

ML Interpretability and Intrinsic Models

Apr 15, 2020

Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning
Stanford University

Recap

- Major Fairness Criteria

- Fairness Through Unawareness

- Sensitive feature A are being excluded when training ML models

- Demographic Parity

- Probabilities of distributing the favorable outcome across groups are the same

$$P(\hat{Y} = 1 | \underline{A} = 1) = P(\hat{Y} = 1 | \underline{A} = 0)$$

- Equal Opportunity

- Probabilities of distributing the favorable outcome to the qualified members across groups are the same

$$P(\hat{Y} = 1 | \underline{A} = 0, \underline{Y} = 1) = P(\hat{Y} = 1 | \underline{A} = 1, \underline{Y} = 1)$$

- Equal Odds

- Probabilities of distributing the favorable outcome to both qualified and unqualified members across groups are the same

$$P(\hat{Y} = 1 | \underline{A} = 0, \underline{Y}) = P(\hat{Y} = 1 | \underline{A} = 1, \underline{Y})$$

Recap

- Fair Representation Learning
 - Prejudice Removing Regularizer

$$-\mathcal{L}(\mathcal{D}; \boldsymbol{\Theta}) + \eta R(\mathcal{D}, \boldsymbol{\Theta}) + \frac{\lambda}{2} \|\boldsymbol{\Theta}\|_2^2$$

Loss of the Model Fairness Regularizer L2 Regularizer

- Prejudice Removing Regularizer Minimizes Mutual Information

$$PI = \sum_{Y,S} \hat{Pr}[Y, S] \ln \frac{\hat{Pr}[Y, S]}{\hat{Pr}[S] \hat{Pr}[Y]}$$

- $PI = 0 \Rightarrow \hat{Y} \perp\!\!\!\perp S \Rightarrow \text{Demographic Parity}$

Outline

- Fair Representation Learning
- ML Interpretability
- Intrinsically Interpretable Models
 - Simple interpretable models
 - Intrinsically interpretable techniques for deep learning
- Interpretability Concepts
 - Intrinsic and post hoc methods
 - model-specific and model-agnostic methods
 - Local and global interpretable methods
 - Interpretability and performance trade-offs

Fair Representation Learning

- How Do We Test the Fairness of Deep Representation Z?
 - Adversarial Learning

I want to find the best representation for my task.



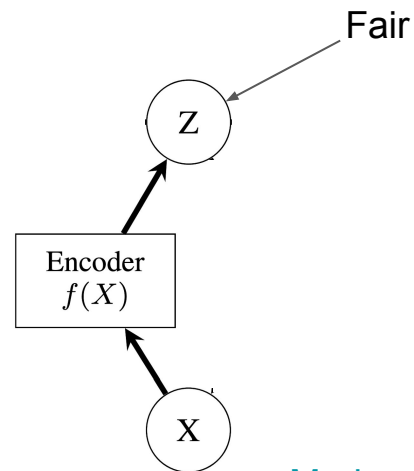
Adversarial Learning

I want to find the worst representation that can reconstruct A



Fair Representations

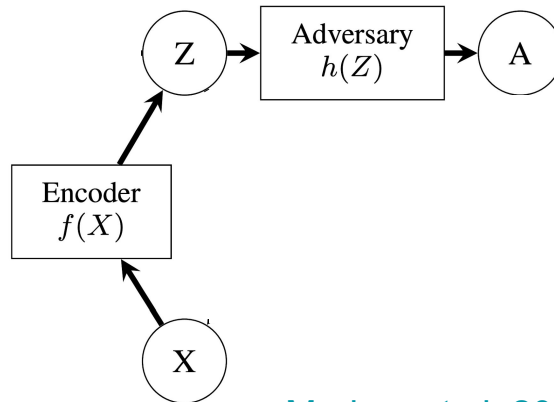
- How Do We Make a Deep Representation Z Fair?



[Madras et al, 2018](#)

Fair Representations

- How Do We Make a Deep Representation Z Fair?
 - $Z = f(X)$
 - Test and see if a good amount of A can be reconstructed from Z
 - Compare A with $h(z)$

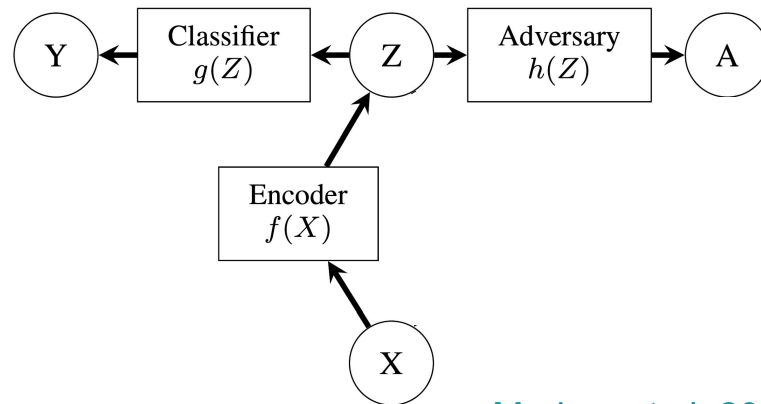


[Madras et al, 2018](#)

Fair Representations

- How Do We Make a Deep Representation Z Fair?
 - $Z = f(X)$
 - Test and see if a good amount of A can be reconstructed from Z
 - Compare A with $h(z)$

- Properties of Deep Representations
 - Achieve good performance for downstream task that generates $y=g(z)$

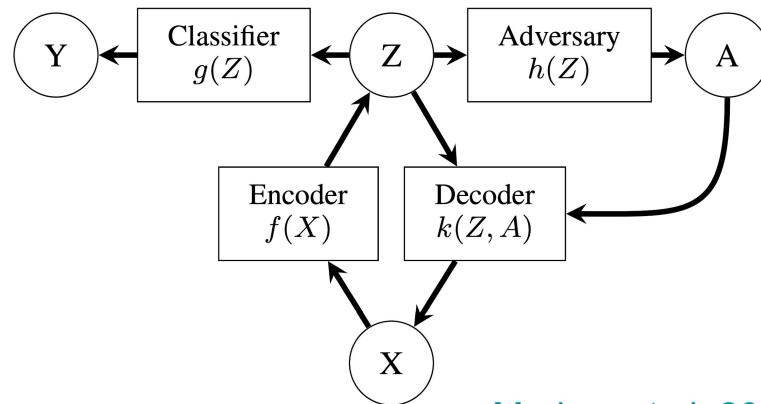


[Madras et al, 2018](#)

Fair Representations

- How Do We Make a Deep Representation Z Fair?
 - $Z = f(X)$
 - Test and see if a good amount of A can be reconstructed from Z
 - Compare A with $h(z)$

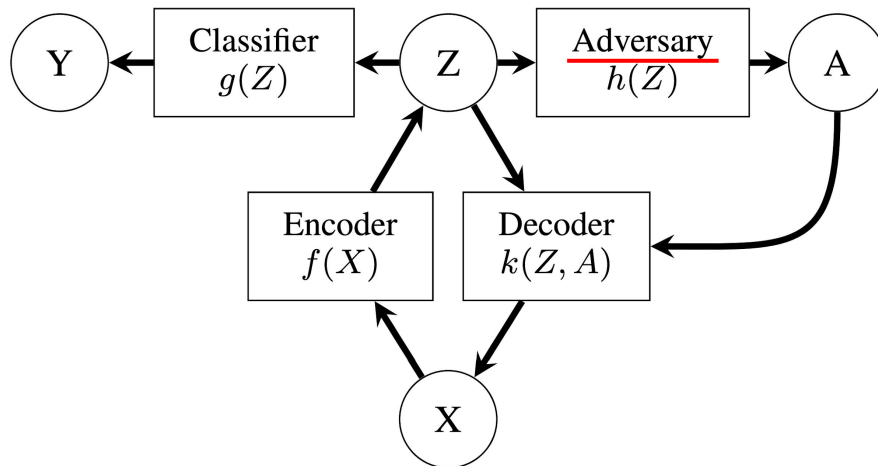
- Properties of Deep Representations
 - Achieve good performance for downstream task that generates $y=g(z)$
 - Has the Ability to Reconstruct $X = k(Z, A)$



[Madras et al, 2018](#)

Fairness Through Adversarial Learning

- Adversarial Learning
 - Models are trained using objectives that compete with each other



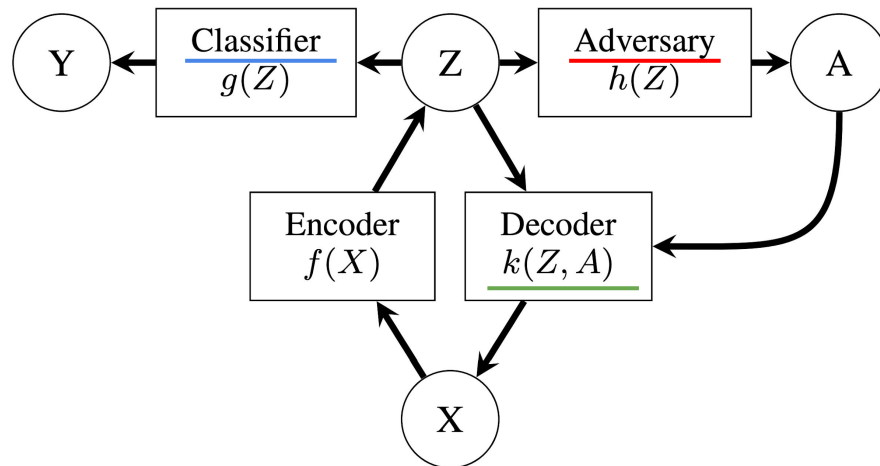
$$\underset{f, g, k}{\text{minimize}} \underset{h}{\text{maximize}} \mathbb{E}_{X, Y, A} [L(f, g, h, k)]$$

[Madras et al, 2018](#)

Fairness Through Adversarial Learning

- Adversarial Learning

$$L(f, g, h, k) = \alpha \underline{L_C(g(f(X, A)), Y)} + \beta \underline{L_{Dec}(k(f(X, A), A), X)} + \gamma \underline{L_{Adv}(h(f(X, A)), A)}$$



$$\underset{f, g, k}{\text{minimize}} \underset{h}{\text{maximize}} \mathbb{E}_{X, Y, A} [L(f, g, h, k)]$$

[Madras et al, 2018](#)

Loss for Learning Fair Representations

- Adversarial Loss for Demographic Parity with Group $\mathcal{D}_0, \mathcal{D}_1$

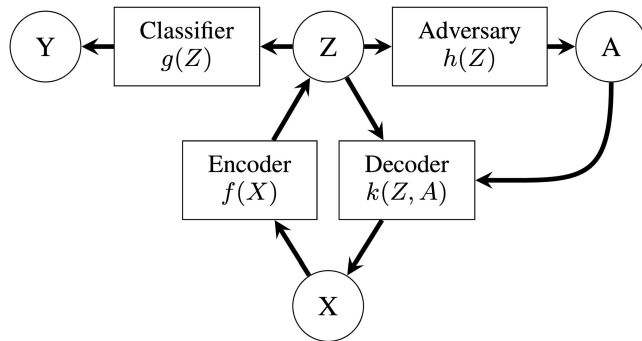
$$L_{Adv}^{DP}(h) = 1 - \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a) \in \mathcal{D}_i} |h(f(x,a)) - a|$$

Demographic Parity: $P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$

- Adversarial Loss for Equality of Odds with Group $\mathcal{D}_i^j = \{(x, y, a) \in \mathcal{D} | a = i, y = j\}$

$$L_{Adv}^{EO}(h) = 2 - \sum_{(i,j) \in \{0,1\}^2} \frac{1}{|\mathcal{D}_i^j|} \sum_{(x,a) \in \mathcal{D}_i^j} |h(f(x,a)) - a|$$

Equality of Odds: $P(\hat{Y} = 1|A = 0, Y) = P(\hat{Y} = 1|A = 1, Y)$



[Madras et al, 2018](#)

Discrimination Measures for Representations

$$\mathcal{Z}_1 = p(Z|A = 1) \quad \mathcal{Z}_0 = p(Z|A = 0) \quad \mathcal{Z}_a^y = p(Z|A = a, Y = y)$$

- Demographic Parity

$$\Delta_{DP}(g) \triangleq d_g(\mathcal{Z}_0, \mathcal{Z}_1) = |\mathbb{E}_{\mathcal{Z}_0}[g] - \mathbb{E}_{\mathcal{Z}_1}[g]|$$

$$\text{Demographic Parity: } P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$$

