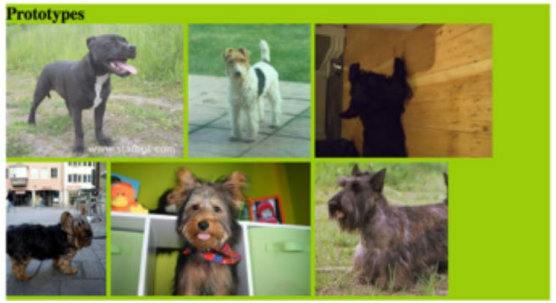
Can a user correctly and efficiently predict the method's results?

Assign a new data point to one of the groups using:

- all images
- prototypes
- prototypes + criticisms
- small set of randomly selected images



a new data point



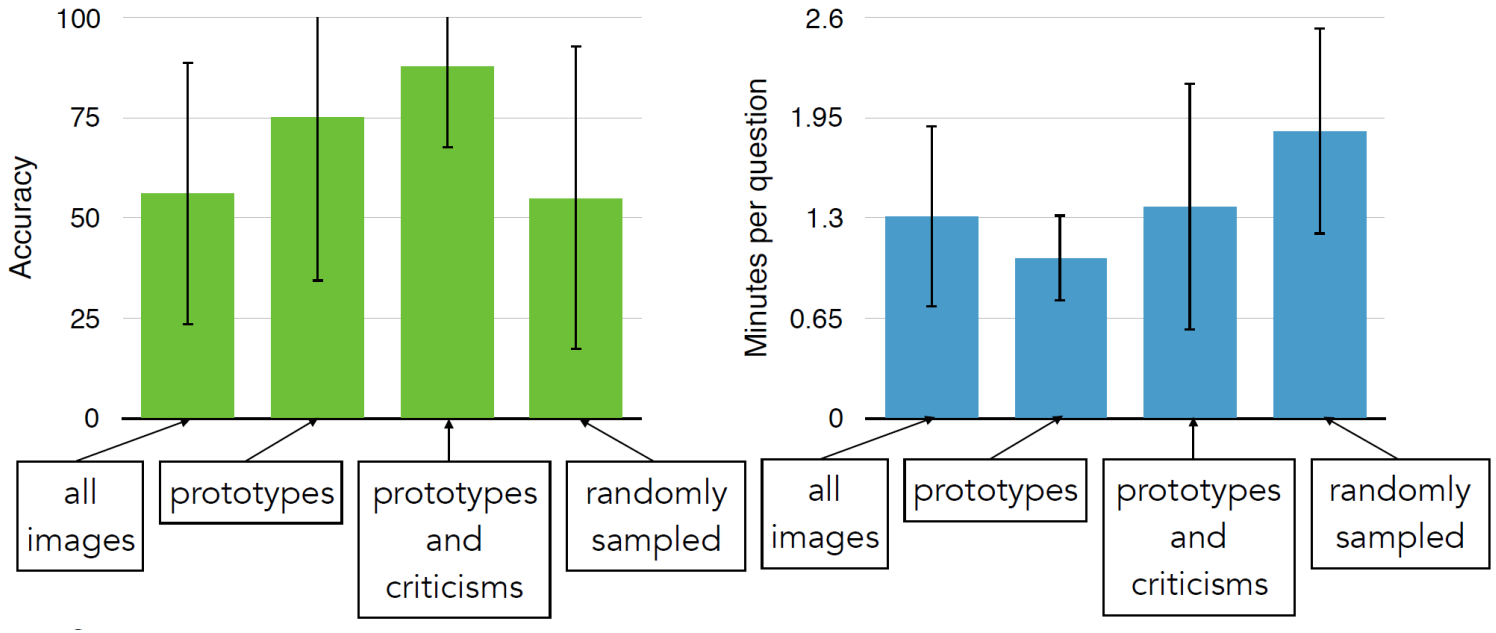
Group 1

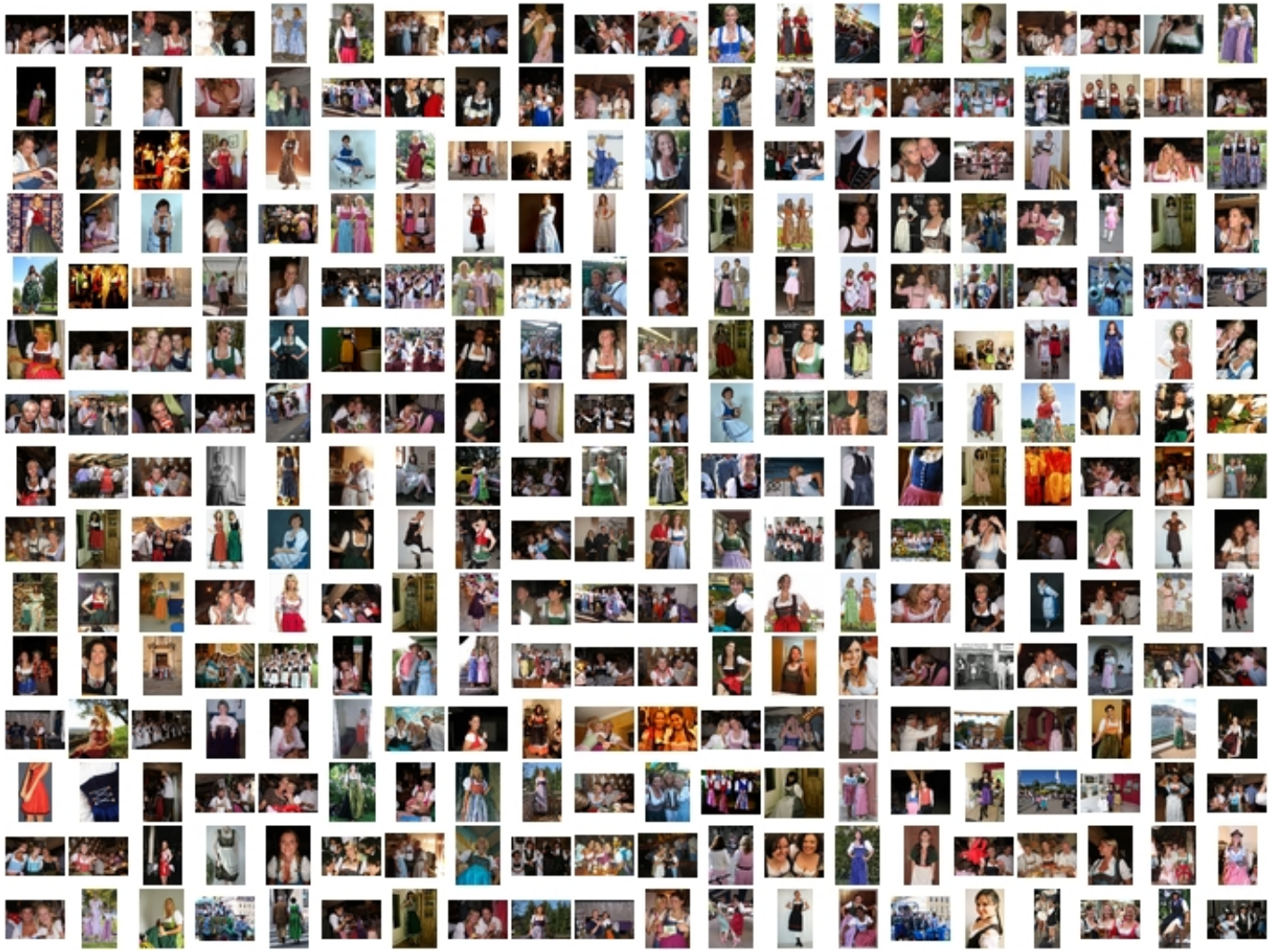
