Interpretable machine learning papers

Presentor: Ji Gao

@ https://qdata.github.io/deep2Read/

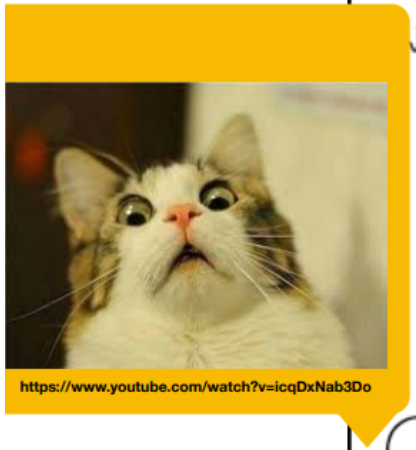
Paper list

- ICML 2017 Tutorial – Interpretable machine learning
- How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation
- Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction
- Sanity Checks for Saliency Maps

ICML 2017 Tutorial

- 0. Interpretation is hard
  - Motivation: Interpretation is important
  - Decision tree example
  - Understand everything = Impossible
- 1. Why & When we need interpretability
- 2. How to achieve interpretability
- 3. How to evaluate interpretability

Machine learning system



https://www.youtube.com/watch?v=icqDxNab3Do

https://xkcd.com/

Cost-effective Health Care (CEHC) built models to predict probability of death for patients [Cooper et al. 97]

- HasAsthma(x) ⇒ LowerRisk for pneumonia (x)

Doctors think he/she is high risk → Aggressive Treatment

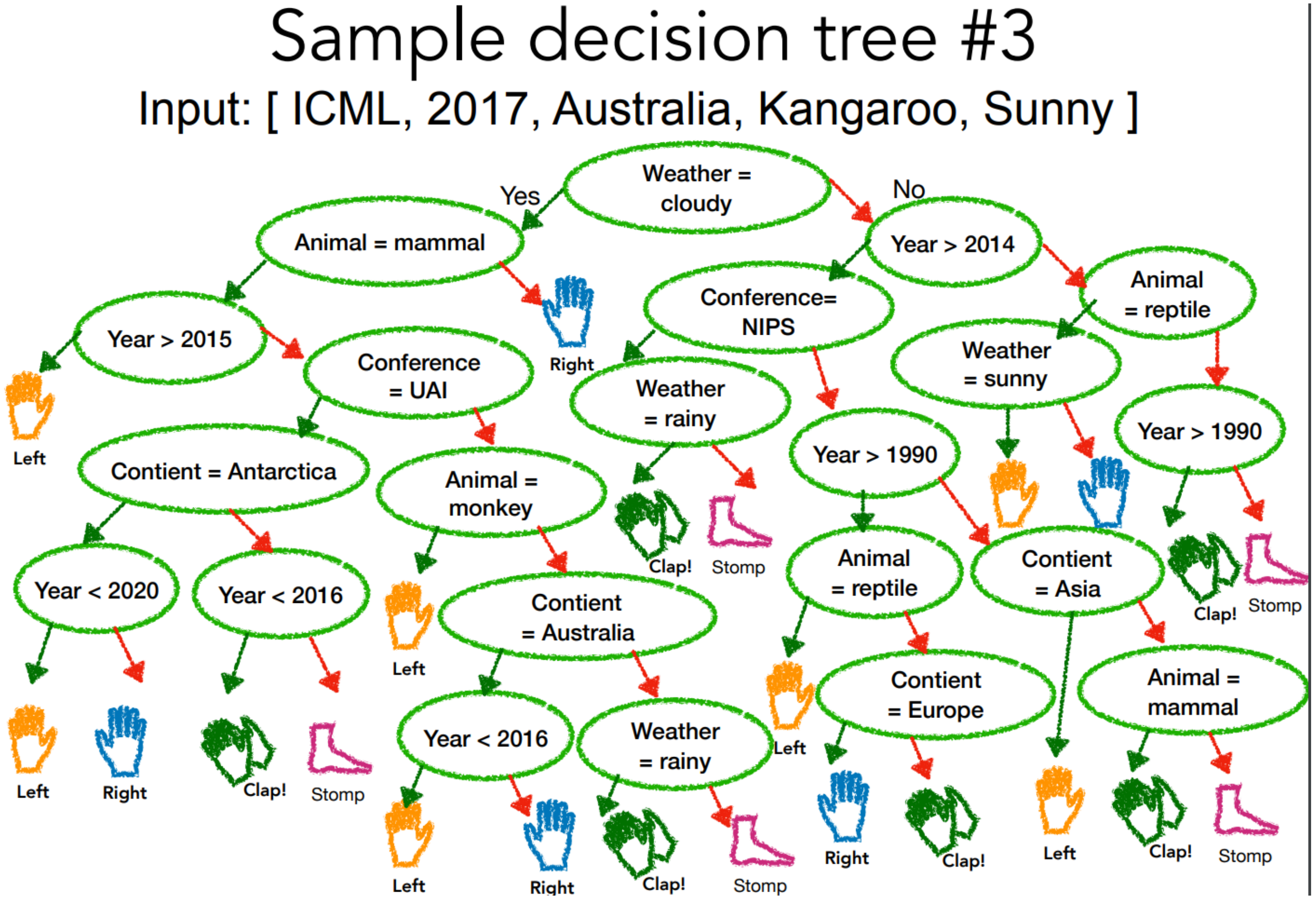
5

https://xkcd.com/

Example borrowed from [Caruana et al. '15]

Decision Tree

- Decision tree is not enough in large scale



If-else/Rule set

IF (sunny and hot) OR (cloudy and hot) OR (sunny and thirsty and bored) OR (bored and tired) OR (thirty and tired) OR (code running) OR (friends away and bored) OR (sunny and want to swim) OR (sunny and friends visiting) OR (need exercise) OR (want to build castles) OR (sunny and bored) OR (done with deadline and hot) OR (need vitamin D and sunny) OR (just feel like it)
THEN go to beach
ELSE work

Understanding all = Impossible

- Interpretability is NOT about understanding all bits and bytes of the model for all data points (we cannot).
- It’s about knowing enough for your downstream tasks.