- Equality of Odds

$$\Delta_{EO}(g) \triangleq |\mathbb{E}_{\mathcal{Z}_0^0}[g] - \mathbb{E}_{\mathcal{Z}_1^0}[g]| + |\mathbb{E}_{\mathcal{Z}_0^1}[1 - g] - \mathbb{E}_{\mathcal{Z}_1^1}[1 - g]|$$

$$\text{Equality of Odds: } P(\hat{Y} = 1|A = 0, Y) = P(\hat{Y} = 1|A = 1, Y)$$

- Equality of Opportunities

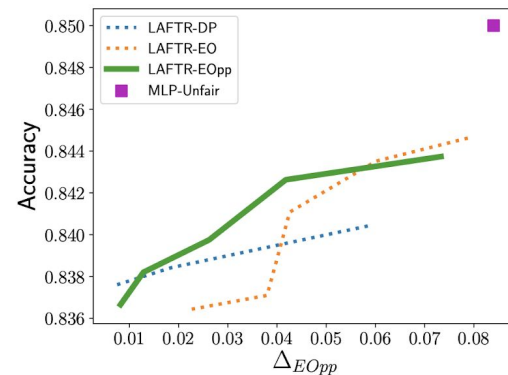
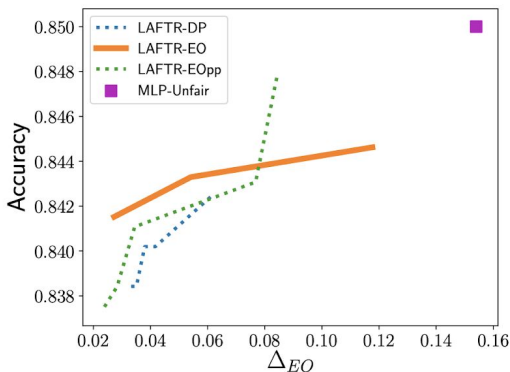
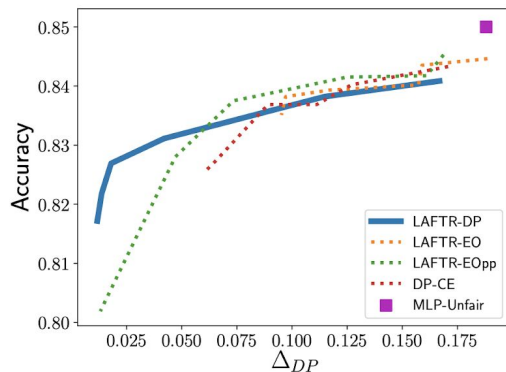
$$\Delta_{EOpp}(g) \triangleq |\mathbb{E}_{\mathcal{Z}_0^0}[g] - \mathbb{E}_{\mathcal{Z}_1^0}[g]|$$

$$\text{Equality of Opportunity: } P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

Accuracy and Fairness on Adult Income Dataset

- Results Generated By Varying γ .

$$L(f, g, h, k) = \alpha L_C(g(f(X, A)), Y) + \beta L_{Dec}(k(f(X, A), A), X) + \gamma L_{Adv}(h(f(X, A)), A)$$

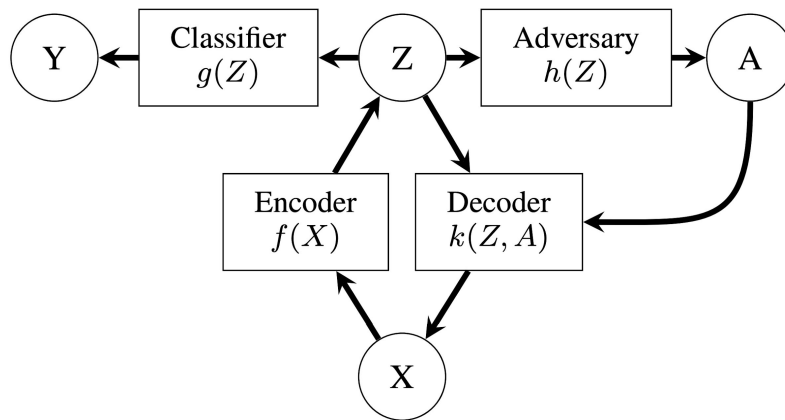


DP-CE - Cross Entropy Adversarial Objective ([Edwards et al., 2016](#))

[Madras et al., 2018](#)

Transferring Fair Representations

- If the Representations Are Fair, All Predictors Should Be Fair!
 - Train f and g based on domain 1 with feature space X
 - Fix f , and train g' on domain 2 with the same feature space X
 - $y=g'(f(x))$ should be a fairness predictor



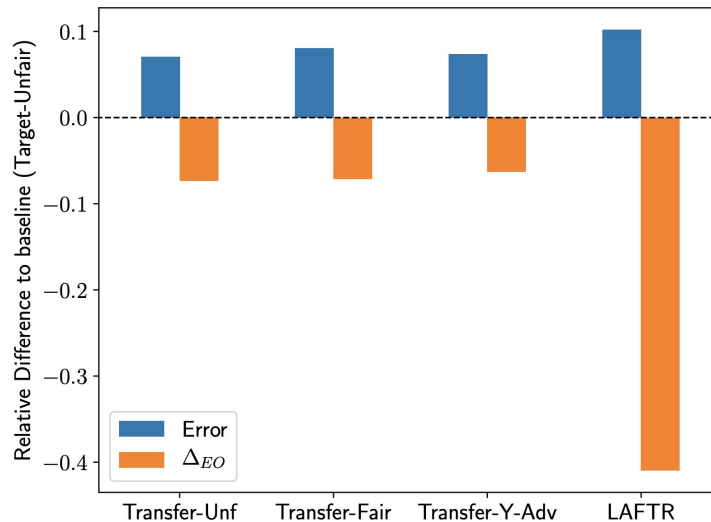
Transfer Fair Representations

- Heritage Health Dataset
 - Comprises insurance claims and physician records
 - Task 1 - Predict Charlson index (prediction of 10 year survival of patients) trained using equalized odds adversarial objective
 - Task 2 - Same input, task becomes predicting a patient's insurance claim corresponding to a specific medical condition

Transfer- unf - MLP with no fairness constraints

Transfer- fair - MLP with fairness constraints in [Bechavod et al, 2017](#)

Transfer - Y - Adv baseline in [Zhang et al, 2018](#)



[Madras et al, 2018](#)

Discussions

- What Are the Pros and Cons of Prejudice Removing Regularizer and Adversarial Learning for Fairness?

Comparisons: Regularization and Adversarial Learning

	Prejudice Removing Regularizer	Adversarial Learning
Pros	Minimal modifications to training procedure	Transferable representations
		Can be applied to many different fairness criteria
Cons	Can only be applied to Demographic Parity	Adversarial loss can be difficult to train

Next Fairness Lectures

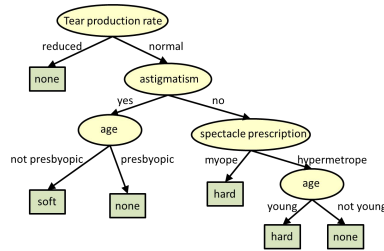
- May 6 Fairness Through Input Manipulations
- May 8 Fair NLP
- May 13 Fairness for Vision Representations

Outline

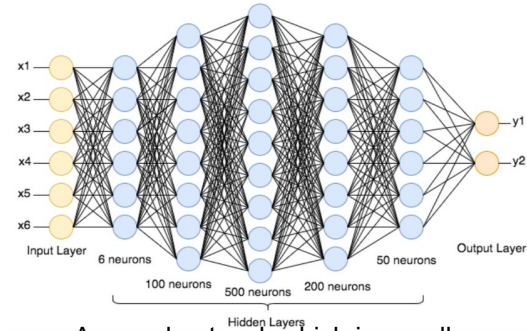
- Fair Representation Learning
- ML Interpretability
- Intrinsically Interpretable Models
 - Simple interpretable models
 - Intrinsically interpretable techniques for deep learning
- Interpretability Concepts
 - Intrinsic and post hoc methods
 - model-specific and model-agnostic methods
 - Local and global interpretable methods
 - Interpretability and performance trade-offs

Machine Learning Interpretability

- ML interpretability allows one to examine model's basis in its decision making process.

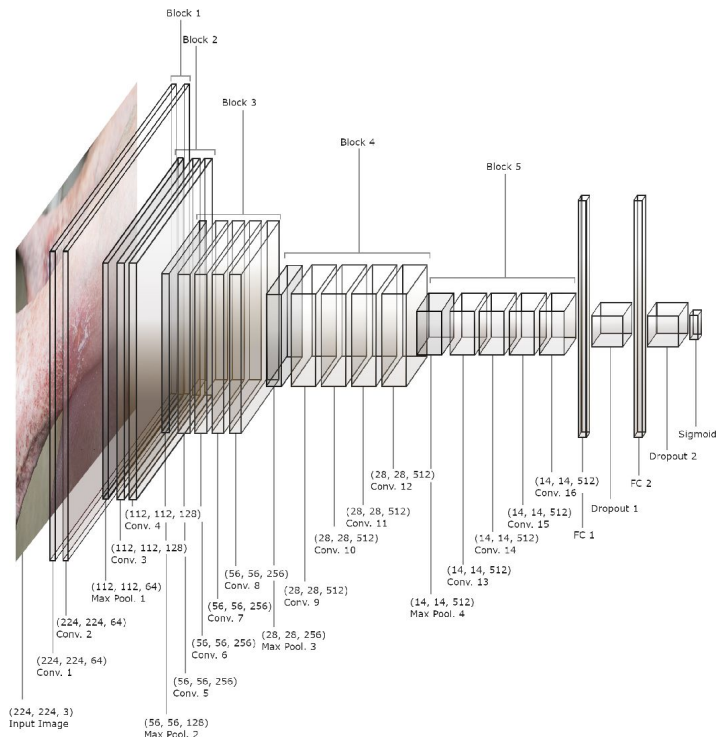


An interpretable tree model to find out the kind of contact lens a person may wear



A neural network which is usually considered a black-box model.

VGG19 Architecture

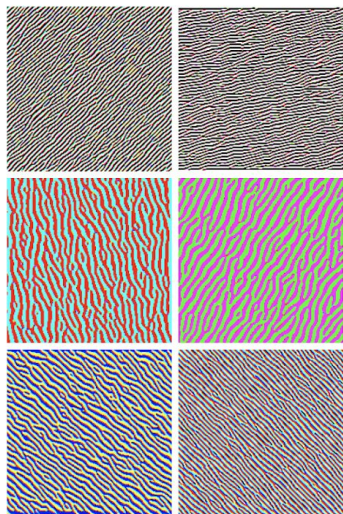


46 layers

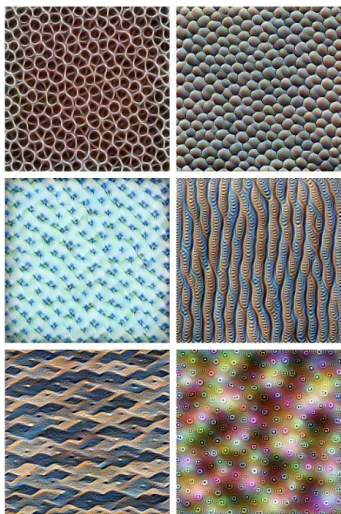
143,667,240 parameters

model size: 575 MB

Visualizations of GoogLeNet



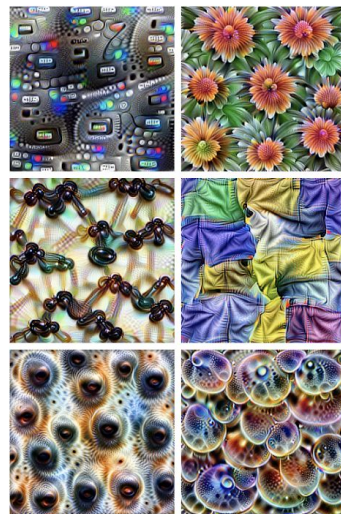
Edges (layer conv2d0)



Textures (layer mixed3a)



Patterns (layer mixed4a)



Parts (layers mixed4b & mixed4c)