VS

Group 2

Study with Human Judges

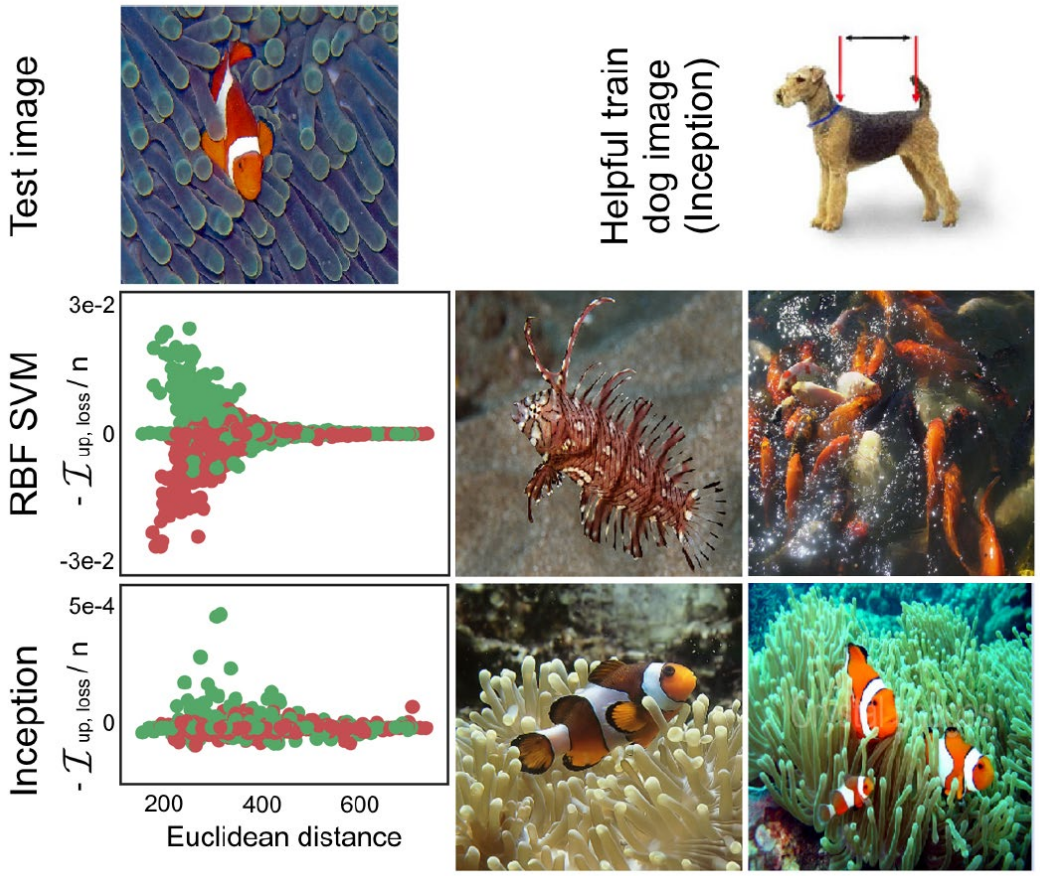
User feedback: “prototype + criticism condition resulted in less confusion from trying to discover hidden patterns in a ton of images, more clues indicating what features are important”





Which examples helped my *trained model* distinguish cats from dogs?