Agenda

1. Why and when?

2. How can we do this?

Interpretation is the process of giving **explanations**

3. How can we measure 'good' explanations?

To Humans

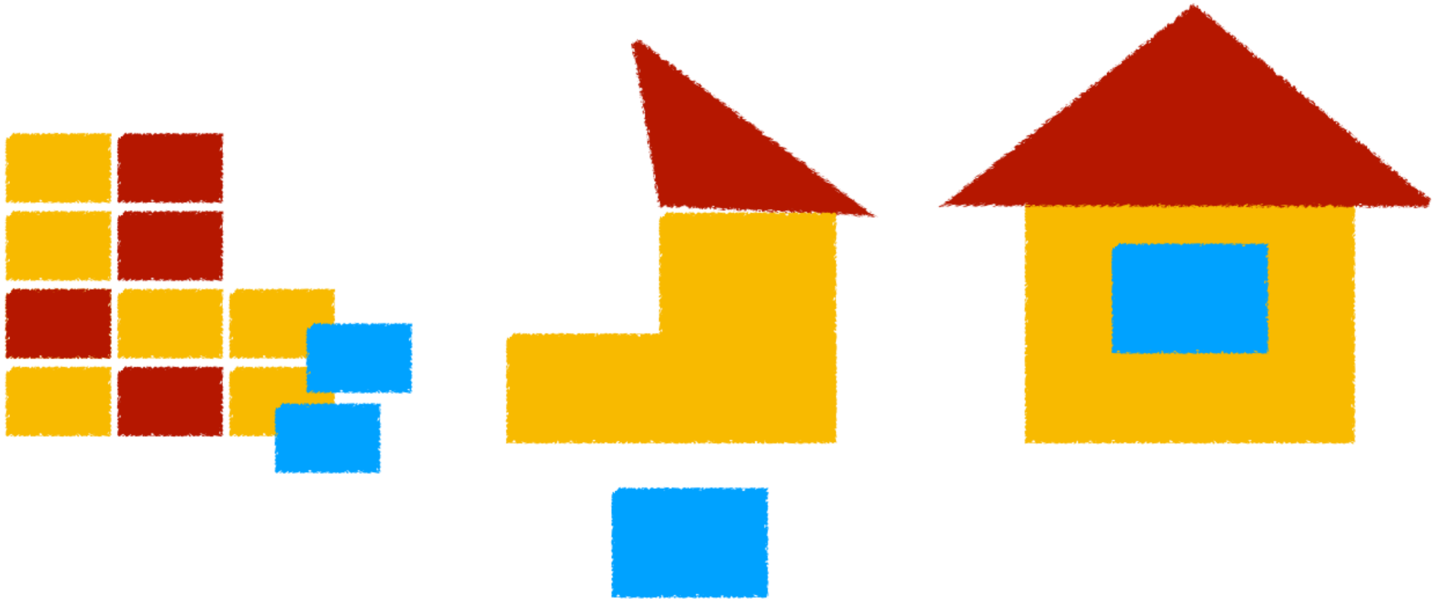
Why & When we need interpretability

- Why? Underspecification (Features are omitted)
- When
  - 1. Safety -> Interpretability helps safety
  - 2. Debugging
  - 3. Mismatched objectives and multi-objective trade-offs
  - 4. Science -> Want to have more discovery
  - 5. Legal/Ethic
  - …

How to achieve interpretability

Types of interpretable methods

Before building any model | **Building a new model** | **After building a model**



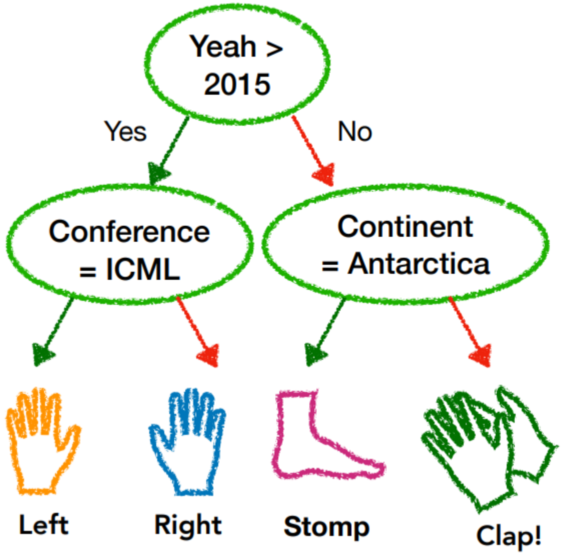
Before building the model

- Data analysis
  - Visualization
  - Exploratory data analysis

Building model

- 1. Rule-based

Building a new model: Rule-based



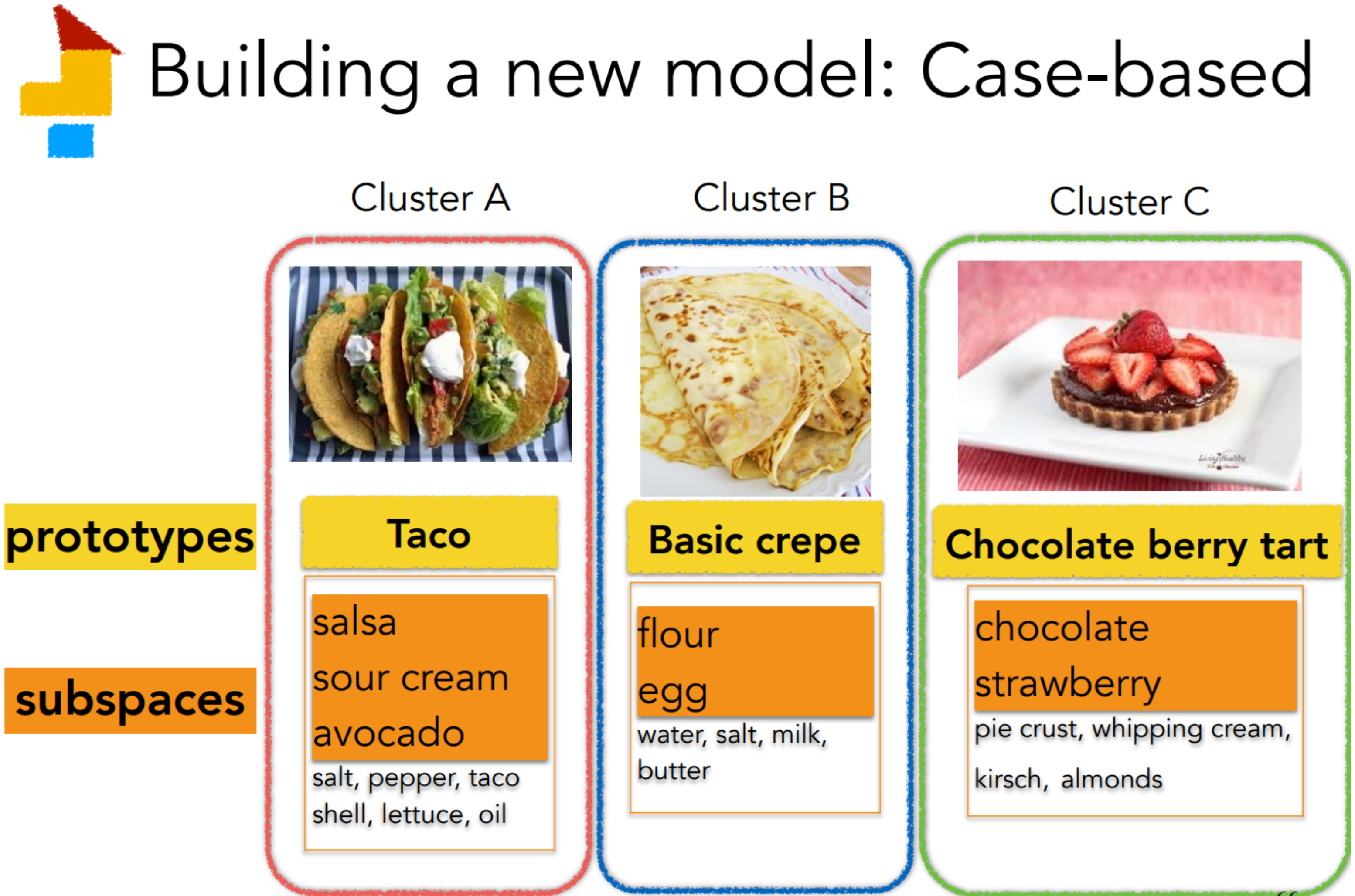
IF (sunny and hot) OR (cloudy and hot)
THEN go to beach
ELSE work

decision trees, rule lists, rule sets

[Breiman, Friedman, Stone, Olshen 84]
[Rivest 87]
[Muggleton and De Raedt 94]
[Wang and Rudin 15]
[Letham, Rudin, McCormick, Madigan '15]
[Hauser, Toubia, Evgeniou, Befurt, Dzyabura 10]
[Wang, Rudin, Doshi-Velez, Liu, Klampfl, MacNeille 17]