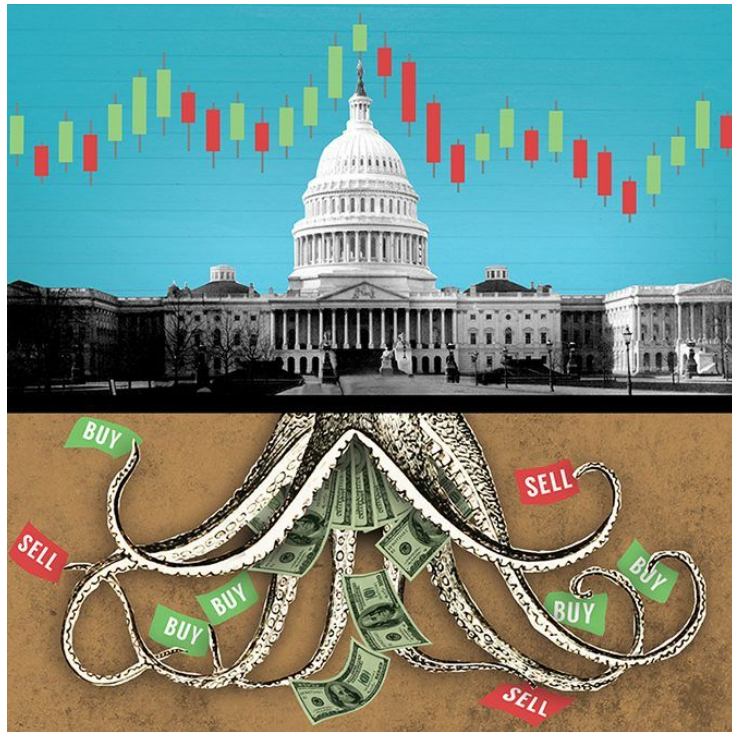


Objects (layers mixed4d & mixed4e)

Reasons for ML Interpretability

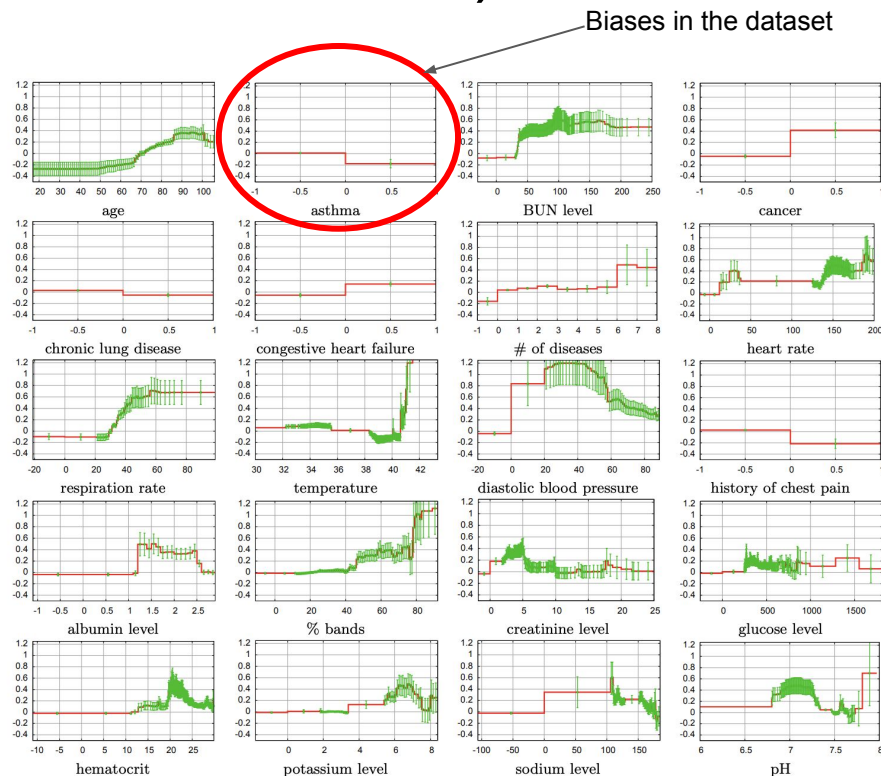
- Our society has been shifted to rely on AI more than ever
 - autonomous vehicles
 - security
 - finance
 - many others
- Who will benefit from ML Interpretability?
 - End Users: enhance trust, understand the consequences of the decisions, e.g., privacy, fairness.
 - Regulatory Agencies: compliance, audits, and accountability.
 - Model Designers: diagnose model performance

Regulating AI Models for Trading



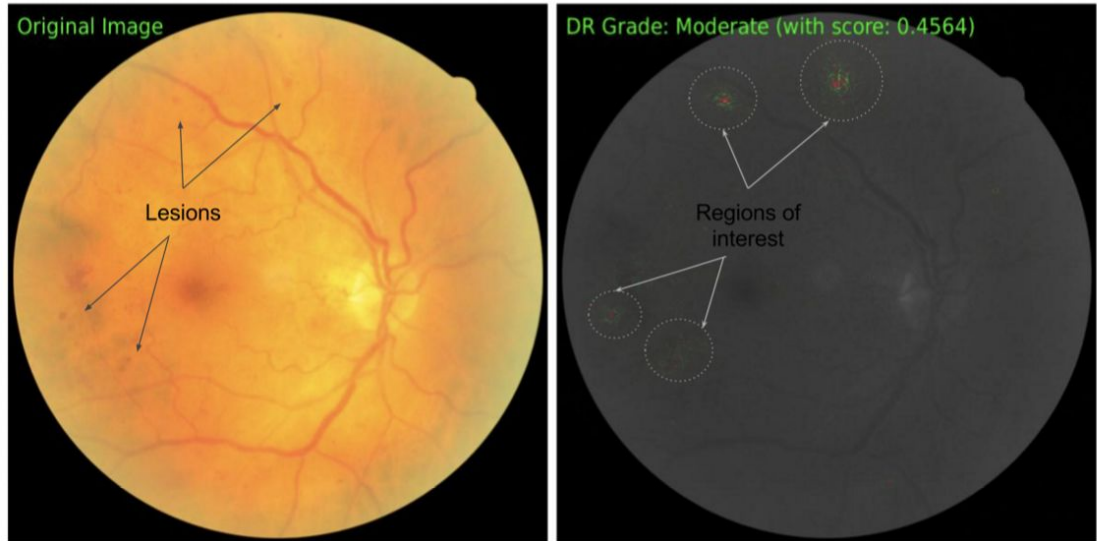
Precision Medicine ([Caruana et al., 2015](#))

- Predict probability of death for patient with pneumonia
 - high probability -> hospital/ICU
 - low probability -> treated as outpatients
- ML models make mistakes
 - biases in the dataset exist for asthma
 - asthma is a serious condition that has to be admitted to hospitals or even ICUs
 - mistakes from neural nets on asthma prevented clinical trials in mid-90's



Medical Imaging ([Sundararajan et al., 2017](#))

- A Diabetic Retinopathy Grade is detected from a retinal fundus image
- Gradient based techniques are used to demonstrate the basis of the model's decisions

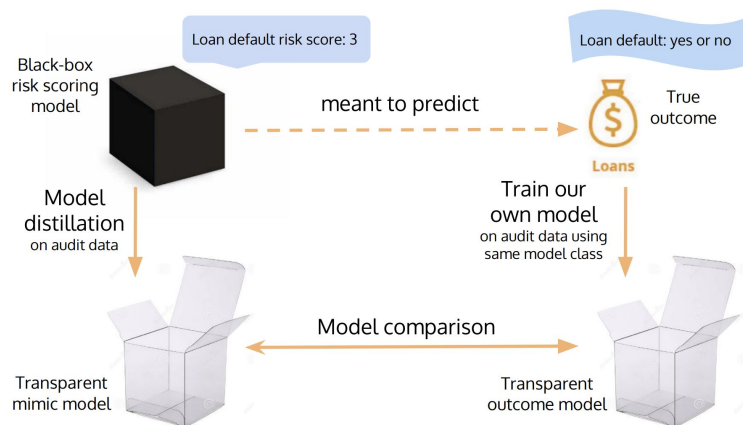


Legal Tool Explanation ([Tan et al., 2018](#))

- A legal case ([Lightbourne, 2017](#)) challenges the use of a software called "COMPAS" when sentencing individuals to prison
 - alleged use of gender and race features in its decision making process
 - algorithm details are considered to be trade secrets and are not transparent

- **Mimicking Model Behaviors**

- Model distillation is used to mimic model behaviors
- Interpretable models are used to explain the behaviors of black box models



Question Answering ([Seo et al, 2017](#))

- Explanation of question answering systems
 - highlighted keywords on context & questions

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at [Levi's Stadium in the San Francisco Bay Area at Santa Clara, California](#). As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

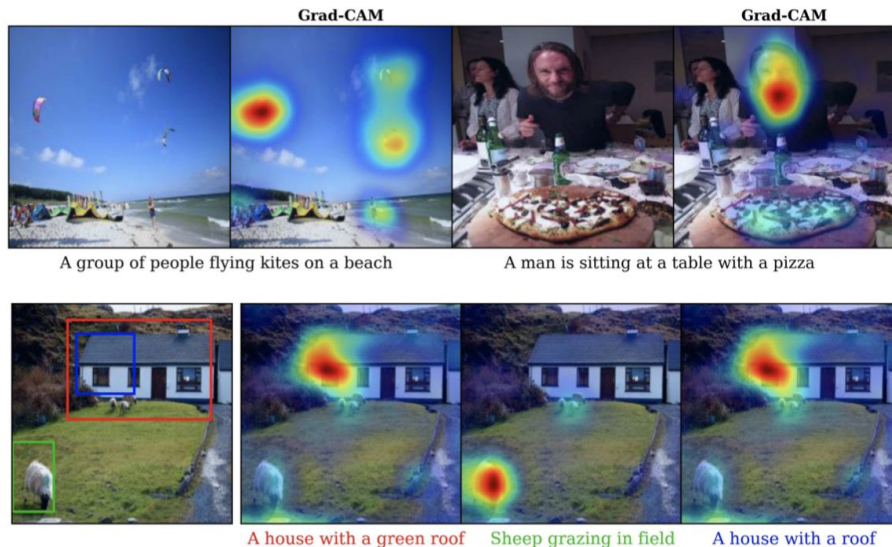
Where		at, the, at, Stadium, Levi, in, Santa, Ana
did		[]
Super		Super, Super, Super, Super, Super
Bowl		Bowl, Bowl, Bowl, Bowl, Bowl
50		50
take		
place		
?		initiatives

There are [13](#) natural reserves in Warsaw—among others, Bielany Forest, Kabaty Woods, Czerniaków Lake. About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the otter, beaver and hundreds of bird species. There are also several lakes in Warsaw—mainly the oxbow lakes, like Czerniaków Lake, the lakes in the Łazienki or Wilanów Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent—the majority are emptied before winter to clean them of plants and sediments.

How		[]
many		hundreds, few, among, 15, several, only, 13, 9
natural		natural, of
reserves		reserves
are		are, are, are, are, are, includes
there		[]
in		[]
Warsaw		Warsaw, Warsaw, Warsaw
?		inter species

Image Caption Generation ([Selvaraju et al., 2017](#))

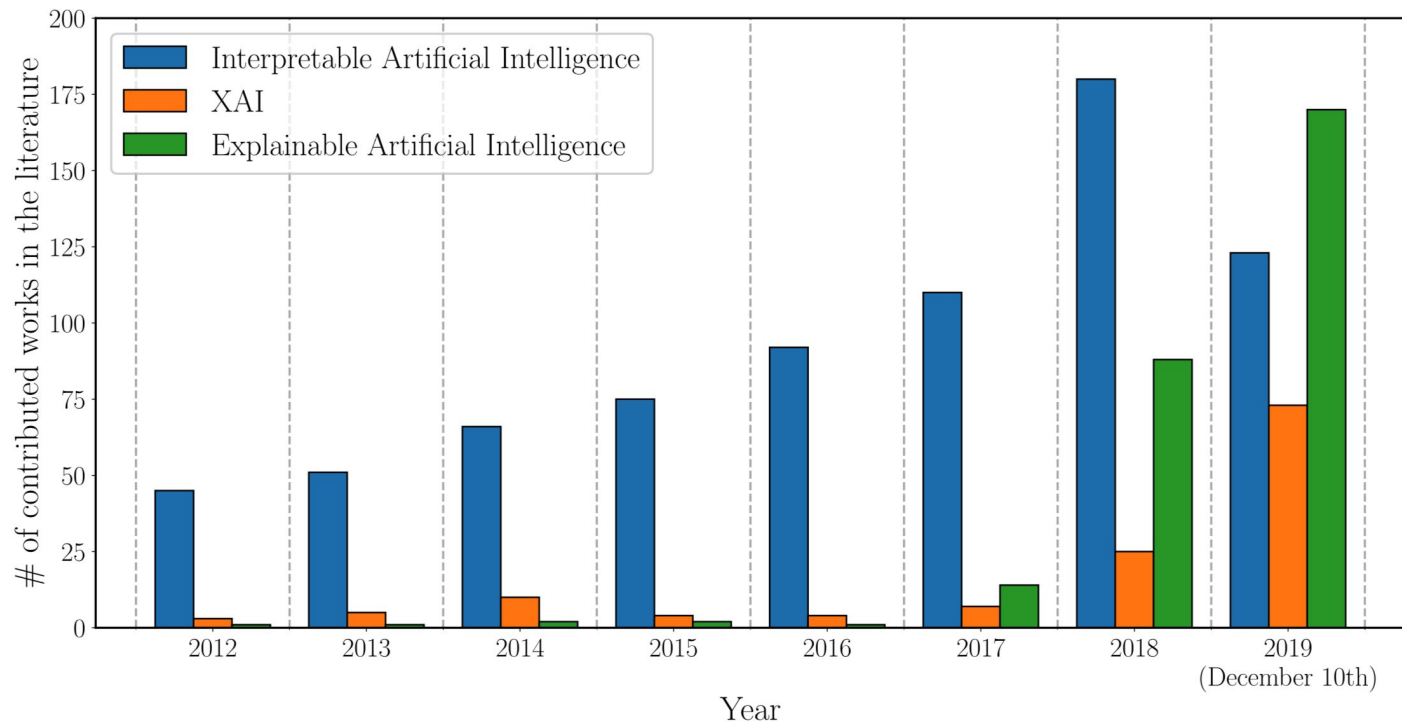
- Highlighted regions explaining an image caption generation algorithm.