Selecting examples using influence functions



- Efficiently approximation ablation with respect to each sample, i.e., how much does the model change if I add/remove this sample?
- Understanding model behavior; most helpful “fish” training examples for predicting the fish

Koh and Liang (2017), Understanding Black-box Predictions via Influence Functions

Robustness, training set attacks

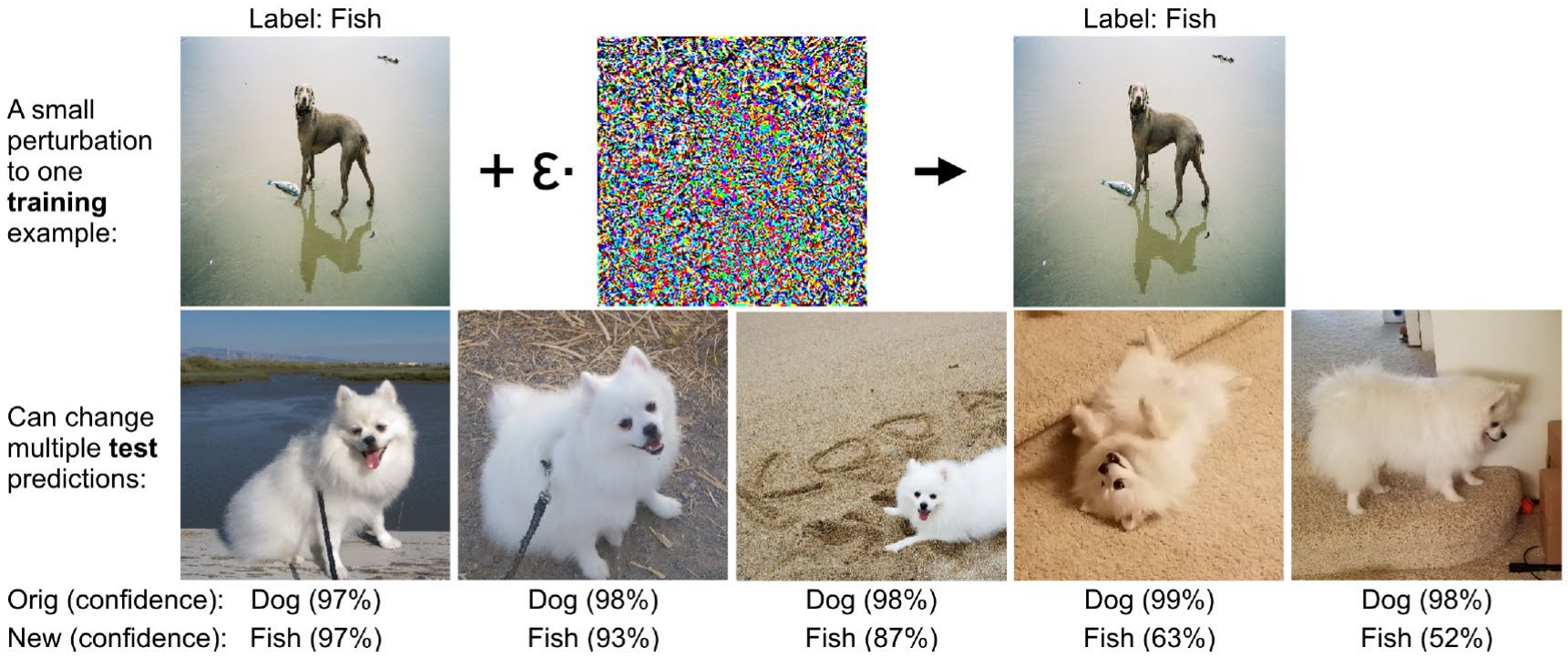
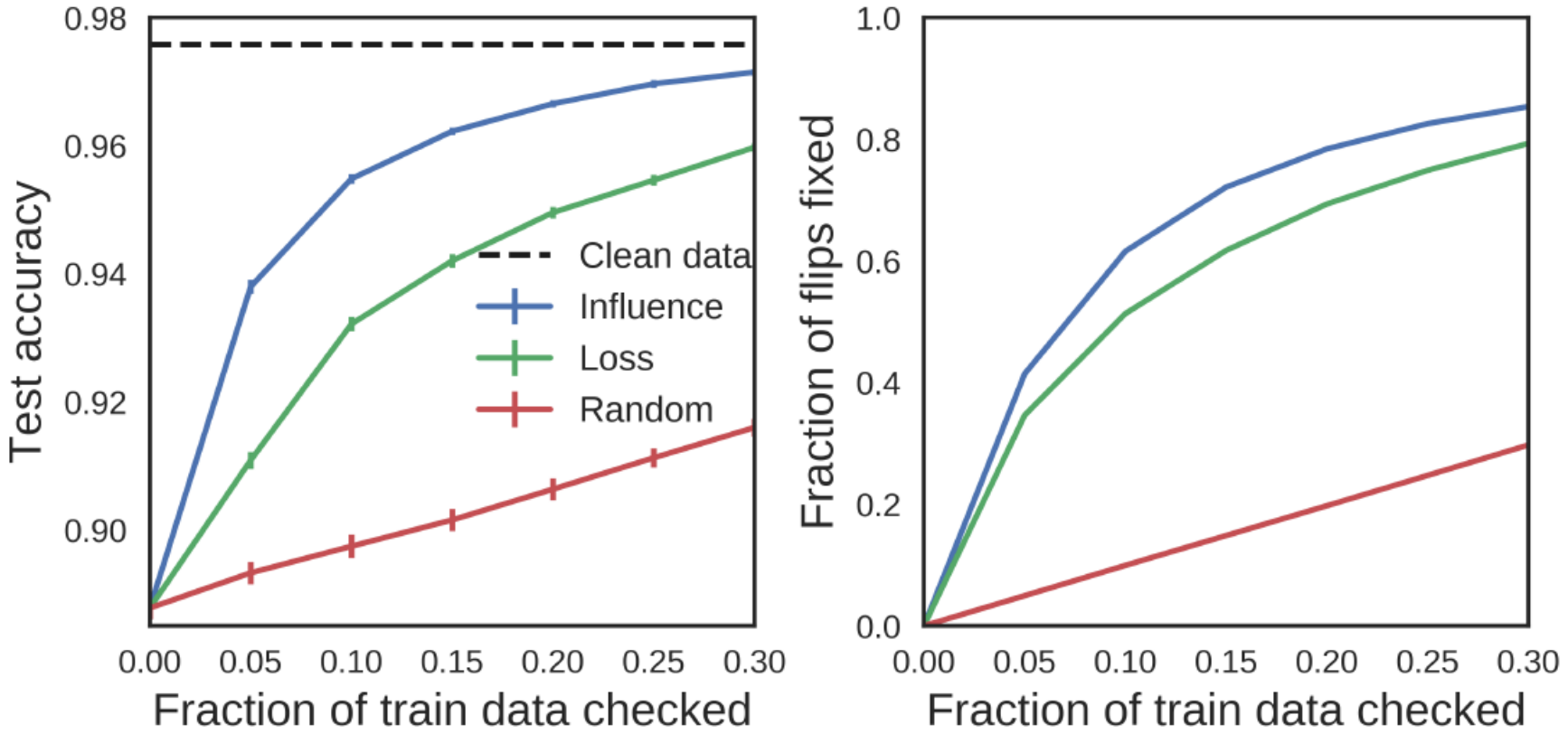