Building model

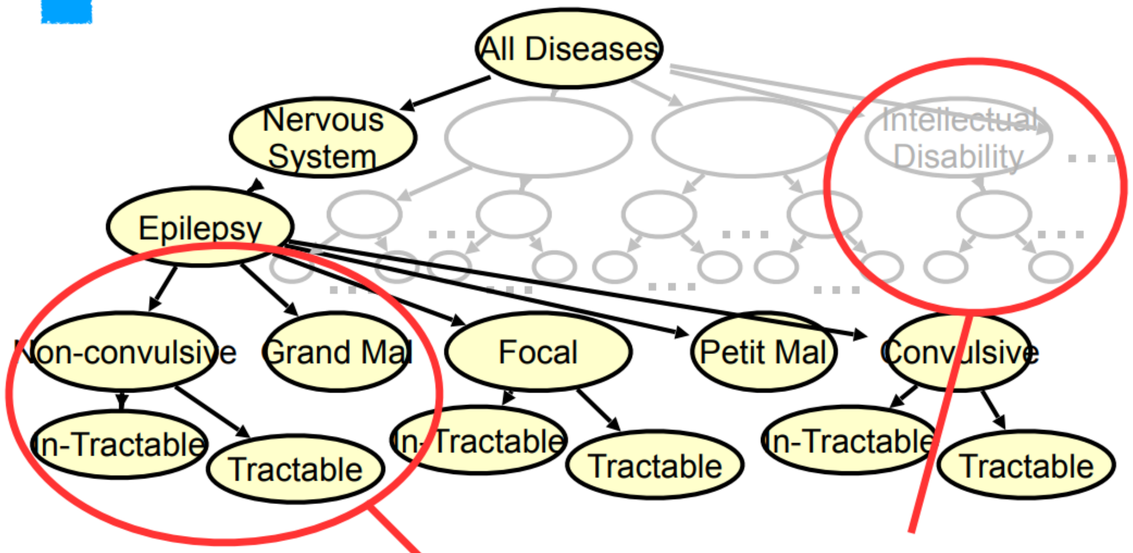
- 2. Case-basd



Building model

- 3. Sparsity based

Building a new model: Sparsity-based



Correlations across subtrees: may be a single cause manifesting in multiple aspects. Model that!

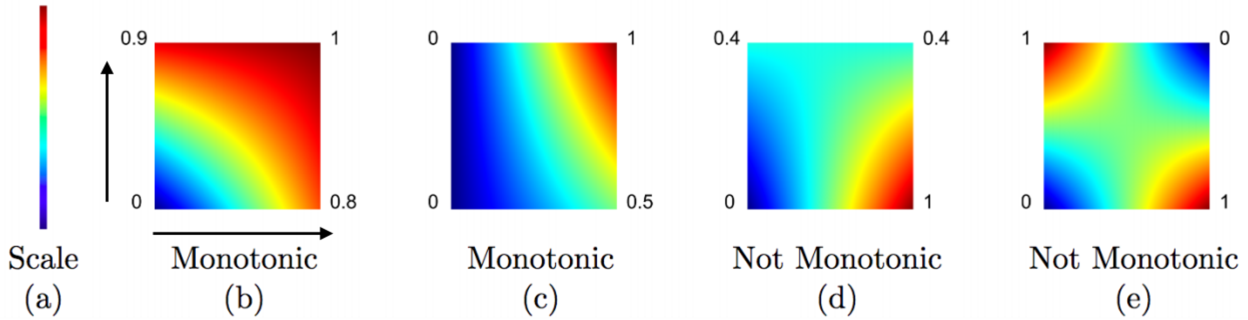
$$\Pr(\text{ data }) = \text{Mult}\Big(\underbrace{\boxed{\text{patient-subtype}}}_{\Theta_n}\ \underbrace{\boxed{\text{Subtype-concept}}}_{\Phi_k}\ \underbrace{\boxed{\text{concept-diagnosis}}}_{T_c}\Big)$$

Graph-sparse LDA [Doshi-Velez et al.'15]

Building model

- 4. Monotonicity

Building a new model: Monotonicity



- Learn piecewise monotonic function within a user specified lattice (intervals) [Gupta et al. '16]

- Monotonic neural networks by constraining weights [Neumann et al.'13, Riihimaki and Vehtari '10]

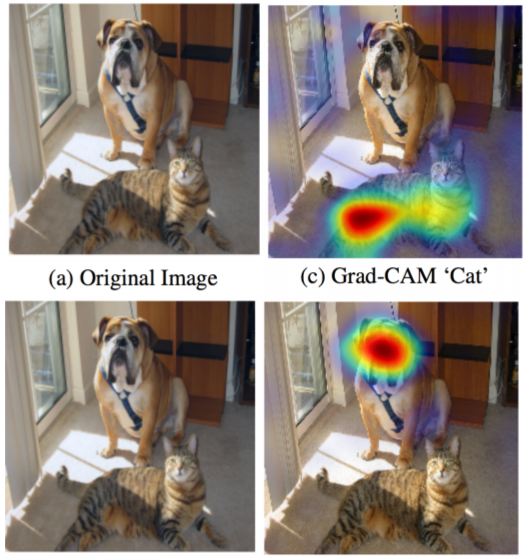
After building a model

- Analyze the result:
  - Sensitivity analysis
    - Saliency
  - mimic/surrogate models
  - Investigation on hidden layers

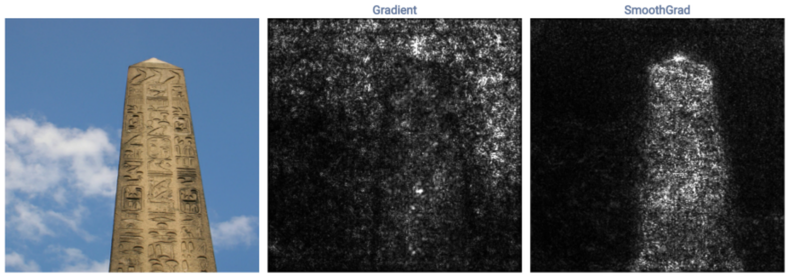
Saliency

After building a model: Saliency/attribution Maps

Grad-CAM [Selvaraju et al. 16]



SmoothGrad [Smilkov et al. 17]



Integrated gradients [Sundararajan et al. 17]