Right to Explanation

- Credit Scores in United States
 - Equal Credit Opportunity Rights (Regulation B of the [Code of Federal Regulations](#))
 - Creditors are required to notify applicants of action taken with statement of *specific*
- European Union General Data Protection Regulation
 - GDPR 1995 provided a legally disputed form of a right to an explanation [Recital 71](#)
 - "The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing..."
- France
 - In a decision taken on the basis of an algorithmic treatment, the rules that define that treatment and its "principal characteristics" must be communicated to the citizen upon request
 - the degree and the mode of contribution of the algorithmic
 - the data processed and its source
 - the treatment parameters, and where appropriate, their weighting
 - the operations carried out by the treatment.

Surge in Explainable Research ([Arrieta et al., 2019](#))



Outline

- Fair Representation Learning
- ML Interpretability
- Intrinsically Interpretable Models
 - Simple interpretable models
 - Intrinsically interpretable techniques for deep learning
- Interpretability Concepts
 - Intrinsic and post hoc methods
 - model-specific and model-agnostic methods
 - Local and global interpretable methods
 - Interpretability and performance trade-offs

Intrinsically interpretable models

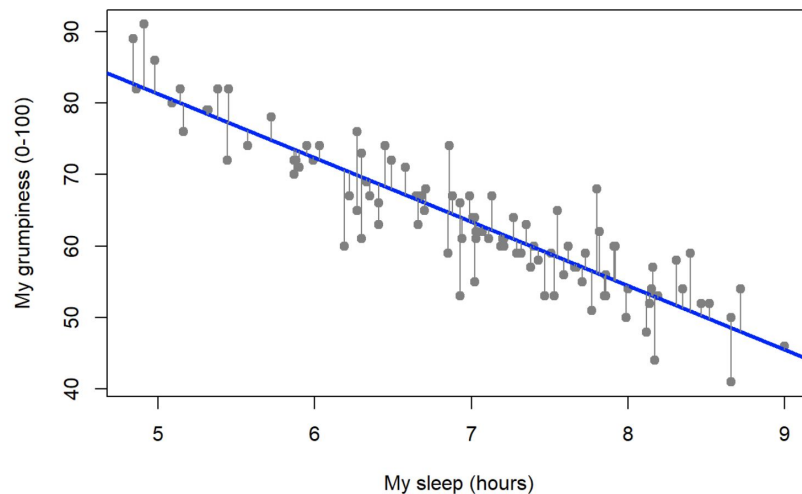
- Models that are interpretable by design
- No post-processing steps are needed to achieve interpretable.

Linear Regression

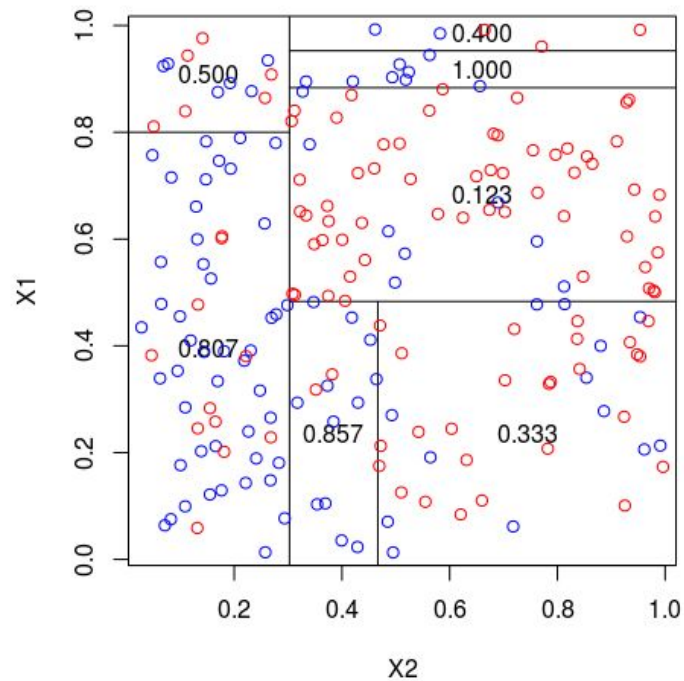
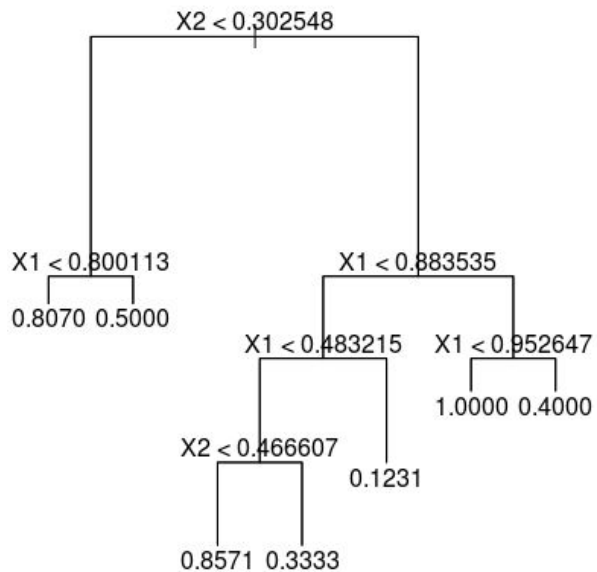
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$



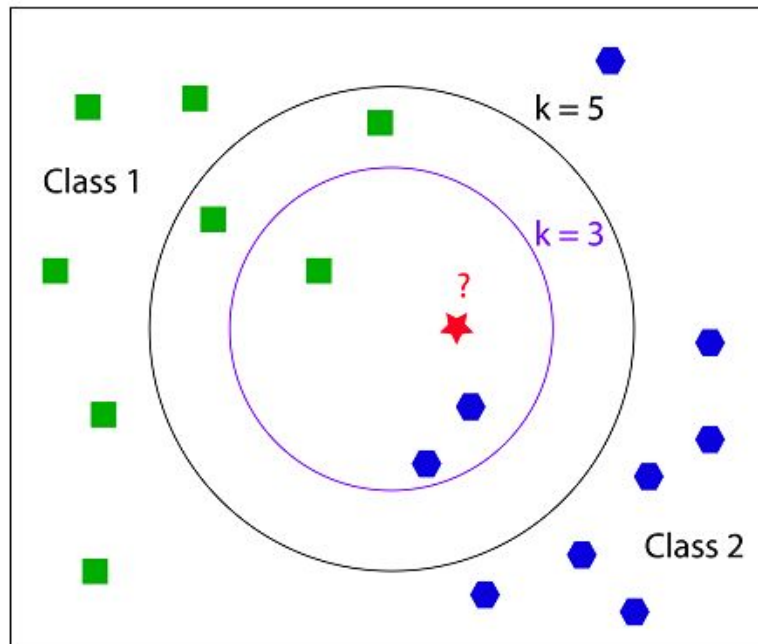
interpretable components



Decision Trees

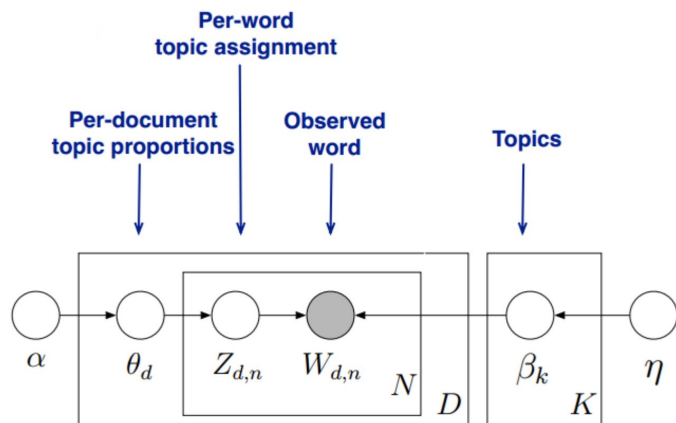


K-Nearest Neighbors



Bayesian Models

Latent Dirichlet Allocation ([Blei et al., 2003](#))



Documents

XXXX XXXX I purchased a vehicle from XXXX
XXXX XXXX XXXX which I traded in my
XX/XX/XXXX Volvo. I then signed contract and
release of liability to the dealer. I still have the
contract. Three years later I received a letter
from a collection agency that I owe them XXXX
dollars for the car I traded in, that was towed
from XXXX XXXX XXXX XXXX said at the time
the car was still in my name. So I went back to
the dealer and the dealer before was sold to
another company. I spoke with XXXX XXXX and
did what they told me and it is still on my credit
report. I am really frustrated on what I am going
through. The collectors will not listen to me.
What can I do. The agency is XXXX Collections
in XXXX XXXX California.

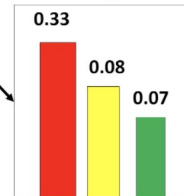
Topics β_k

car	0.23
vehicle	0.18
finance	0.09
...	...

collect	0.25
agenc	0.13
recover	0.05
...	...

receiv	0.23
letter	0.17
send	0.1
...	...

Topic proportions θ_d



Outline

- ML Interpretability
- Intrinsically Interpretable Models
 - Simple interpretable models
 - Intrinsically interpretable techniques for deep learning
- Interpretability Concepts
 - Intrinsic and post hoc methods
 - model-specific and model-agnostic methods
 - Local and global interpretable methods
 - Interpretability and performance trade-offs

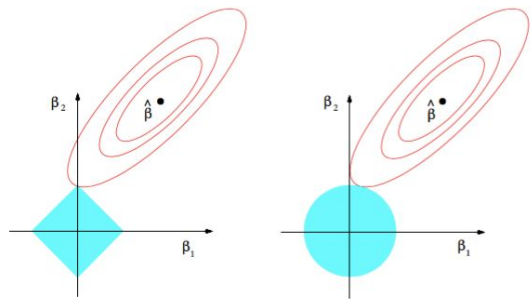
Sparsity

- Controls the sparsity of model parameters when learning a model
- Popular choices
 - L1 regularization

$$||\mathbf{w}||_1 = |w_1| + |w_2| + \dots + |w_N|$$

- L2 regularization

$$||\mathbf{w}||_2 = (w_1^2 + w_2^2 + \dots + w_N^2)^{\frac{1}{2}}$$



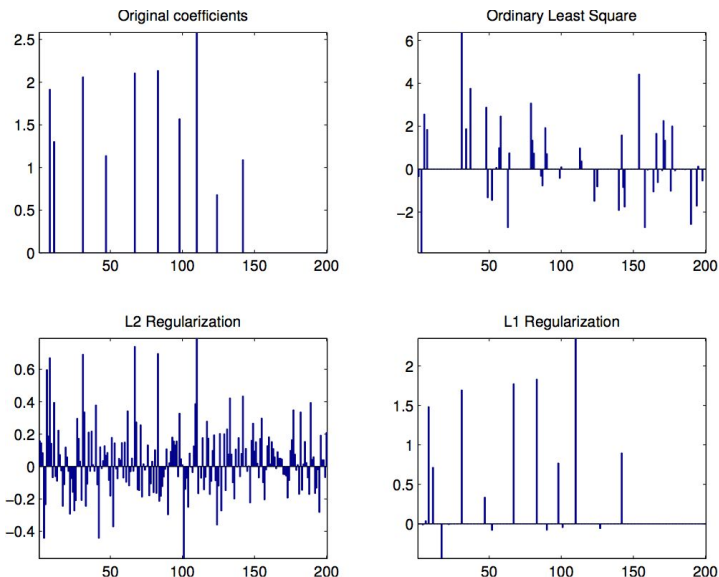
Sparsity for Interpretable Linear Regression

- In the case of linear regression
 - $\hat{y} = w_1x_1 + w_2x_2 + \dots + w_Nx_N + b$
- Linear regression with L1 regularization