Figure 5. **Training-set attacks.** We targeted a set of 30 test images featuring the first author's dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, we created a visually-imperceptible change to the particular training image (shown on top) that flipped predictions on 16 test images.

Fixing mislabeled examples



Model-dependent similarity

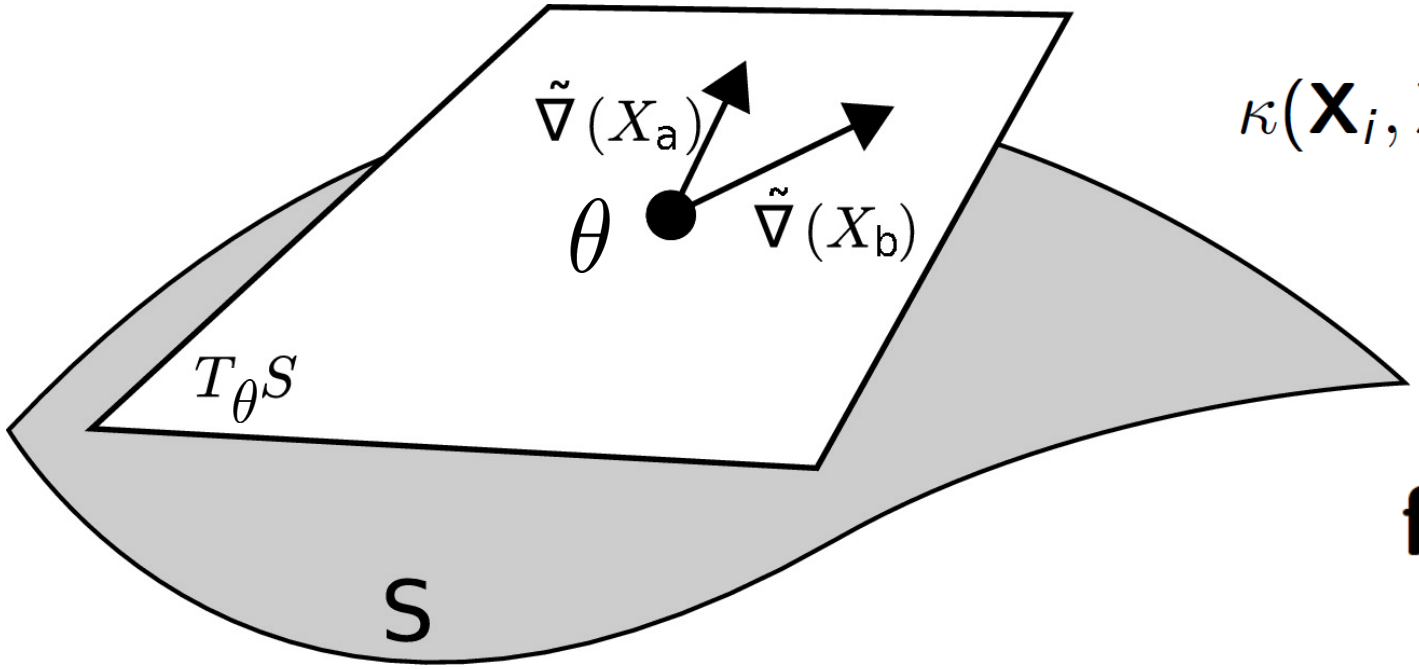
- Represent each data point by “how much it changes the model”
  - Naïve approach: *leave-one-out* re-training
  - Infinitesimal changes to parameter ➔ gradient

- ***Fisher Kernels [Jakkola & Haussler, 1999]:***
  Similarity of two points ➔ similarity between induced model gradients
  - Measures effect of slight perturbation of the data with respect to the two points
  - Equivalent to “influence functions” [Koh & Liang, 2017]

- Kernel similarity is ***model agnostic*** – requires only the function and gradients
- Under MMD selection: prototypes selected wrt. of largest marginal change in model

Jaakkola, T., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. NIPS
Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. ICML