Drawback of saliency

Number of books bought

Price of a book

$$e = a \cdot p \cdot c$$

Total Profit

Currency

	2016	2017	Only this feature changed
a	4	5	(5-4)*1* 3 = 3
p	1	2	4*(2-1)*3 = 12
c	3	4	4*1*(4-3) = 4
e	12	40	19

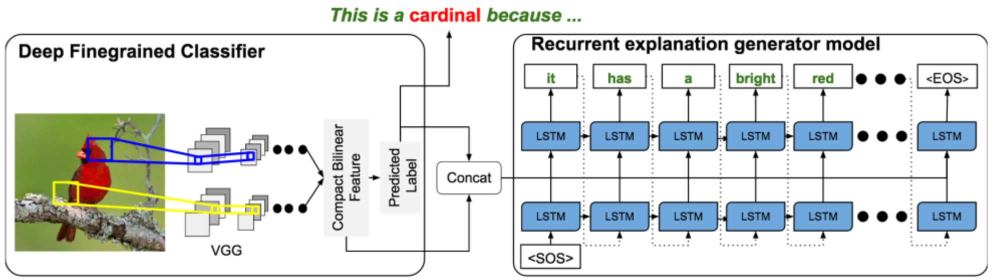
What?!

Increase in e: 28! Where is my 9?

Mimic models

After building a model: Mimic models

- Model compression or distillation [Bucila et al. ’06, Ba et al. ’14, Hinton et al. ’15]
- Visual explanations [Hendricks et al. ’16]

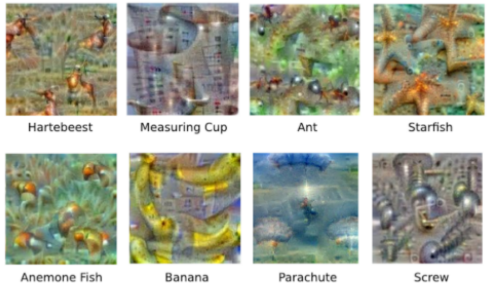


Investigation on hidden layer

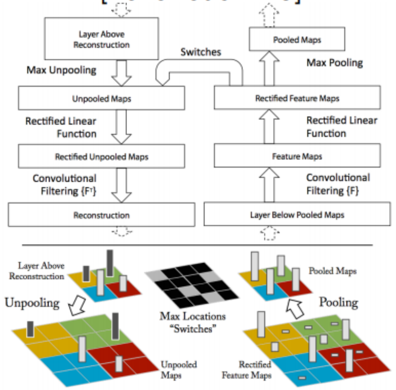
- Investigation on hidden layers
- Issues:
  - A. They may be lack of actionable insights
  - B. It is unclear if visualizing neuron vs. per layer vs. per subspaces is more meaningful than others
  - C. A golden dataset with detailed labels with human concepts are often not available

After building a model: Investigation on hidden layers

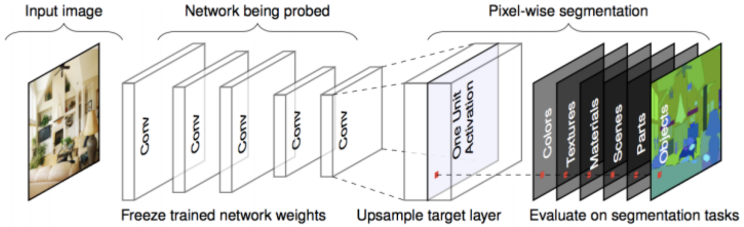
[Dosovitskiy et al. '16]



[Zeiler et al. '13]



[Bau and Zhou et al. '17]



Evaluation of interpretability

How are we measuring explanation quality now?

"You know it when you see it"

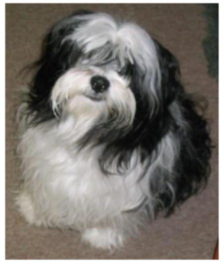
Generalized additive models (GAMs) are the gold standard for intelligibility when low-dimensional terms are considered [4, 5, 6]. Standard GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j), \qquad (1)$$

where g is the link function and for each term f_j, $E[f_j] = 0$. Generalized linear models (GLMs), such as logistic regres-

accurate, yet are highly interpretable. These predictive mod
the form of sparse *decision lists*, which consist of a series o
statements where the *if* statements define a partition of a s
and the *then* statements correspond to the predicted outcom
Because of this form, a decision list model naturally provides

Give human a task, then measure how well they do



Q. Which group does this new data point belong to?

A. Group 1

B. Group 2

[K. 16]

Evaluation

Spectrum of evaluation

	Function-based	Cognition-based	Application-based
Quantitative	How sparse are the features?	What factor should change to change the outcome?	How much did we improve patient outcomes?
Qualitative	Does it look reasonable?	What are the discriminative features?	Do scientists find the explanations useful?

Evaluation

Spectrum of evaluation

Problem-related Factors

1. Global vs. Local
2. Time budget
3. Severity of underspecification

Method-related factors

4. Cognitive chunks
5. Audience training

Cognition-based

What factor should change to change the outcome?

What are the discriminative features?

e.g.,
Humans capacity as function of factors,
Set of factors that carries over well to application

Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction

NIPS 15’ Been Kim, Doshi-Velez Finale, Julie Shah

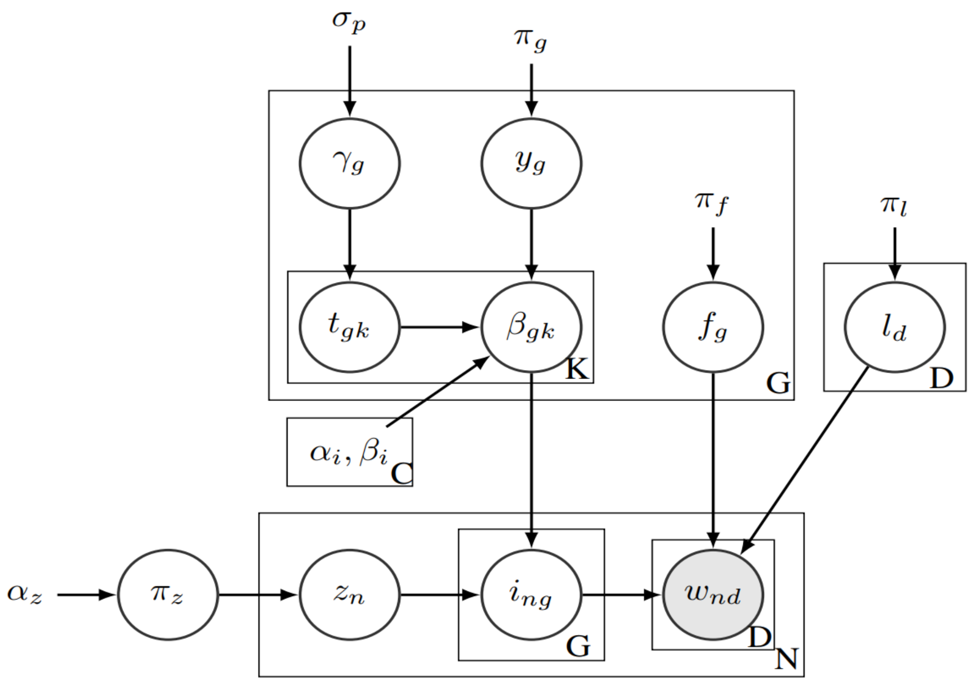
- Task: feature selection
- Mind the Gap Model: A graphical model that extracts distinguishing features with interpretability

Setting

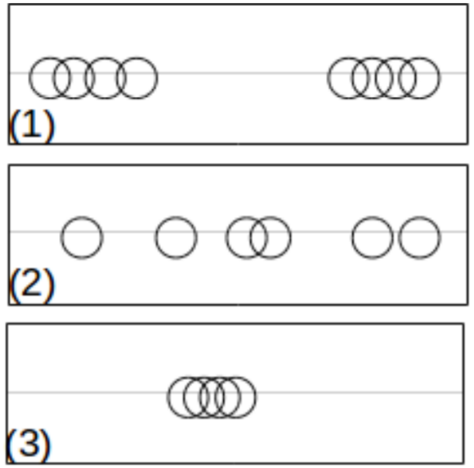
- Dataset: N observations and D binary features
- Goal: Divide the N observations into K clusters while simultaneously returning a comprehensive list of what sets of dimensions D are important for distinguishing between the clusters.