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^N |w_i|$$

- Linear Regression with L2 regularization

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^N w_i^2$$

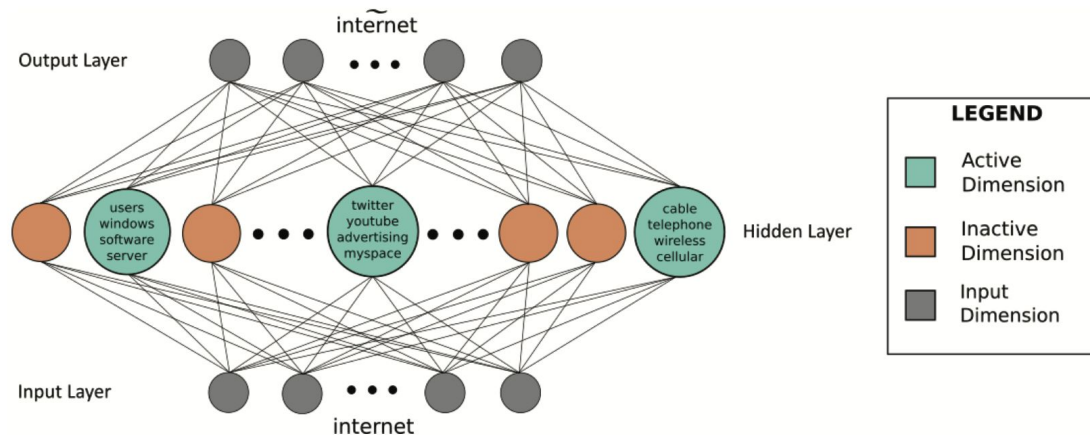


Interpretable Neural Embeddings ([Subramanian et al., 2018](#))

- Use Sparse Autoencoder to generated interpretable word embeddings

$$\tilde{x} = g(z^x)$$

$$z^x = f(x)$$



Interpretable Neural Embeddings ([Subramanian et al., 2018](#))

- Add sparse constraints to the auto-encoder
 - Unit-wise sparsity constraint enforces sparsity for each unit
 - Layer-wise sparsity constraint controls the number of activations for each layer

$$\mathcal{L} = \underbrace{\sum_{x_i} (|x_i - \tilde{x}_i|^2)}_{\text{reconstruction loss}} + \underbrace{\sum_h z_h^{x_i} \times (1 - z_h^{x_i})}_{\text{unit-wise sparsity constraint}} + \underbrace{\max(0, \sum_m (\sigma(z_h^{x_i})/m - \rho)/N)}_{\text{layer-wise sparsity constraint}}$$

target
activation rate
↓

Clustering of Words for the Top Participating Dimension

	Initial GloVe vectors	Initial word2vec vectors
mathematics	<u>intelligence, government, foreign, security</u> <u>kashmir, algorithms, heat, computational</u> <u>robes, tito, aviation, backward, dioceses</u>	<u>leukemia, enterprises, wingspan, info, booker</u> <u>ore, greens, badminton, hymns, clay</u> <u>asylum, intercepted, skater, rb, flats</u>
remote	<u>thousands, residents, palestinian, police</u> <u>kashmir, algorithms, heat, computational</u> <u>tamil, guerrilla, spam, rebels, infantry</u>	<u>basilica, sensory, ranger, chapel, memorials</u> <u>microsoft, sr, malaysia, jan, cruisers</u> <u>capt, obey, tents, overdose, cognitive, flats</u>
internet	<u>thousands, residents, palestinian, police</u> <u>intelligence, government, foreign, security</u> <u>nhl, writer, writers, drama, soccer</u>	<u>cardinals, tsar, papal, autobiography, befriends</u> <u>gases, gov, methane, graph, buttons</u> <u>longitude, carr, precipitation, snowfall, homer</u>
	SPINE w/ GloVe	SPINE w/ word2vec
mathematics	<u>sciences, honorary, faculty, chemistry, bachelor</u> <u>university, professor, graduate, degree, bachelor</u> <u>mathematical, equations, theory, quantum</u>	<u>algebra, exam, courses, exams, math</u> <u>theorem, mathematical, mathematician, equations</u> <u>doctorate, professor, doctoral, lecturer, sociology</u>
remote	<u>territory, region, province, divided, district</u> <u>wilderness, ski, camping, mountain, hiking</u> <u>rugged, mountainous, scenic, wooded, terrain</u>	<u>villages, hamlet, villagers, village, huts</u> <u>mountainous, hilly, impoverished, poorest, populated</u> <u>button, buttons, click, password, keyboard</u>
internet	<u>windows, users, user, software, server</u> <u>youtube, mspace, twitter, advertising, ads</u> <u>wireless, telephone, cellular, cable, broadband</u>	<u>hacker, spam, pornographic, cyber, pornography</u> <u>browser, app, downloads, iphone, download</u> <u>cellular, subscriber, verizon, broadband, subscribers</u>

Performance on Intrusion Detection Test

- Human annotators are asked to select odd words from a group.

Select the odd one out:

- ☐ grandchildren
- ☐ sons
- ☐ visual
- ☐ grandson
- ☐ granddaughter

Sample Question for Intrusion
Detection Test

GloVe (original)	SPOWV (w/ GloVe)	SPINE (w/ GloVe)
22.97	28.18	68.35
Word2vec (original)	SPOWV (w/ word2vec)	SPINE (w/ word2vec)
26.08	41.75	74.83

Precision scores

GloVe (original)	SPOWV (w/ GloVe)	SPINE (w/ GloVe)
76%, 22%	74%, 21%	83%, 47%
Word2vec (original)	SPOWV (w/ word2vec)	SPINE (w/ word2vec)
77%, 18%	79%, 28%	91%, 48%

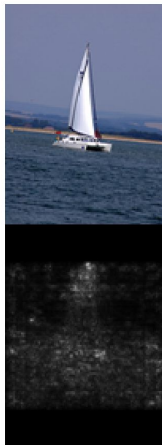
Inter-annotator agreement
across different models

Regularization for Saliency Maps ([Plumb et al, 2019](#))

- Saliency Maps
 - A gradient based method to generate visualizations to interpret deep neural networks

x

$$e(f, x) = \frac{\partial f(y|X)}{\partial X} \bigg|_{X=x}$$

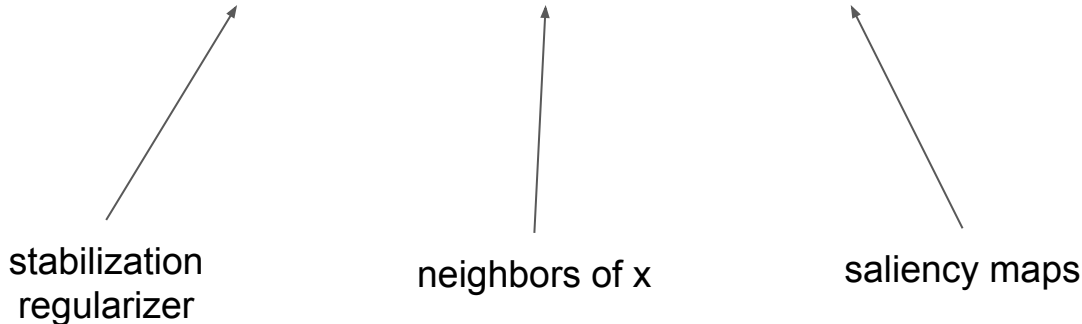


Regularization for Saliency Maps ([Plumb et al, 2019](#))

- Saliency Maps
 - $e(x, f)$ generates a saliency map for a deep learning model f on a given image x
- Stabilization Regularization
 - Stabilizes the saliency map of x and its nearby samples

$$S(f, x) = \mathbb{E}_{x' \sim N_x} ||e(f, x) - e(f, x')||_2^2$$

stabilization
regularizer



neighbors of x

saliency maps

Regularization for Saliency Maps ([Plumb et al, 2019](#))

- Saliency Maps
 - $e(x, f)$ generates a saliency map for a deep learning model f on a given image x
- Stabilization Regularization
 - Stabilizes the saliency map of x and its nearby samples

stabilization
regularizer

$$S(f, x) = \mathbb{E}_{x' \sim N_x} \|e(f, x) - e(f, x')\|_2^2$$

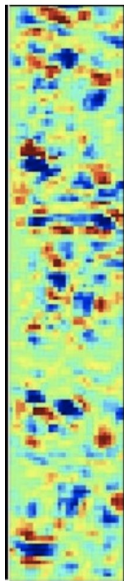
final loss
function

$$\mathcal{L} = \frac{1}{N} \sum_i \mathcal{L}(f(x_i), y_i) + \gamma S(f, x_i)$$

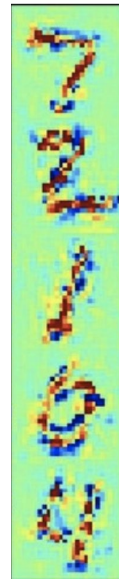
Regularization for Saliency Maps ([Plumb et al, 2019](#))



mnist samples



saliency map without
regularization

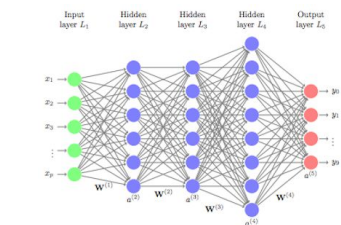


saliency map with
regularization

Bayesian Deep Learning

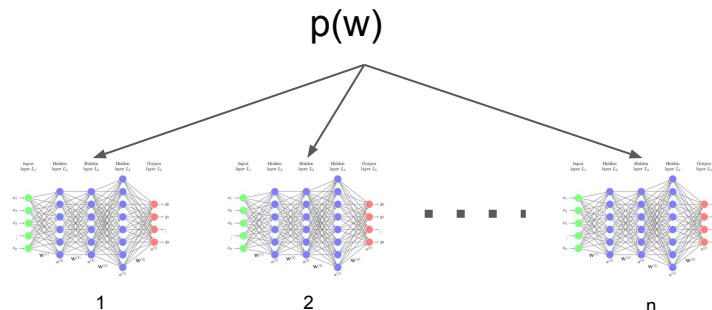
- Modeling the Distributions of Neural Network Parameters
 - A distribution of neural networks co-exist at any time
 - Networks are initialized using a prior and each play a role in modeling uncertainties

A Deep Neural Network with Deterministic Parameters



$$w \sim \text{uniform}(-1, 1)$$

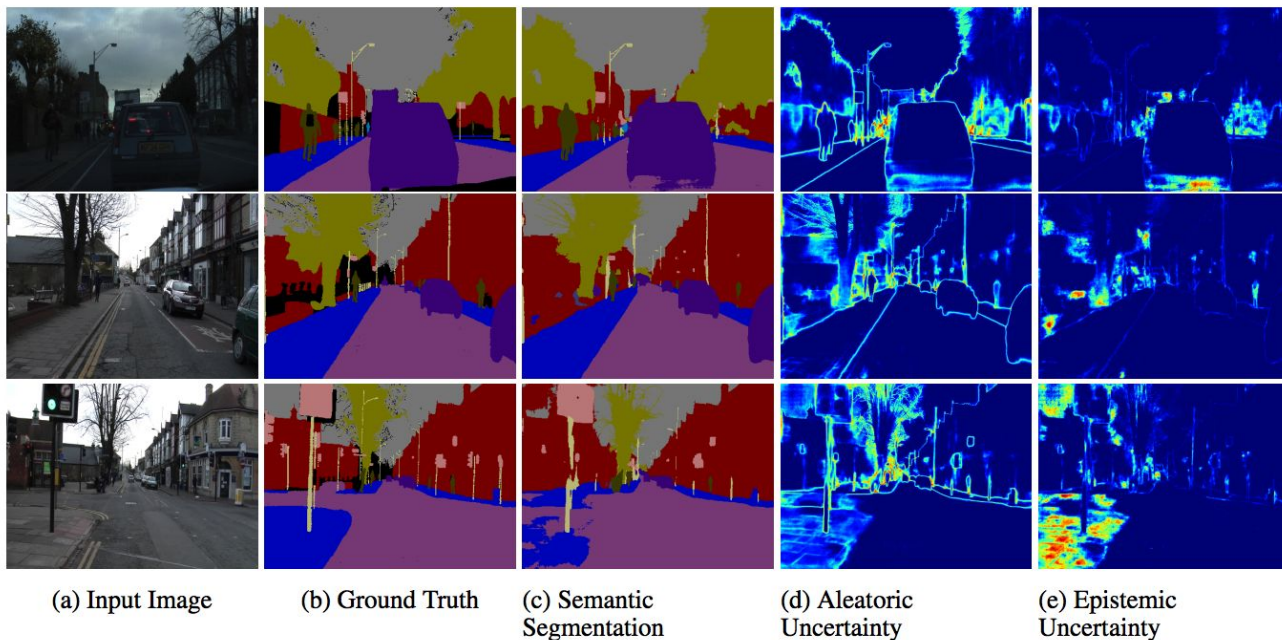
A Bayesian Deep Neural Network with a distribution over parameters