Fisher Kernel

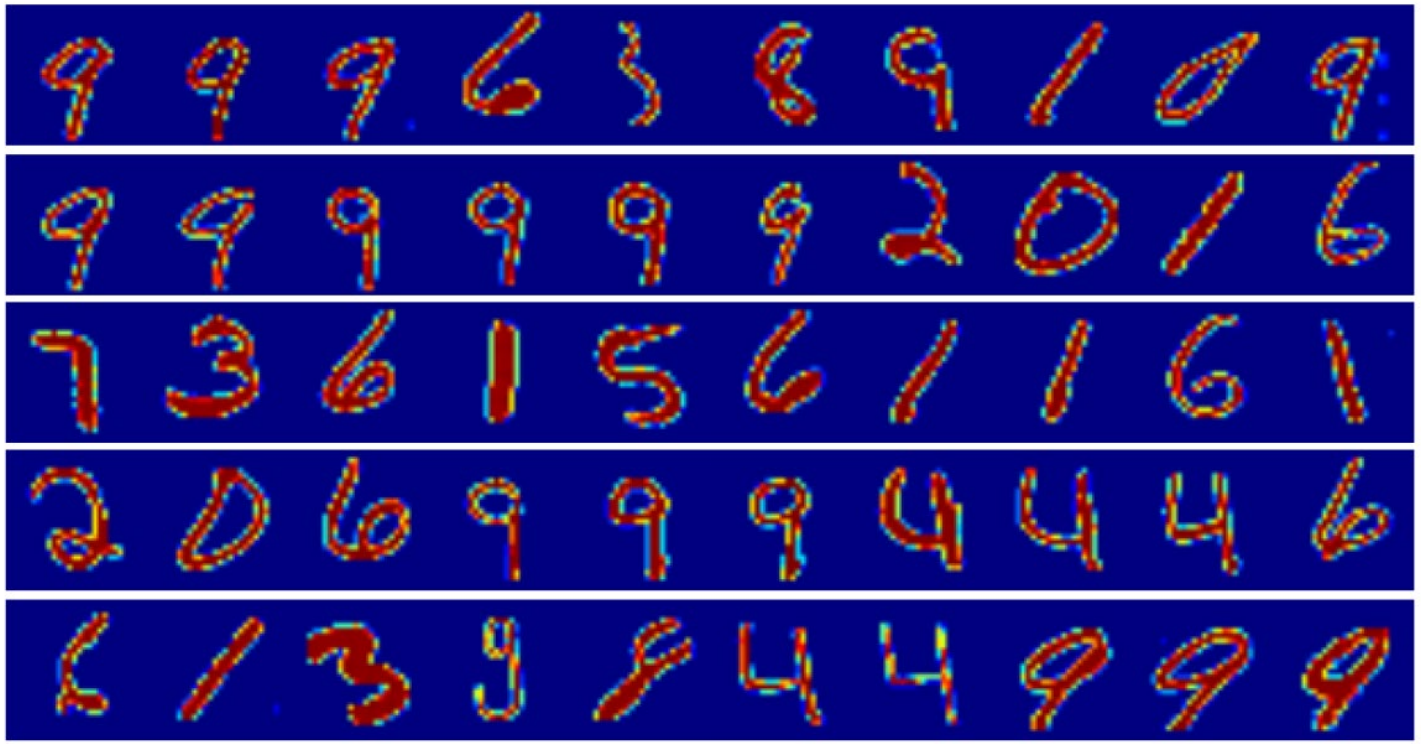


$$\kappa(\mathbf{X}_i, \mathbf{X}_j) := \mathbf{f}_i^\top \mathcal{I}^{-1} \mathbf{f}_j,$$

$$\mathbf{f}_i := \frac{\partial \log p(\mathbf{X}_i|\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}}$$

$$\mathcal{I} := \mathbb{E}_{p(\mathbf{X})}\left[\frac{\partial \log p(\mathbf{X}|\theta)}{\partial \theta}^\top \frac{\partial \log p(\mathbf{X}|\theta)}{\partial \theta}\right]$$

Figure: Sanchez, J., & Redolfi, J. (2015). Exponential family Fisher vector for image classification. *Pattern Recognition Letters*, *59*, 26-32.