Graphical model

- g – group
- y_g - group g is selected or not selected
- l_d - the group to which dimension d belongs



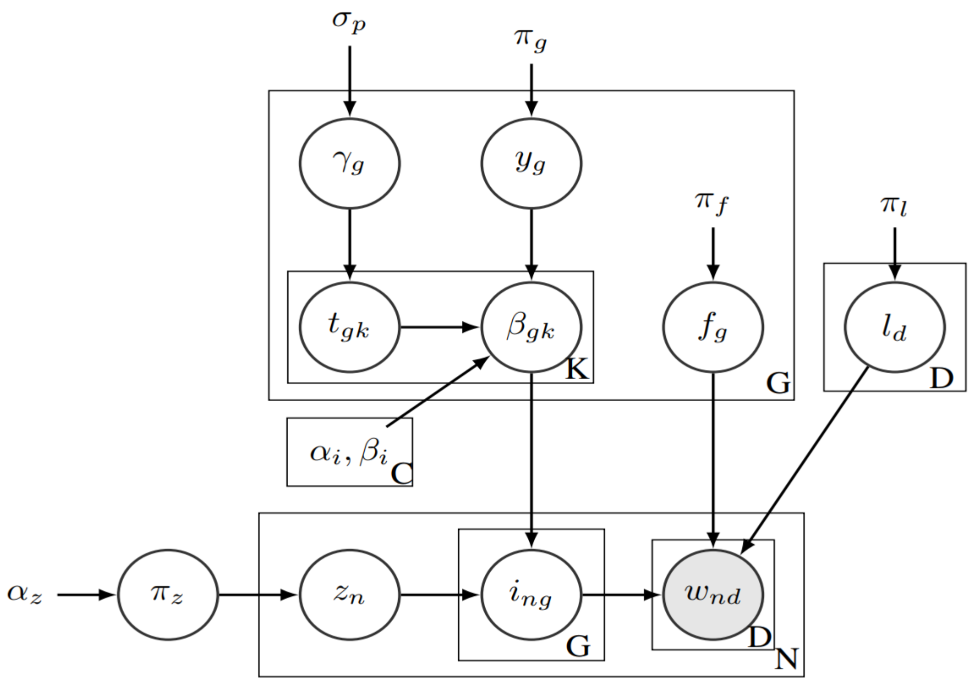
(a) Mind the gap graphical model



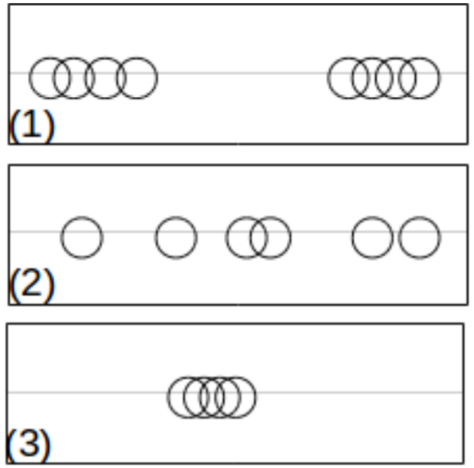
(b) Cartoon describing emissions from important dimensions. In our case, we define importance by separability—or a gap—rather than simply variance. Thus, we distinguish panel (1) from (2) and (3), while [17] distinguishes between (2) and (3).

Graphical model

- g – group
- f_{g} - or/and. Each feature only in one group
- i_{ng} - group g shown in sample n
- $w_{nd}=1$ if associated features also present in the sample



(a) Mind the gap graphical model



(b) Cartoon describing emissions from important dimensions. In our case, we define importance by separability—or a gap—rather than simply variance. Thus, we distinguish panel (1) from (2) and (3), while [17] distinguishes between (2) and (3).

Example

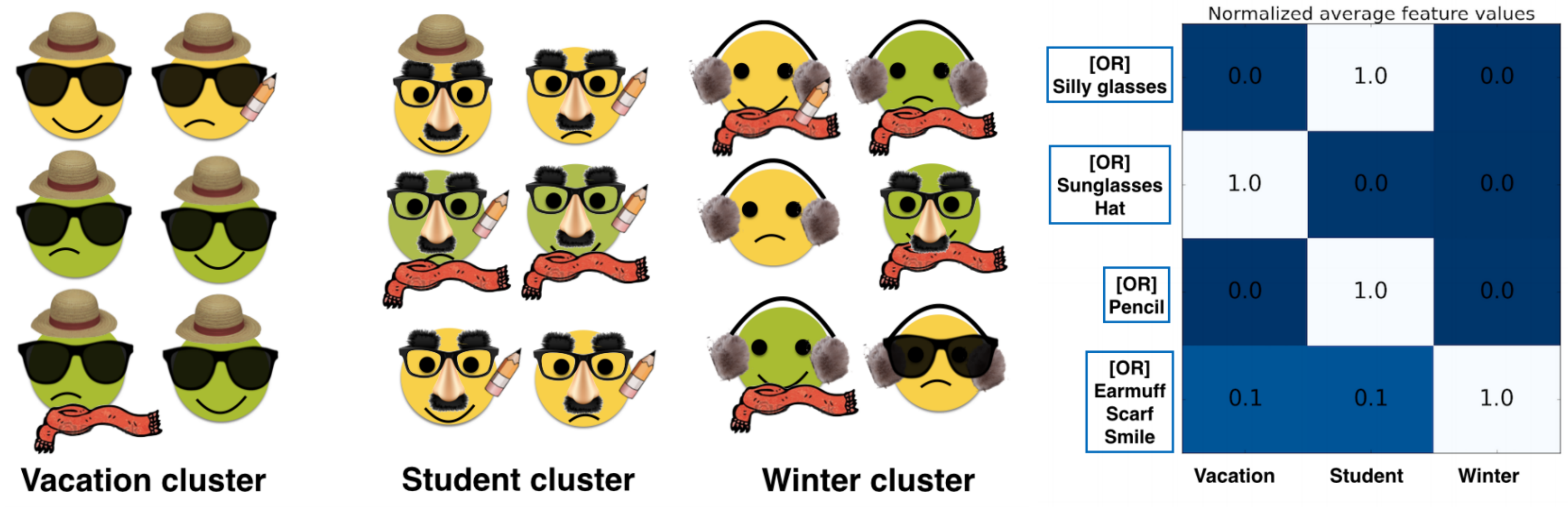


Figure 2: Motivating examples with cartoons from three clusters (vacation, student, winter) and the distinguishable dimensions discovered by the MGM.