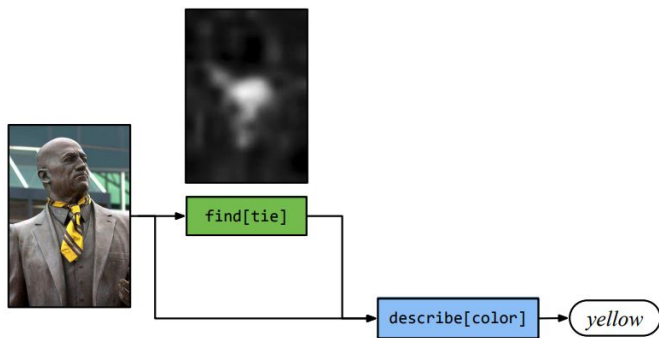
$$w^i \sim \mathcal{N}(0, I)$$

Epistemic and Heteroscedastic Uncertainty ([Kendall et al, 2017](#))

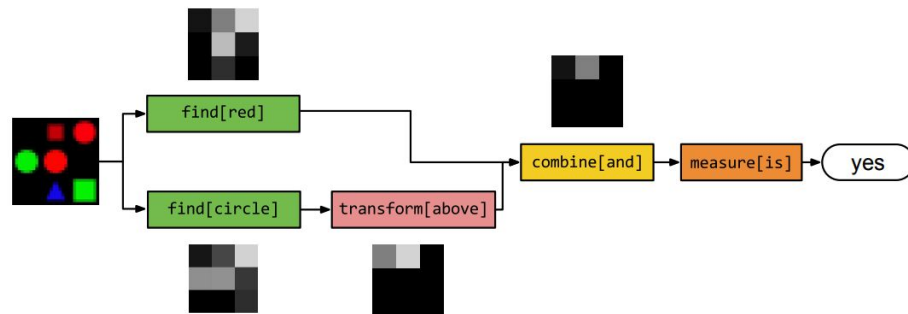
- Increased aleatoric uncertainty on object boundaries and for objects far from the camera.
- increased epistemic uncertainty for semantically and visually challenging pixels



Neural Modular Networks ([Andreas et al, 2016](#))



What color is his tie?

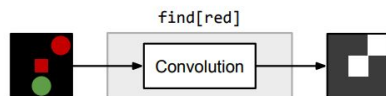


Is there a red shape above a circle?

Neural Modular Networks ([Andreas et al, 2016](#))

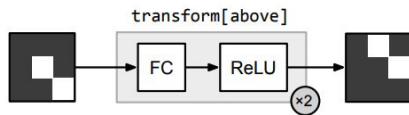
Find

Image \rightarrow Attention



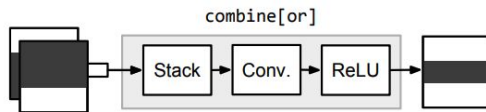
Transform

Attention \rightarrow Attention



Combine

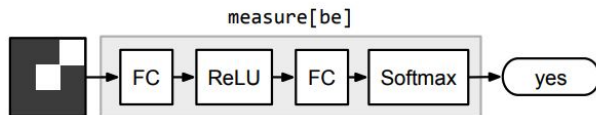
Attention \times Attention \rightarrow Attention



Neural Modular Networks ([Andreas et al, 2016](#))

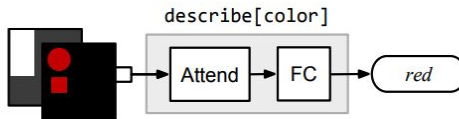
Measure

Attention \rightarrow Label





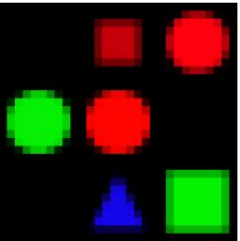


Describe

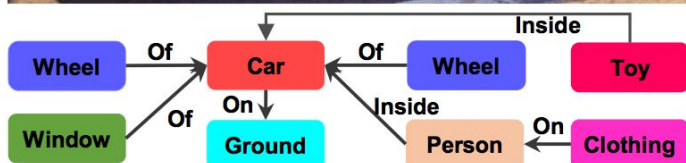
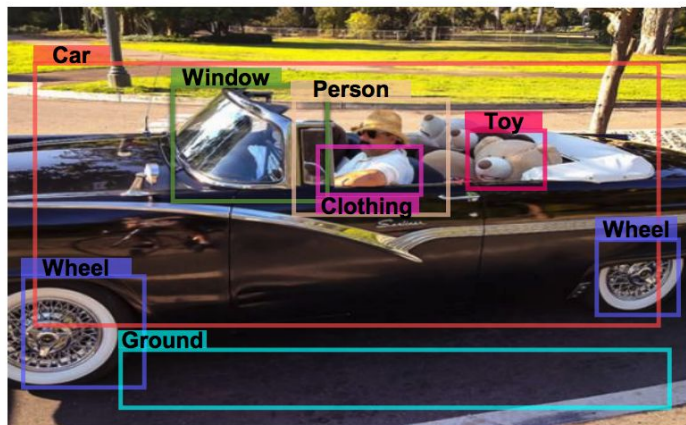
Image \times Attention \rightarrow Label



Neural Modular Networks ([Andreas et al, 2016](#))

				
<i>how many different lights in various different shapes and sizes?</i>	<i>what is the color of the horse?</i>	<i>what color is the vase?</i>	<i>is the bus full of passengers?</i>	<i>is there a red shape above a circle?</i>
<code>describe[count](find[light])</code>	<code>describe[color](find[horse])</code>	<code>describe[color](find[vase])</code>	<code>describe[is](combine[and](find[bus], find[full])</code>	<code>measure[is](combine[and](find[red], transform[above](find[circle])))</code>
four (four)	brown (brown)	green (green)	yes (yes)	yes (yes)

Neural Logic Induction Learning ([Yang et al, 2020](#))



“An object that has wheels and windows is a car”

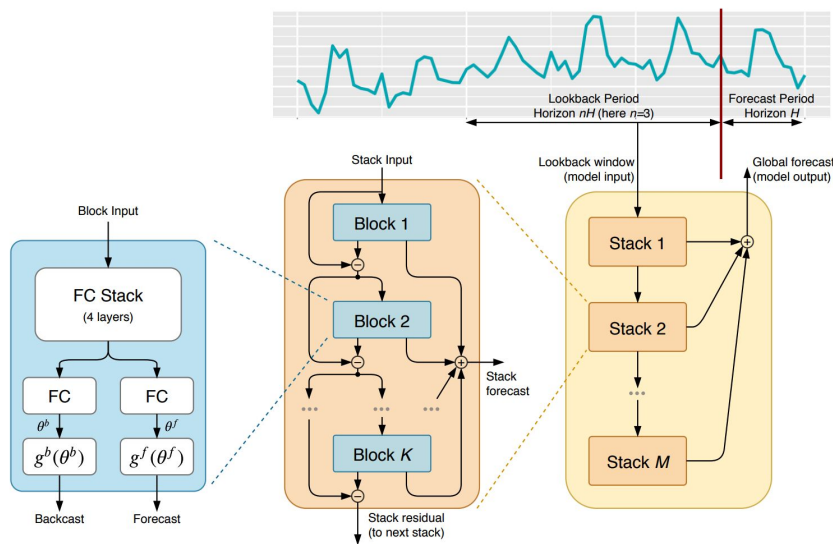
$$\text{Car}(\text{img1}) \leftarrow \text{Of}(\text{img2}, \text{img3}) \wedge \text{Window}(\text{img4}) \wedge \text{Of}(\text{img5}, \text{img6}) \wedge \text{Wheel}(\text{img7})$$

“An object that is inside the car with clothing is a person”

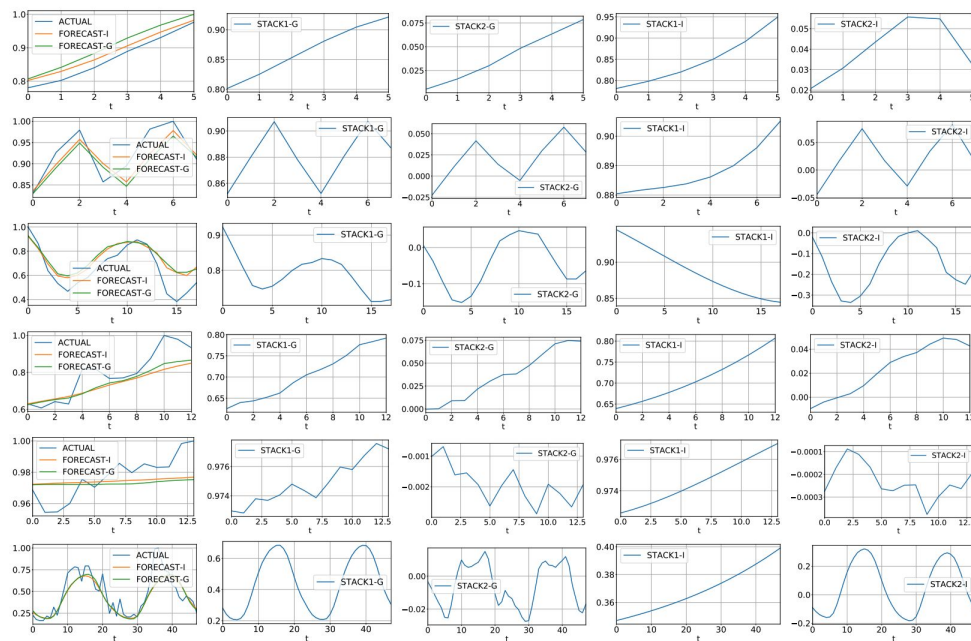
$$\text{Person}(\text{img8}) \leftarrow \text{Car}(\text{img9}) \wedge \text{Inside}(\text{img10}, \text{img11}) \wedge \text{On}(\text{img12}, \text{img13}) \wedge \text{Clothing}(\text{img14})$$

Neural Basis Expansion ([Oreshkin et al, 2020](#))

- Time series forecasting
 - Given historical data, predict future values.



Neural Basis Expansion ([Oreshkin et al, 2020](#))



(a) Combined

(b) Stack1-G

(c) Stack2-G

(d) StackT-I

(e) StackS-I

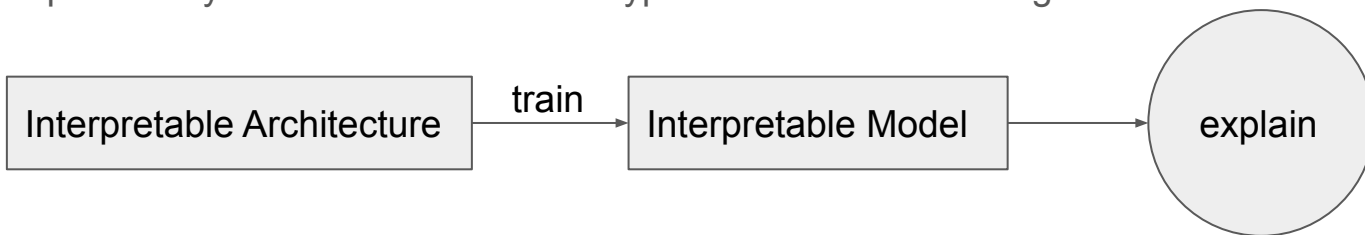
Outline

- Fair Representation Learning
- ML Interpretability
- Intrinsically Interpretable Models
 - Simple interpretable models
 - Intrinsically interpretable techniques for deep learning
- **Interpretability Concepts**
 - Intrinsic and post hoc methods
 - model-specific and model-agnostic methods
 - Local and global interpretable methods
 - Interpretability and performance trade-offs

Intrinsic and Post Hoc Interpretability

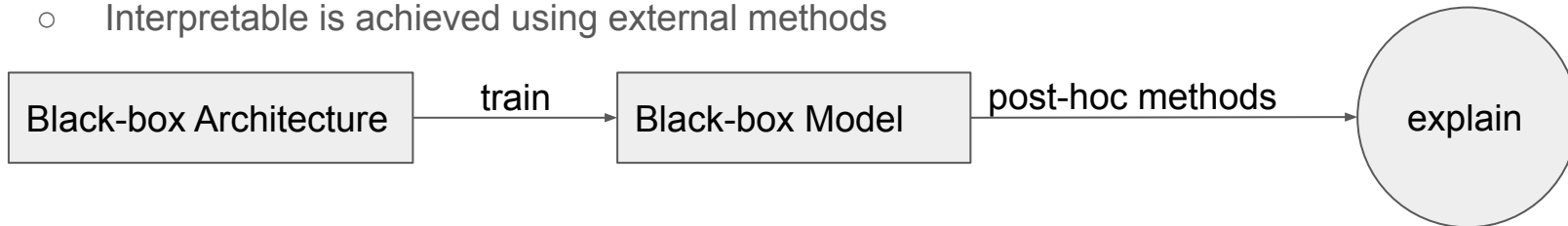
- Intrinsically interpretable models

- Interpretable is achieved by model design
- ML models are explainable by itself
- Explainability is often achieved as a byproduct of model training