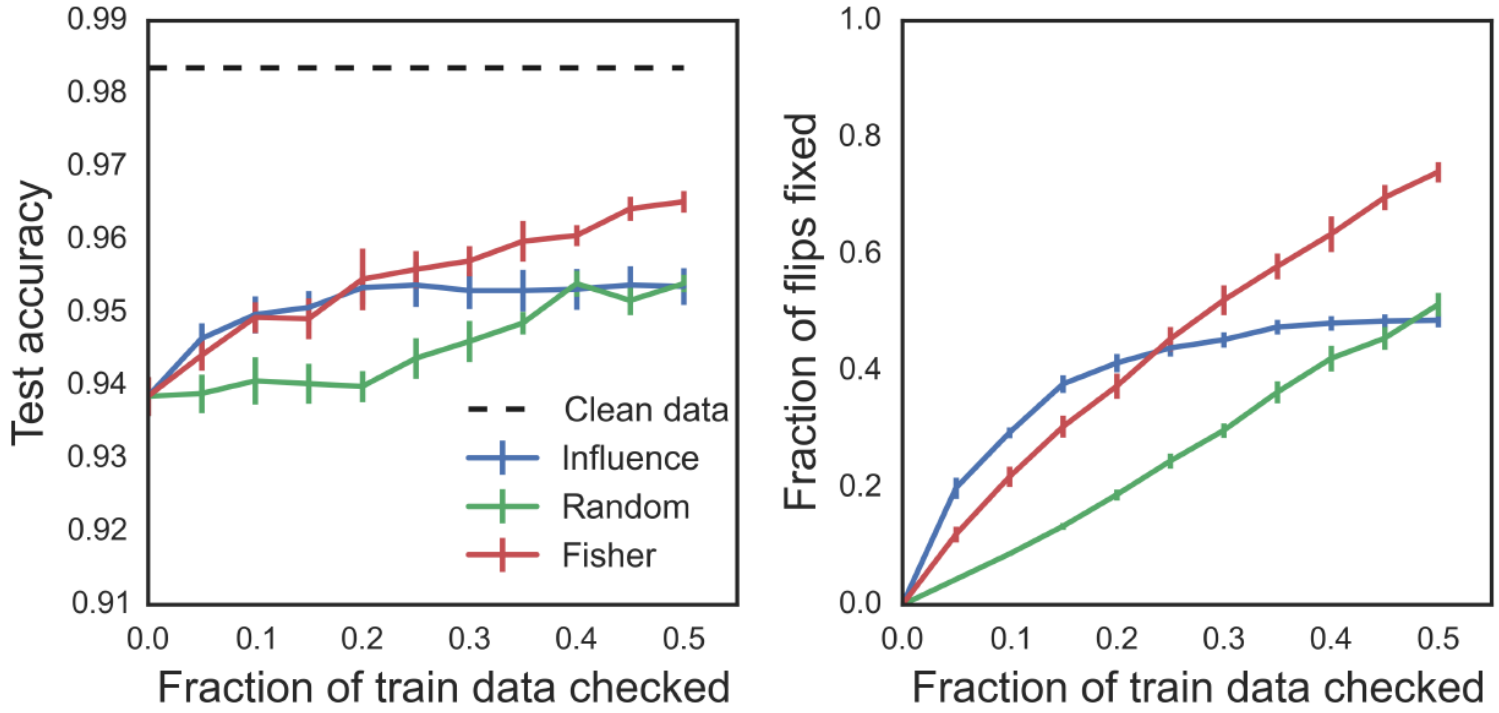


“Influential” samples in neural network classifier

- Trained a 2 layer convolutional neural network to predict digit
- “Which training examples are most important for distinguishing 4s and 9s when building a classifier for all the digits?”
- Prototypes best explain model, i.e., adding/removing these will change model the most

Khanna, Kim, Ghosh, **K.** (2019), Interpreting Black Box Predictions using Fisher Kernels

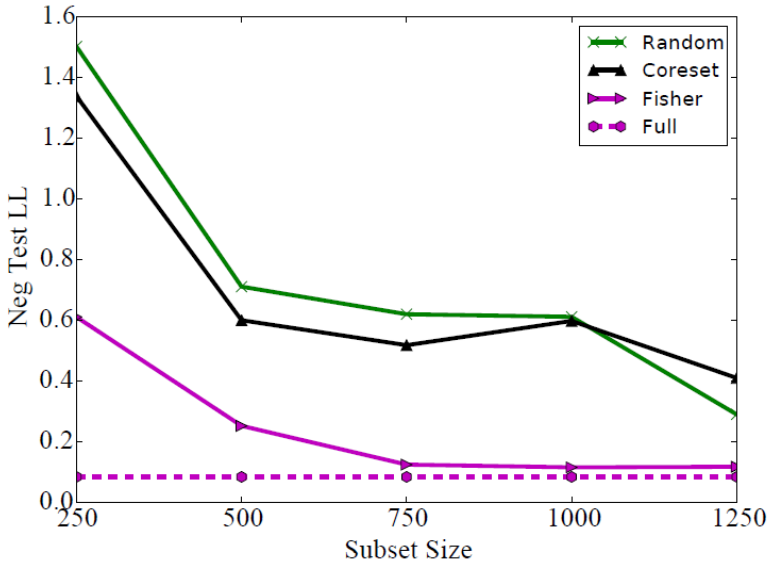
Identify and fix mislabeled data

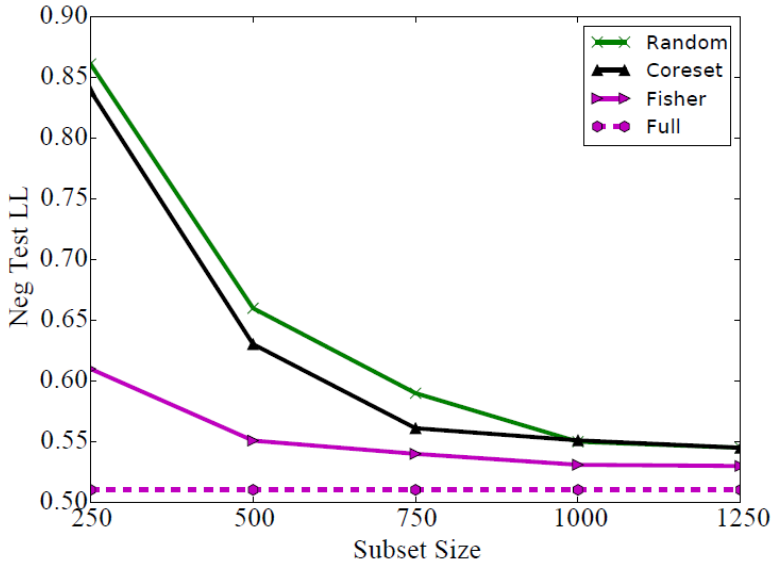


Intuition: wrongly labelled data has large influence on model

Summarization for a logistic regression model

- ***Machine Teaching:*** selected subset replaces full dataset





Benchmark data: Chemreact, Covtype

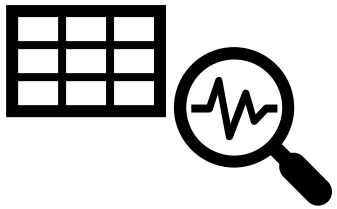
Recourse

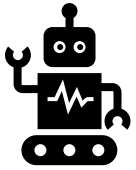
Individual Recourse

Joshi, Kim, **K,** Ghosh. (2019), REVISE. In Prep.