Experiment

- Animals - 21 biological and ecological properties of 101 animals
- Recipes - 56 recipes, with 147 total ingredients
- Diseases - 184 patients with at least 200 diagnoses

Result

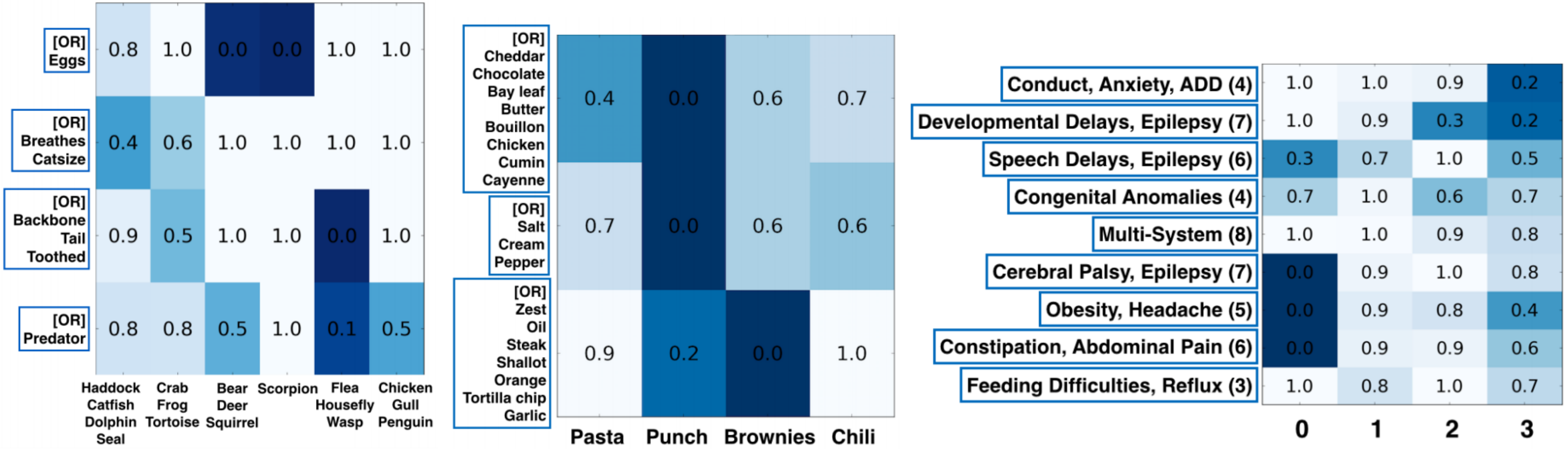


Figure 3: Results on real-world datasets: animal dataset (left), recipe dataset (middle) and disease dataset (right). Each row represents an important feature. Lighter boxes indicate that the feature is likely to be present in the cluster, while darker boxes are unlikely to be present.

How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation

Menaka Narayanan*1 , Emily Chen*1 , Jeffrey He*1 , Been Kim2 , Sam Gershman1 and Finale Doshi-Velez

- Given an input, an explanation, and an output, is the output consistent with the input and the supposed rationale?
- Study the effect of different explanations on human: For example, is a longer evaluation makes people harder to understand?
- If we understand that, it helps to generate better explanations

Definition of explanation

- In the form of ***Decision sets***:

weekend and raining → sad
spinach or chocolate → gas (which the alien hates)
sad → vegetables and candy or spices

Figure 1: Example of a decision set explanation.

- Each line contains a clause in disjunctive normal form (an or-of-ands) of the inputs, which, if true, provides a way to verify the output (also in disjunctive normal form).

Test interface

The alien's preferences:

checking the news and coughing → windy
snowing or humid and weekend → spices or vegetables and grains
embarrassed and grouchy or raining → dairy or vegetables
snowing or windy and energetic → candy or dairy and fruit
grouchy or weekend and windy → spices or grains and fruit

Is the alien happy with his meal?

Observations: Saturday, coughing, checking the news

Recommendation: bagel, rice, strawberry

Ingredients:

- **Vegetables:** okra, carrot, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

○ Yes ○ No

Submit Answer

(a) Recipe Domain

The alien's diagnosis:

frowning or upset stomach → flu season
flu season and October → hives
shrugging or hives → fast heart rate
fast heart rate and feverish or anemic or shortness of breath → vitamins and stimulants or laxatives
bleeding or anemic and fatigued → painkillers and tranquilizers or vitamins
headache or feverish and anemic → laxatives and painkillers or stimulants
high blood pressure and allergies or fast heart rate and bleeding → tranquilizers or vitamins and stimulants

Is the alien happy with his prescription?

Observations: anemic, October, frowning

Recommendation: Vipryl, Setoxin, Votasol

Disease Medications:

- **antibiotics:** Aerove, Adenon, Athoxin
- **painkillers:** Poxin, Parola, Pelapin
- **vitamins:** Vipryl, Vyorix, Votasol
- **stimulants:** Silvax, Setoxin, Soderal
- **tranquilizers:** Trasmin, Tydesol, Texopal
- **laxatives:** Lantone, Lezanto, Lexerol