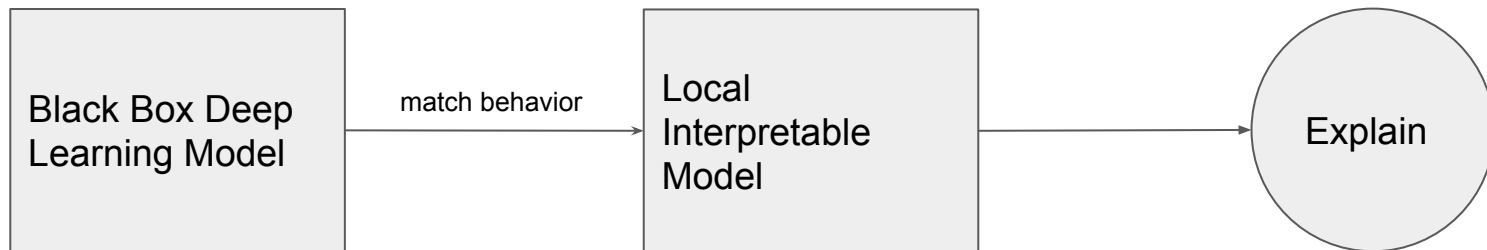
- Post Hoc/Model-specific methods

- Explainability is often achieved after the model is trained
- Interpretable is achieved using external methods



Post Hoc Interpretability

- One of the way to achieve Post Hoc Interpretability is to deploy a local proxy model
- We will introduce more about Post Hoc Interpretable methods in the next lecture.



Model Specific and Model Agnostic Methods

- Model Specific Methods

- Techniques that can be used for a specific architecture
- Usually preferable when you have the ability to design your own model
- Model specific techniques might compromise the performance of your model
- Requires training the model using a dataset
- Intrinsic methods are by definition model specific

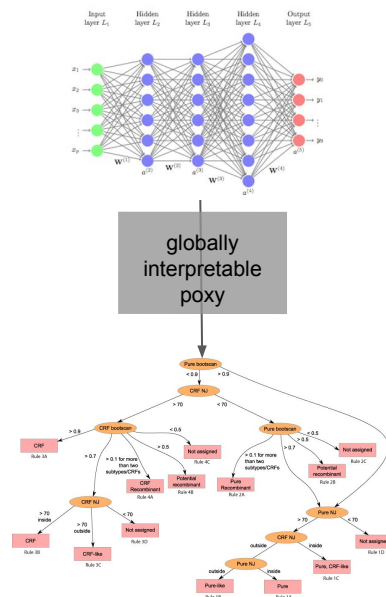
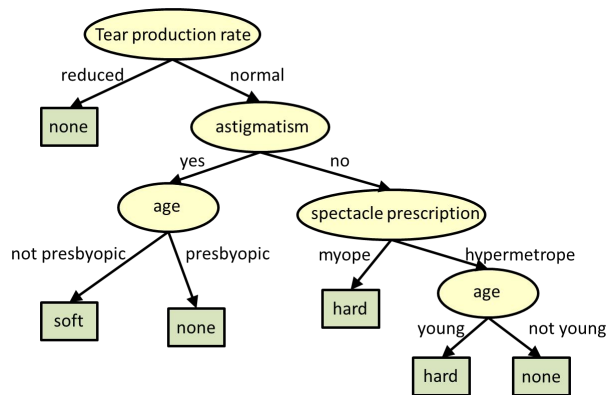
- Model-agnostic Methods

- Techniques that can be used across many black box models
- Model-agnostic methods will not affect the performance of your model
- Do not require training the model
- Will be covered in the next lecture
- Post hoc methods are usually model-agnostic

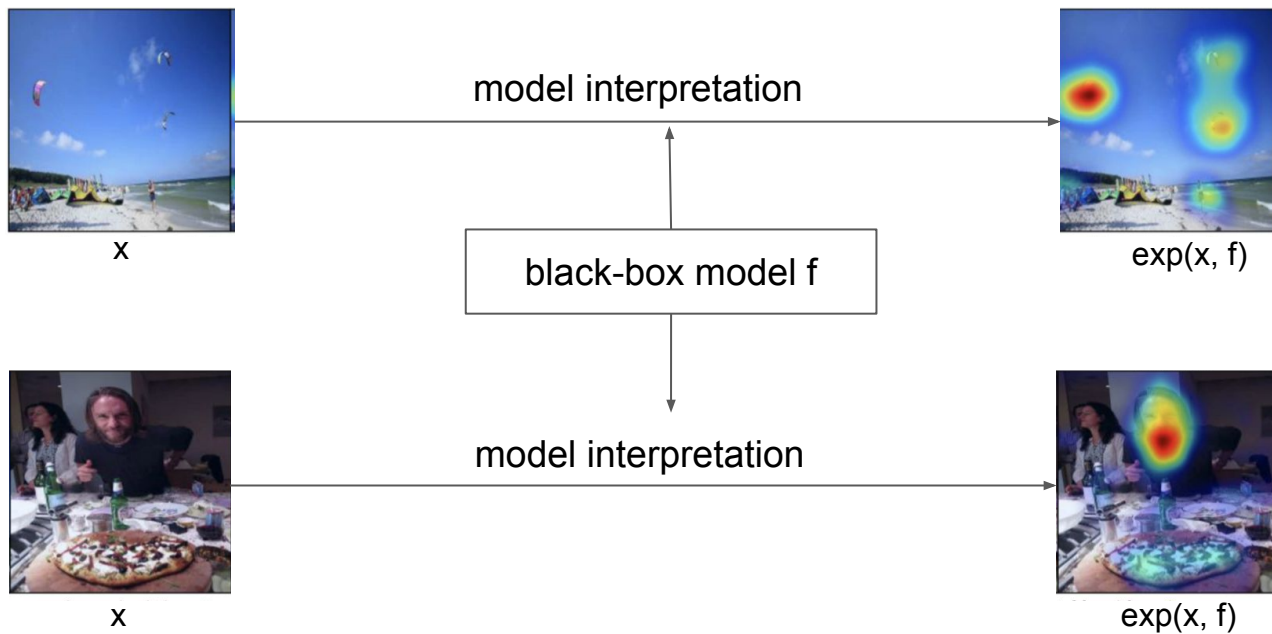
Global and Local Interpretability

- Global Interpretability
 - Explains the entire ML model at once from input to prediction
 - 1) Holistic Model Interpretability
 - 2) Modular Level Interpretability
 - e.g., Decision Trees, Linear regression
- Local Interpretability
 - Explain how predictions change for when input changes
 - 1) For a single prediction
 - 2) for a group of predictions

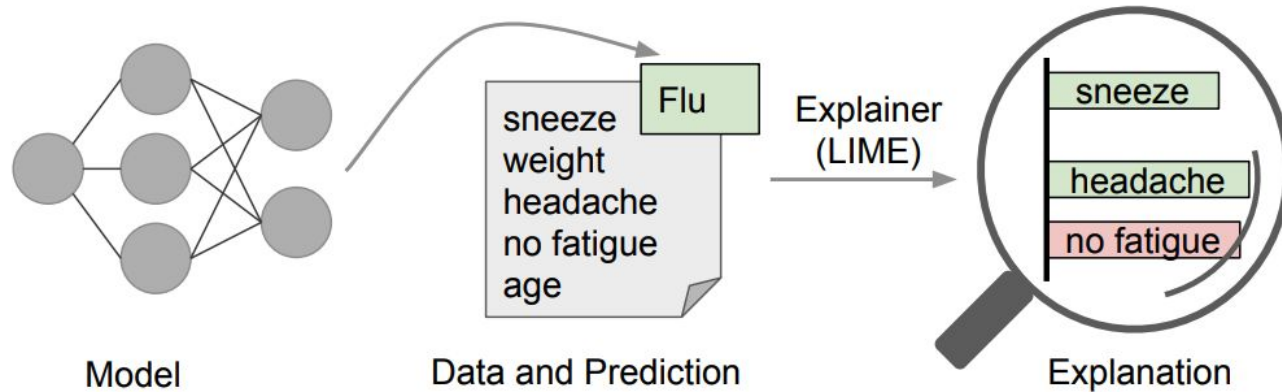
Global Interpretability



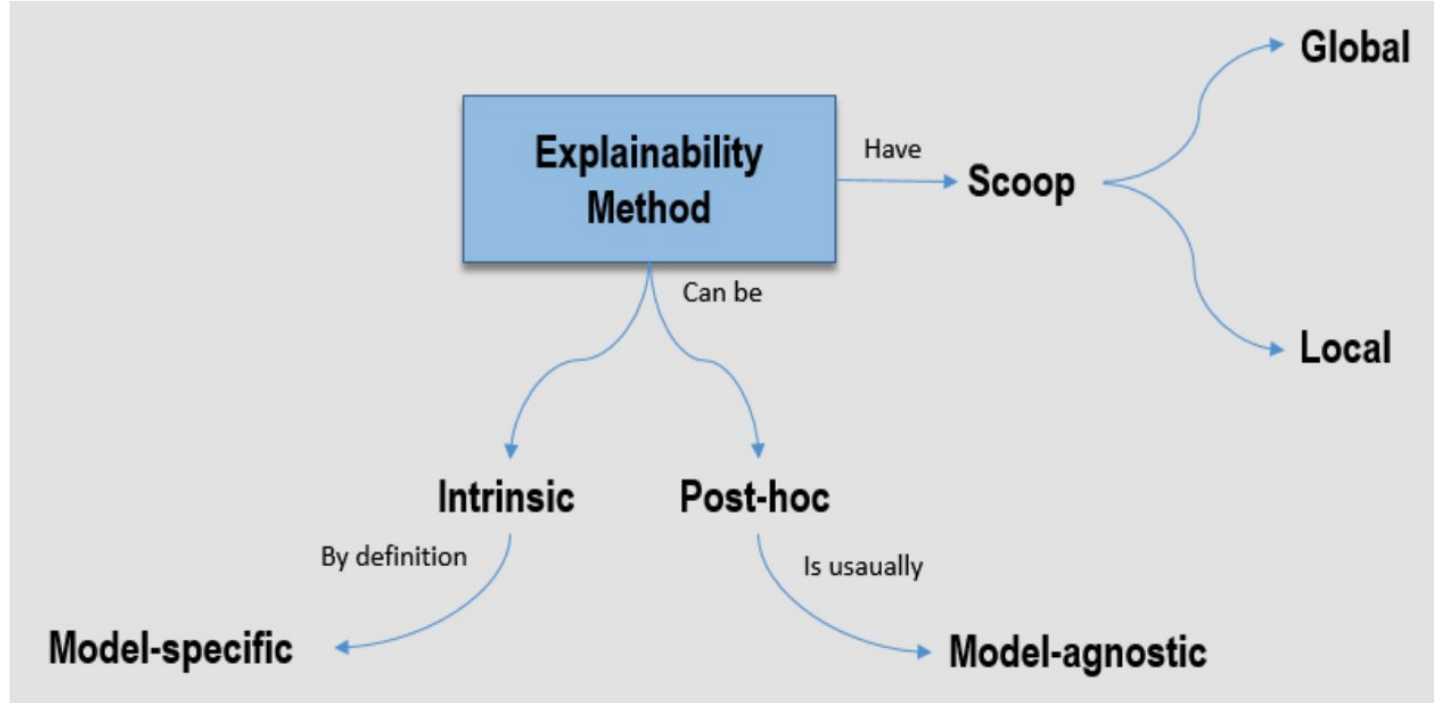
Local Interpretability



Local Interpretability



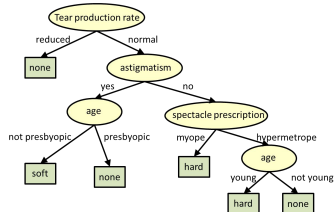
An Ontology of AI Explainability ([ADADI el al, 2018](#))