Will the customer default on a credit loan re-payment next month?

Attributes	Original	Under Recourse
Max Bill Amount Over Last 6 Months	930	782.7
Max Payment Amount Over Last 6 Months	40.0	50
Months With High Spending Over Last 6 Months	6.0	0.0
Total Months Overdue	16.0	3.0





Is customer likely to default?

Yes

Recourse

No

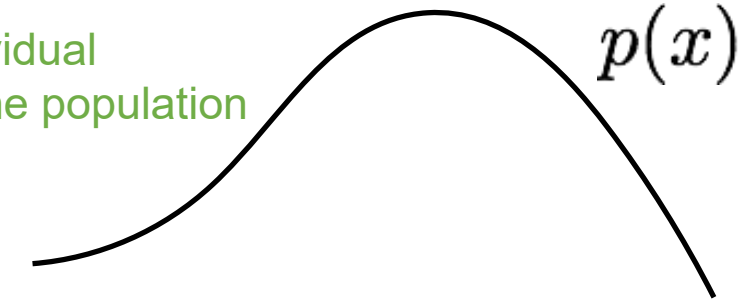
Individual recourse should be “realistic”

- Feasible: don’t propose that individuals change race, gender

Feasibility

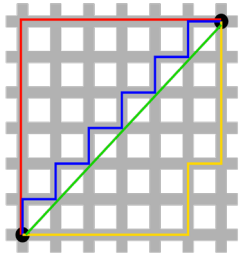
Race
Gender
Age ↑ } **fixed**

Looks like a real individual
,i.e., a sample from the population



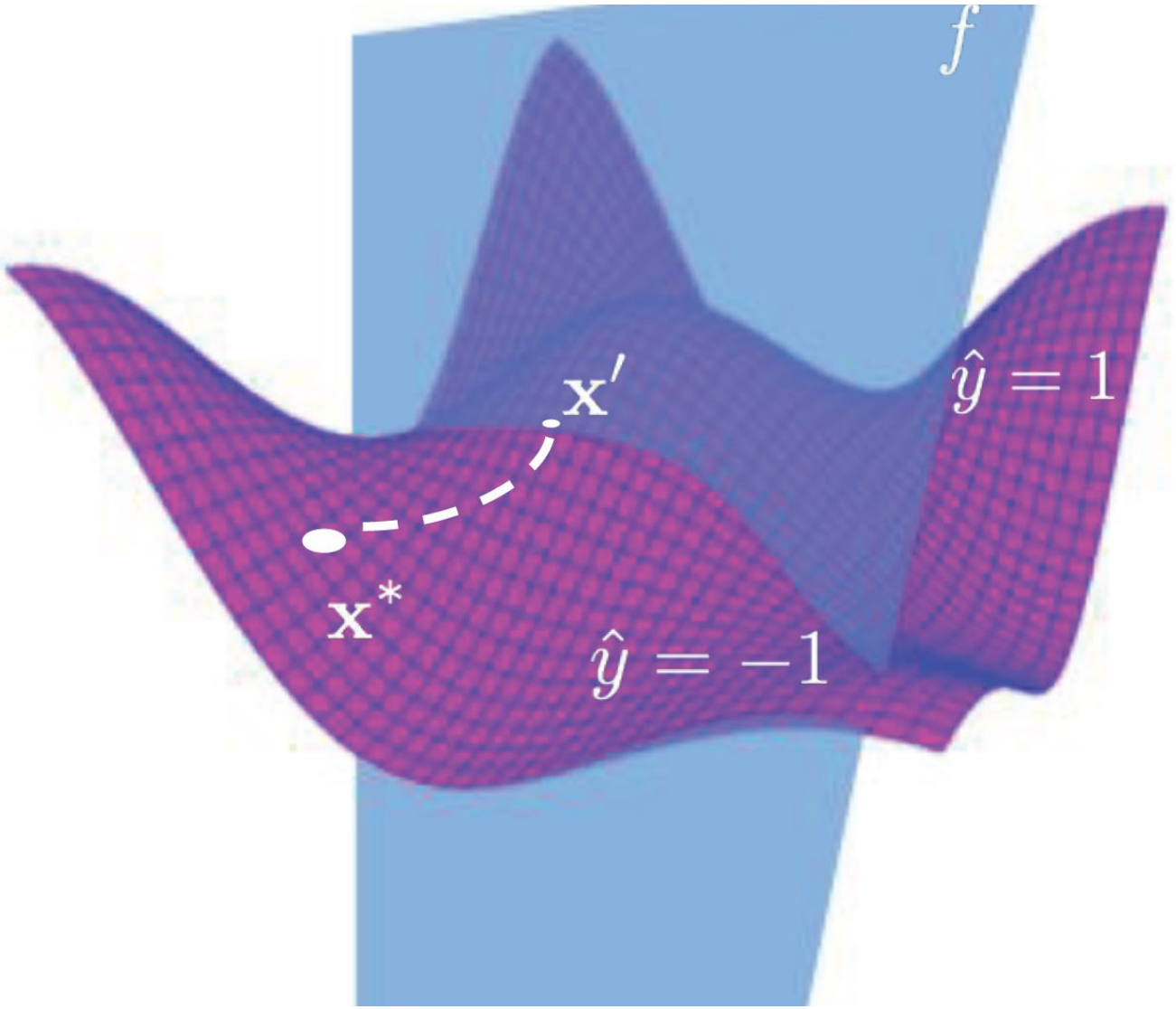
- Seek the smallest set of feasible changes that changes outcome