○ Yes ○ No

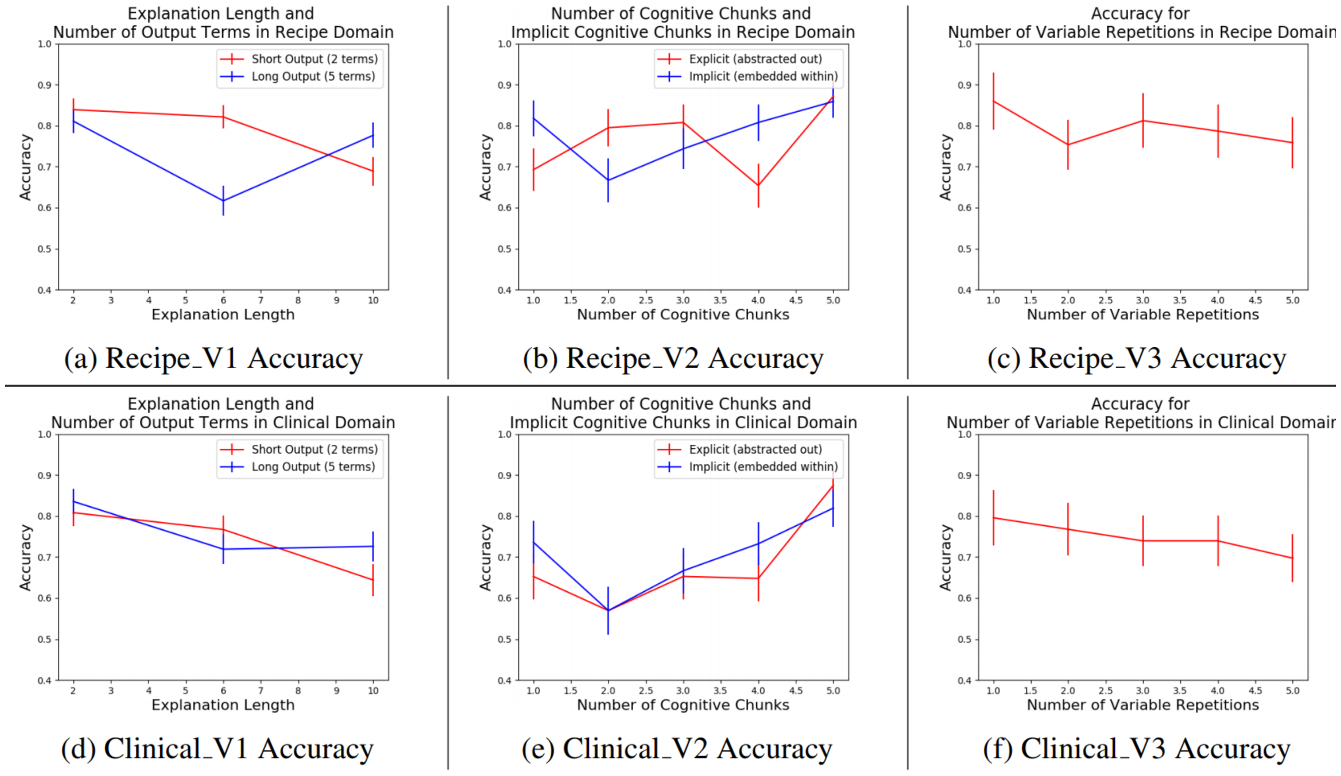
(b) Clinical Domain

Variables

- V1: Explanation Size - Number of lines of explanation
- V2: Creating New Types of Cognitive Chunks – Number of terms
- V3: Repeated terms: How many time a certain term repeated

Experiment

- A total of 600 subjects
- On 6 experiments: 3 Variables on 2 situations



(a) Recipe_V1 Accuracy (b) Recipe_V2 Accuracy (c) Recipe_V3 Accuracy

(d) Clinical_V1 Accuracy (e) Clinical_V2 Accuracy (f) Clinical_V3 Accuracy

Figure 3: Accuracy across the six experiments. Vertical lines indicate standard errors.

Experiment result

Accuracy

	Recipe		Clinical	
Factor	weight	p-value	weight	p-value
Explanation Length (V1)	-0.0116	0.00367	-0.0171	**0.000127**
Number of Output Terms (V1)	-0.0161	0.0629	0.00685	0.48
Number of Cognitive Chunks (V2)	0.0221	0.0377	0.0427	**0.00044**
Implicit Cognitive Chunks (V2)	0.0147	0.625	0.0251	0.464
Number of Variable Repetitions (V3)	-0.017	0.104	-0.0225	0.0506

Response Time

	Recipe		Clinical	
Factor	weight	p-value	weight	p-value
Explanation Length (V1)	3.77	**2.24E-34**	3.3	**5.73E-22**
Number of Output Terms (V1)	1.34	0.0399	1.68	0.0198
Number of Cognitive Chunks (V2)	8.44	**7.01E-18**	4.6	**1.71E-05**
Implicit Cognitive Chunks (V2)	-15.3	**2.74E-08**	-11.8	**0.000149**
Number of Variable Repetitions (V3)	2.4	**0.000659**	2.13	0.0208

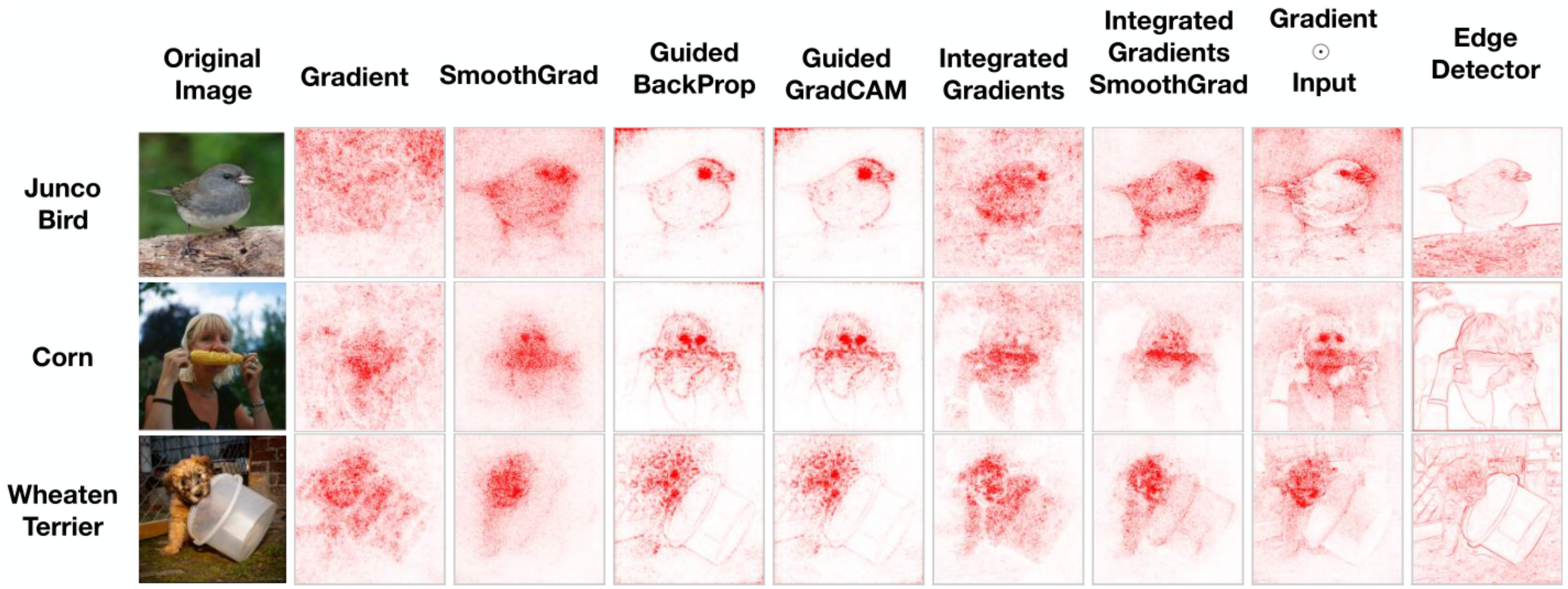
Subjective Evaluation

	Recipe		Clinical	
Factor	weight	p-value	weight	p-value
Explanation Length (V1)	-0.165	**5.86E-16**	-0.186	**1.28E-19**
Number of Output Terms (V1)	-0.187	**2.12E-05**	-0.0335	0.444
Number of Cognitive Chunks (V2)	-0.208	**1.93E-05**	-0.0208	0.703
Implicit Cognitive Chunks (V2)	0.297	0.0303	0.365	0.018
Number of Variable Repetitions (V3)	-0.179	**5.71E-05**	-0.149	**0.000771**

Sanity Checks for Saliency Maps

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim

- An assessment of different explanation methods(Based on gradient)