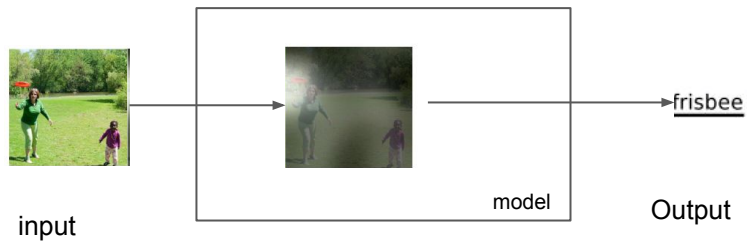
The Big Picture

Intrinsic

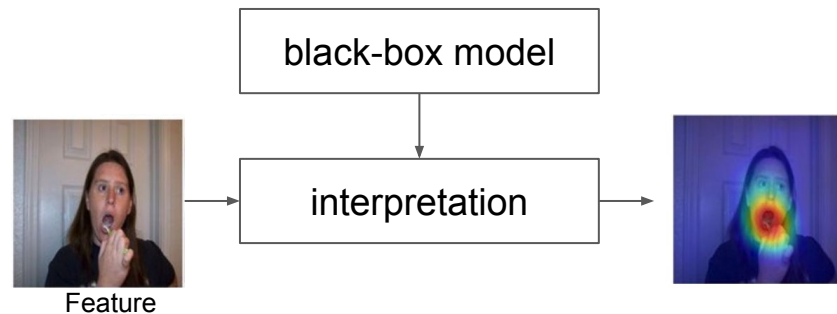
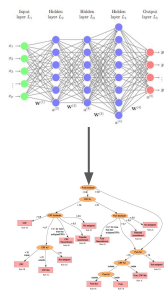
Globally Interpretable



Locally Interpretable



Post Hoc

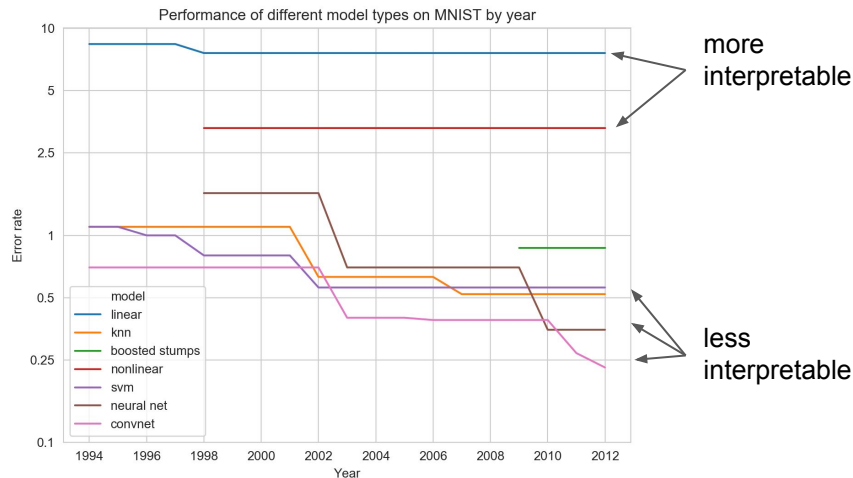


Interpretability and Performance Trade-offs

- highly performed models tend to be less interpretable.
- Can powerful models with complex structures be interpretable at the same time?



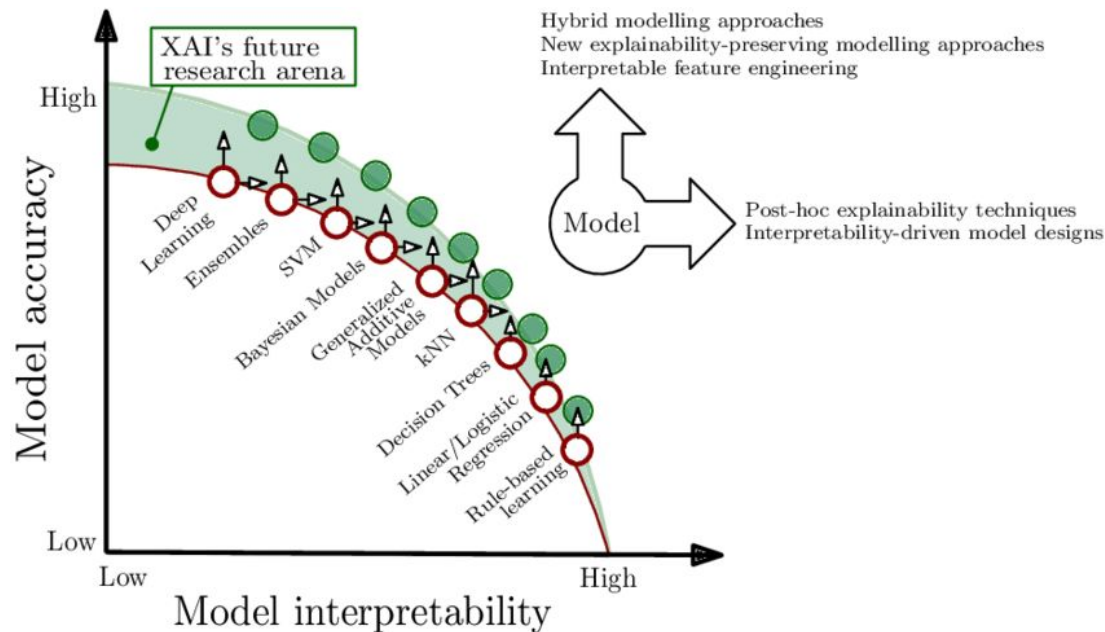
MNIST Dataset



<http://yann.lecun.com/exdb/mnist/>

<https://soph.info/2018/11/08/mnist-history/>

Interpretability and Performance Trade-offs ([Arrieta et al., 2019](#))



Required Reading

- Molnar: Ch 2, Ch 4

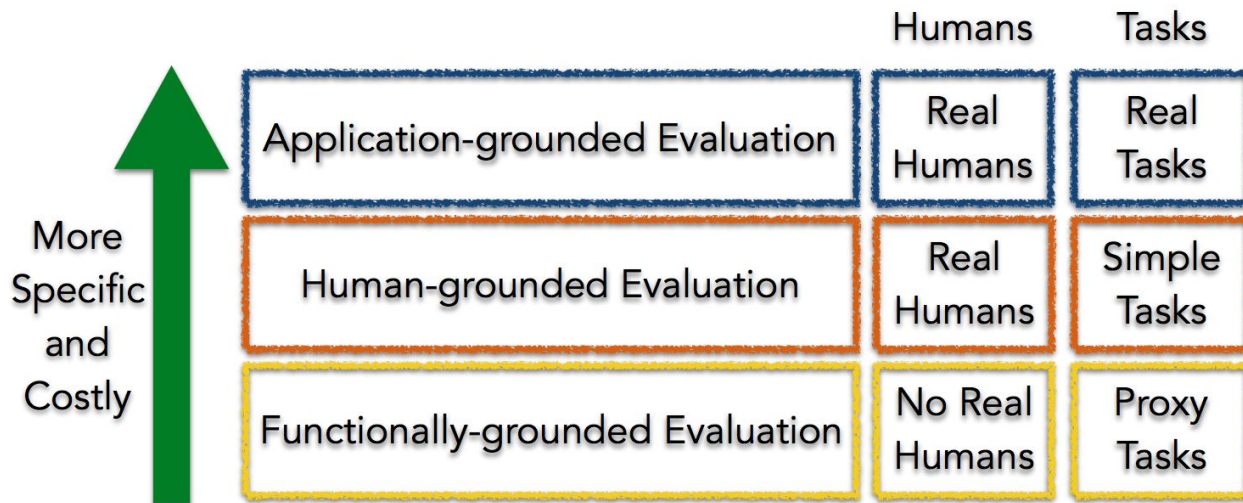
Reading Assignments (Pick One)

- Lipton, Z. C. The mythos of model interpretability. Queue, 2018
- Adadi, Amina, and Mohammed Berrada. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI), IEEE Access 2018
- Doshi-Velez, Finale, and Been Kim. Towards a rigorous science of interpretable machine learning. Arxiv, 2017
- Wang, F., & Rudin, C. Falling rule lists. AIStats, 2015
- Adel, T., Ghahramani, Z., & Weller, A. Discovering interpretable representations for both deep generative and discriminative models, ICML 2018

Next Lecture

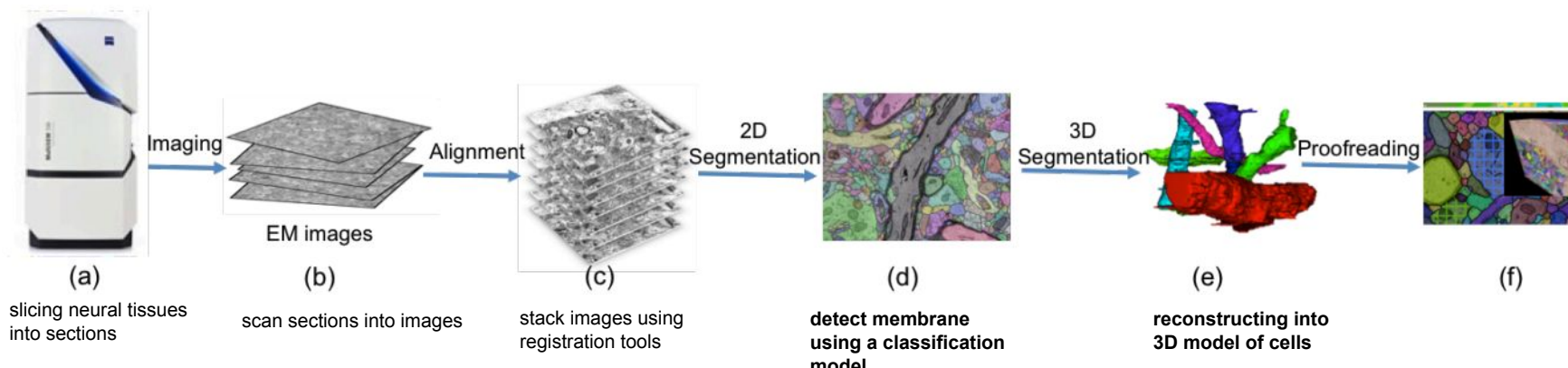
Proxy Models for Post Hoc Interpretability

Evaluations for Interpretability ([Finale Doshi-Velez et al, 2017](#))



Application-Grounded Evaluation

- Examined by Human Experts in a Specialized Domain
 - Interpretable models need to facilitate conducting a real and sophisticated task
- Automatic Neural Reconstruction from Petavoxel of Electron Microscopy Data ([Suissa-Peleg et al, 2016](#))
 - Study the dense structure of the neurons in the brain and their synapses
 - A multi-step process that involves many ML models



Human-Grounded Evaluation

- Examined by a Lay Human in a General Domain
 - Interpretable models are evaluated by average human.
- Explain a model that classifies an article into either "Christianity" or "Atheism" ([Ribeiro et al, 2016](#))
 - Amazon Mechanical Turk workers are asked to the algorithm that has better performance

Example #3 of 6

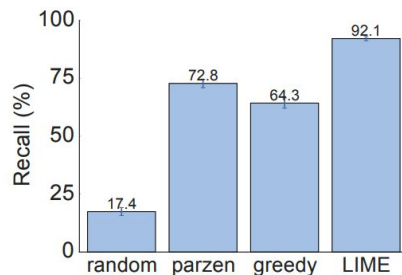
True Class: ● Atheism

[Instructions](#) [Previous](#) [Next](#)

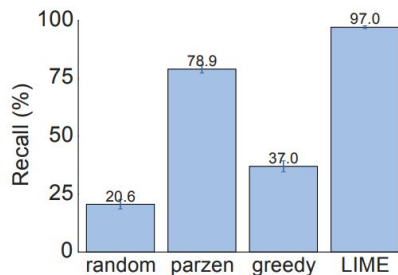
Algorithm 1	Algorithm 2
Words that A1 considers important:	Words that A2 considers important:
<div><div>GOD</div><div>mean</div><div>anyone</div><div>this</div><div>Koresh</div><div>through</div></div>	<div><div>Posting</div><div>Host</div><div>Re</div><div>by</div><div>in</div><div>Nntp</div></div>
Predicted:	Predicted:
<div><div>● Atheism</div><div>Prediction correct: ✓</div></div>	<div><div>● Atheism</div><div>Prediction correct: ✓</div></div>
Document	Document
From: pauld@verdix.com (Paul Durbin) Subject: Re: DAVID CORESH IS! GOD! Nntp-Posting-Host: sarge.hq.verdix.com Organization: Verdix Corp Lines: 8	From: pauld@verdix.com (Paul Durbin) Subject: Re: DAVID CORESH IS! GOD! Nntp-Posting-Host: sarge.hq.verdix.com Organization: Verdix Corp Lines: 8

Functionally-Grounded Evaluation

- Examined using a proxy task
- Compare selected feature from model interpretability against explanatory features ([Ribeiro et al, 2016](#))
 - Explanatory feature are labeled by human as ground truth



(a) Sparse LR



(b) Decision Tree



(a) Original Image

(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*