Small Changes
(under a distance metric)



REVISE

- Generative model of data distribution
- Algorithmic decision (classifier)
- Constrained optimization to identify recourse

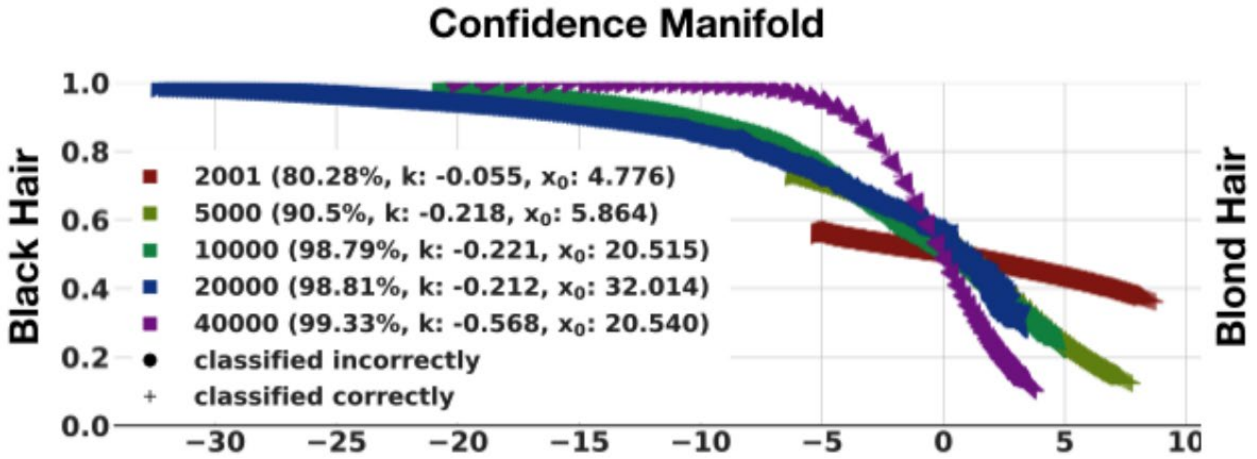


Results – UCI default credit dataset

Attribute	original	REVISE (Linear)	REVISE (MLP)	Ustun et. al. '18 (Linear)
Max Bill Amount Over Last 6 Months	2240.0	3461.2947	1548.9572	-
Max Payment Amount Over Last 6 Months	110.0	100.3251	17.0988	-
Months With High Spending Over Last 6 Months	6.0	0.0547	1.9147	-
Most Recent Bill Amount	2050.0	1768.1843	2059.7888	-
Most Recent Payment Amount	80.0	28.2974	0.0	6010.0
Total Overdue Counts	1.0	1.7552	0.5058	-
Total Months Overdue	12.0	1.05	0.4	-
Others (Marital Status)	0.0	-	-	1

Results – UCI default credit dataset

Attribute	original	REVISE (Linear)	REVISE (MLP)	Ustun et. al. '18 (Linear)
Education Level	University	Graduate	-	-
Max Bill Amount Over Last 6 Months	4000.0	3770.5771	3028.146	-
Max Payment Amount Over Last 6 Months	370.0	241.5032	639.1942	-
Months With Low Spending Over Last 6 Months	0.0	-	0.0745	-
Months With High Spending Over Last 6 Months	6.0	0.0	3.0379	-
Most Recent Bill Amount	3780.0	3122.0967	4995.4946	-
Most Recent Payment Amount	0.0	28.0093	6210.4756	5760.0
Total Overdue Counts	1.0	1.0941	0.7319	-
Total Months Overdue	12.0	1.2939	0.0	-
Others (Marital Status)	0	-	-	1



- Results Celeb-A dataset Counterfactuals of convolutional DNN

Miscellaneous

- Mostly covered supervised ML. Some of these ideas apply to unsupervised learning, RL, …
- Stability of “interpretable” models.
  - Models can be sensitive to correlations. Apply causal interpretation with care!
  - Decision trees and be sensitive to optimization
- Post-hoc interpretability can be unstable / unreliable
  - Saliency maps can be unstable wrt. Data (Adebayo et. al.,2018)
  - LIME can be unstable wrt. Optimization (Adhakari, **K.**, 2019)

Evaluating interpretability

- Initial desiderata (predictability, decomposability, trust) are challenging to evaluate directly
- Compare model properties e.g. sparsity:
  - Pro: are easy to evaluate, objective measure
  - Con: may not reflect useful metrics
- Human evaluation:
  - Pro: more likely to reflect real world utility
  - Con: from survey design to experiment selection, more difficult/costly to implement

Quiz

- Why interpretability is important?
  - Often a proxy solution for an underspecified ML problem
  - For ml algorithm designers: debugging
  - For expert users: trust, safety
  - For end users: fairness, individual recourse
- What are some “interpretable” models?
  - (small) linear models, (small) decision trees, monotonic models
- What are some techniques for post-hoc interpretability?
  - Saliency maps, LIME, TCAV, critics, influential examples

Summary

1

Who cares about interpretability? (should you care?)

2

What is interpretable machine learning?

3

Why do we care about interpretability?

4

How does one build interpretable models?

Thank you

sanmi@Illinois.edu

@sanmikoyejo