Methods

- Gradient $\odot$ Input: Elemental wise product
- Integrated Gradients (IG): $E_{\mathrm{IG}}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x}+\alpha(x-\bar{x}))}{\partial x} d\alpha,$
- Guided Backpropagation (GBP) - negative gradient entries are set to zero while back-propagating through a ReLU unit.
- Guided GradCAM: Based on gradient to the feature map of the last convolutional unit
- SmoothGrad (SG): Smooth the noise from saliency map

$$E_{\mathrm{sg}}(x) = \frac{1}{N} \sum_{i=1}^{N} E(x + g_i)$$

Test 1: Model randomization: Cascading Randomization

- randomize the weights of a model starting from the top layer to bottom

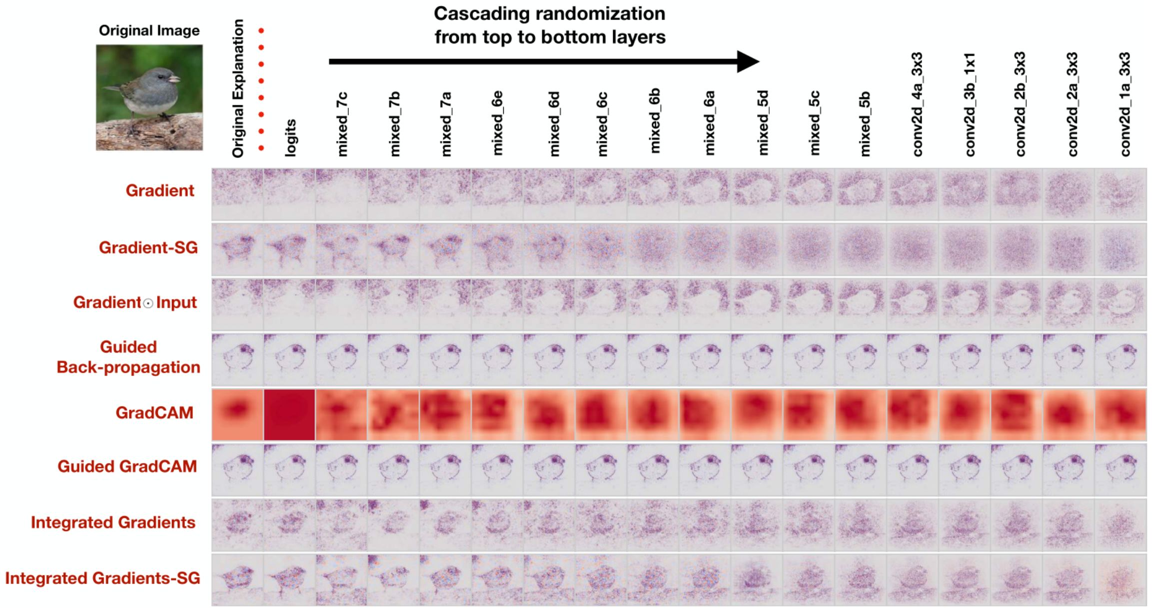
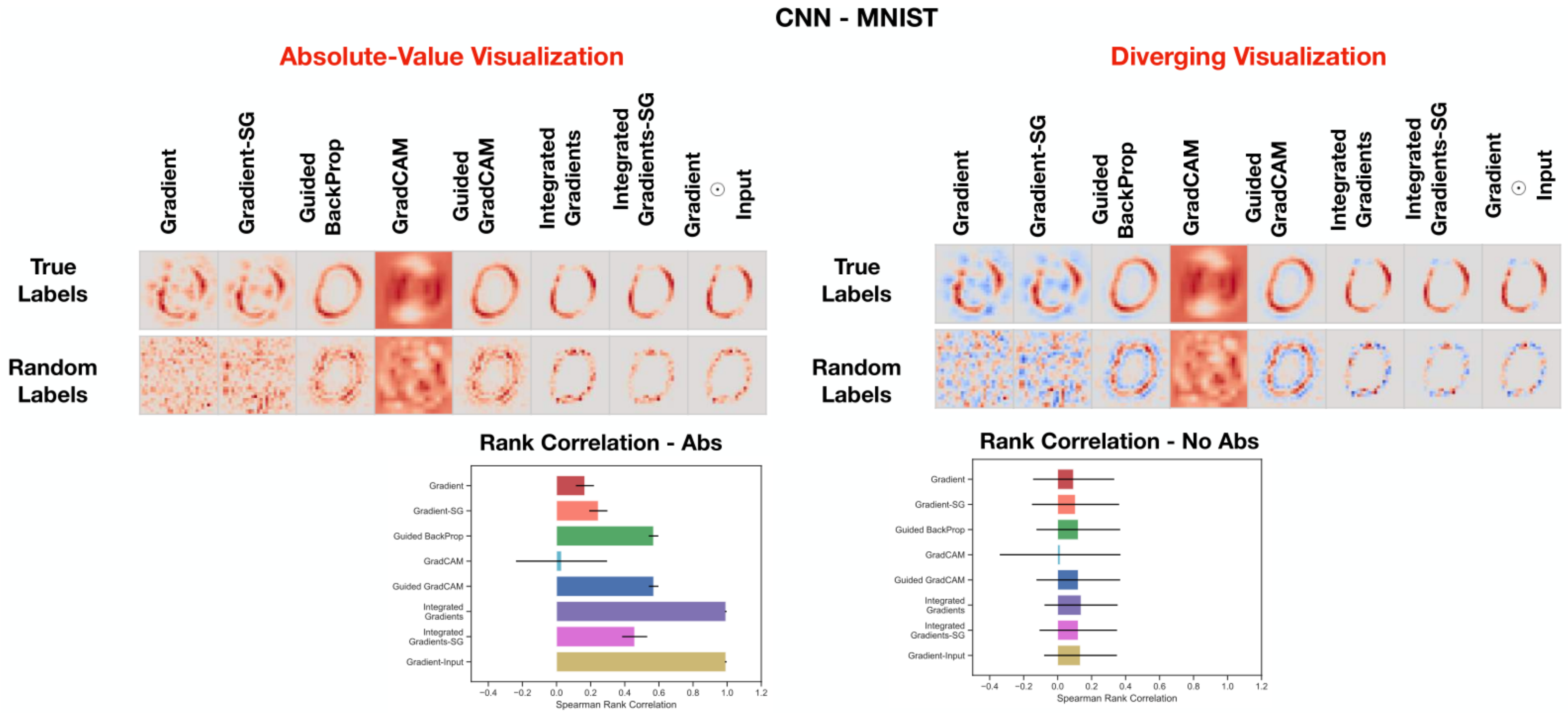


Figure 2: **Cascading randomization on Inception v3 (ImageNet).** Figure shows the original explanations (first column) for the Junco bird. Progression from left to right indicates complete randomization of network weights (and other trainable variables) up to that 'block' inclusive. We show images for 17 blocks of randomization. Coordinate (Gradient, mixed_7b) shows the gradient explanation for the network in which the top layers starting from Logits up to mixed_7b have been reinitialized. The last column corresponds to a network with completely reinitialized weights.

Task 2: Data randomization



Summary

- Some existing saliency methods are ***independent*** both of the model and of the data generating process
- Such methods are unreasonable, because it doesn’t correctly reflect the quality of the model and the method.