O'REILLY®

Artificial Intelligence

DATASCIENCE.COM

Model evaluation in the land of deep learning

May 2, 2018

Host

Pramit Choudhary

@MaverickPramit

www.linkedin.com/in/pramitc/

pramit@datascience.com

DATASCIENCE.COM
SKATER

- Lead data scientist at DataScience.com.
- Currently, exploring better ways to extract, evaluate, and explain the learned decision policies of predictive models. Recently started an open source project, Skater, to improve the process of model interpretation to enable better model evaluation and model security.
- Before joining DataScience.com, used machine learning algorithms to find love for eHarmony customers.

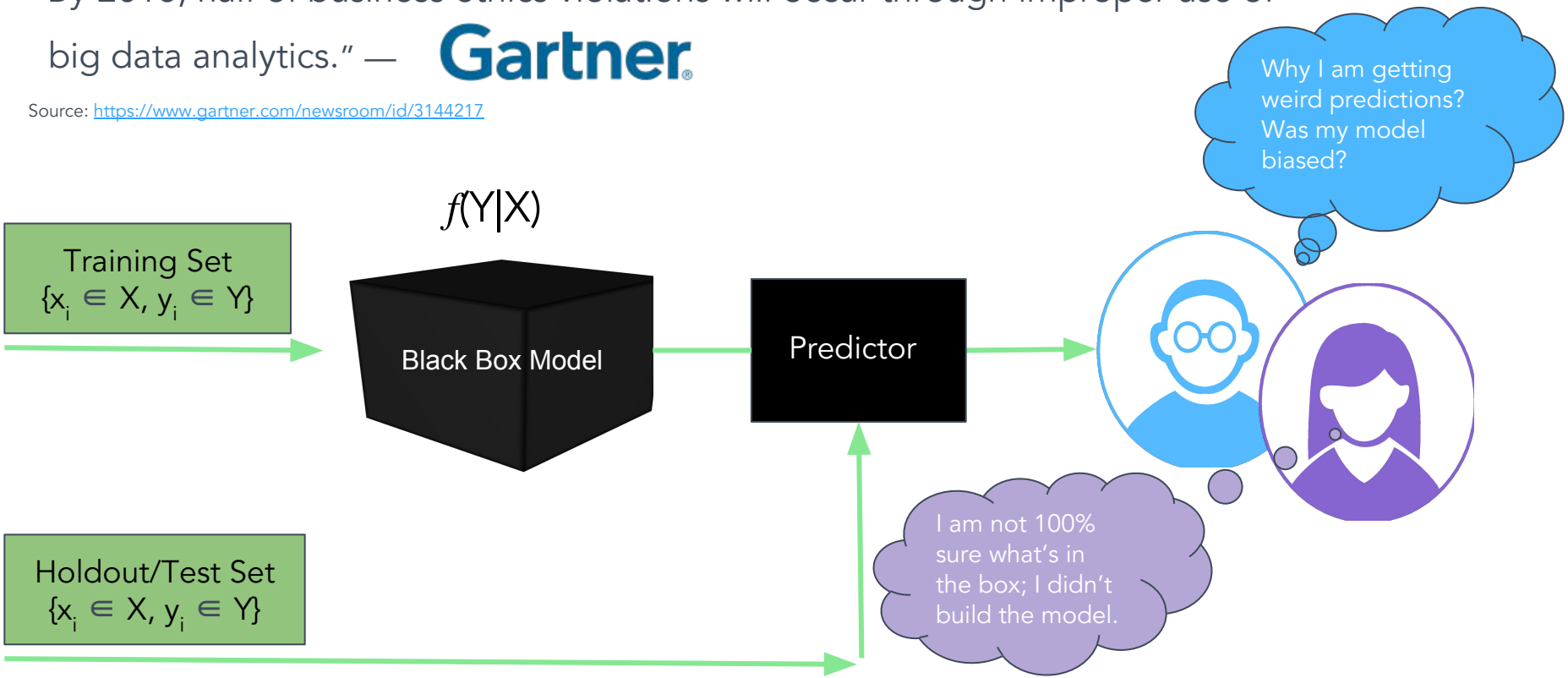
Agenda

- Understand the problem of model opacity
- Define the “what” and “why” of model interpretation
- Define the scope of model interpretation
- Introduce Skater
- How to interpreting Deep Neural Networks(DNNs) model to improve Model Performance with Skater?
- Demo
- Ability to generate and prevent adversarial attacks with Skater
- Q&A
- References

The Problem of Model Opacity

"By 2018, half of business ethics violations will occur through improper use of big data analytics." — Gartner

Source: https://www.gartner.com/newsroom/id/3144217



QUOTE

A collection of visual and/or interactive artifacts that provide a user with sufficient description of the model behavior to accurately perform tasks like evaluation, trusting, predicting, or improving the model.

Assistant Professor Sameer Singh,
University of California, Irvine

What is Model Interpretation?

- Ability to explain and present a model in a way that is **understandable to humans.**

- A Model's result is **self descriptive** and **needs no further explanation**; expressed in terms of input and output.

Why is Model Interpretation Important?

Producer / Model Maker:

- Data scientist/analyst building a model
- Consultants helping clients

Consumer / Model Breaker:

- Business owners or data engineers
- Risk/security assessment managers
- Humans being affected by the model


Ideas collapse.

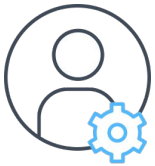
QUOTE

While model interpretation is a hard problem, it's within the role of the data scientist to guide the other stakeholders through different levels of interpretation, recognize the caveats, highlight ambiguities, etc

Paco Nathan,
Director Learning Group, O'Reilly

Motives for Model Interpretation

Producer/Model Maker



- Data Scientist
- Machine Learning Engineer
- Data Analyst
- Statistician

1. **Debugging and improving** an ML system.
2. **Exploring and discovering latent or hidden feature interactions** (useful for feature engineering/selection and resolving preconceptions).
3. Understanding **model variability.**
4. Helps in **model comparison.**
5. Building **domain knowledge** about a particular use case.
6. Bring **transparency** to decision making to enable **trust.**

Consumer/Model Breaker

- Data Science Manager
- Business Owner
- Data Engineer
- Auditors/Risk Managers

1. Explain **the model/algorithm.**
2. Explain **the key features driving the KPI.**
3. **Verify and validate the accountability** of ML learning systems, e.g., causes for false positives in credit scoring, insurance claim frauds.
4. Identify **blind spots to** prevent adversarial attacks or fix dataset errors.
5. **Ability to share** the explanations to consumers of the predictive model.
6. Comply with **data protection regulations**, e.g., EU's GDPR.

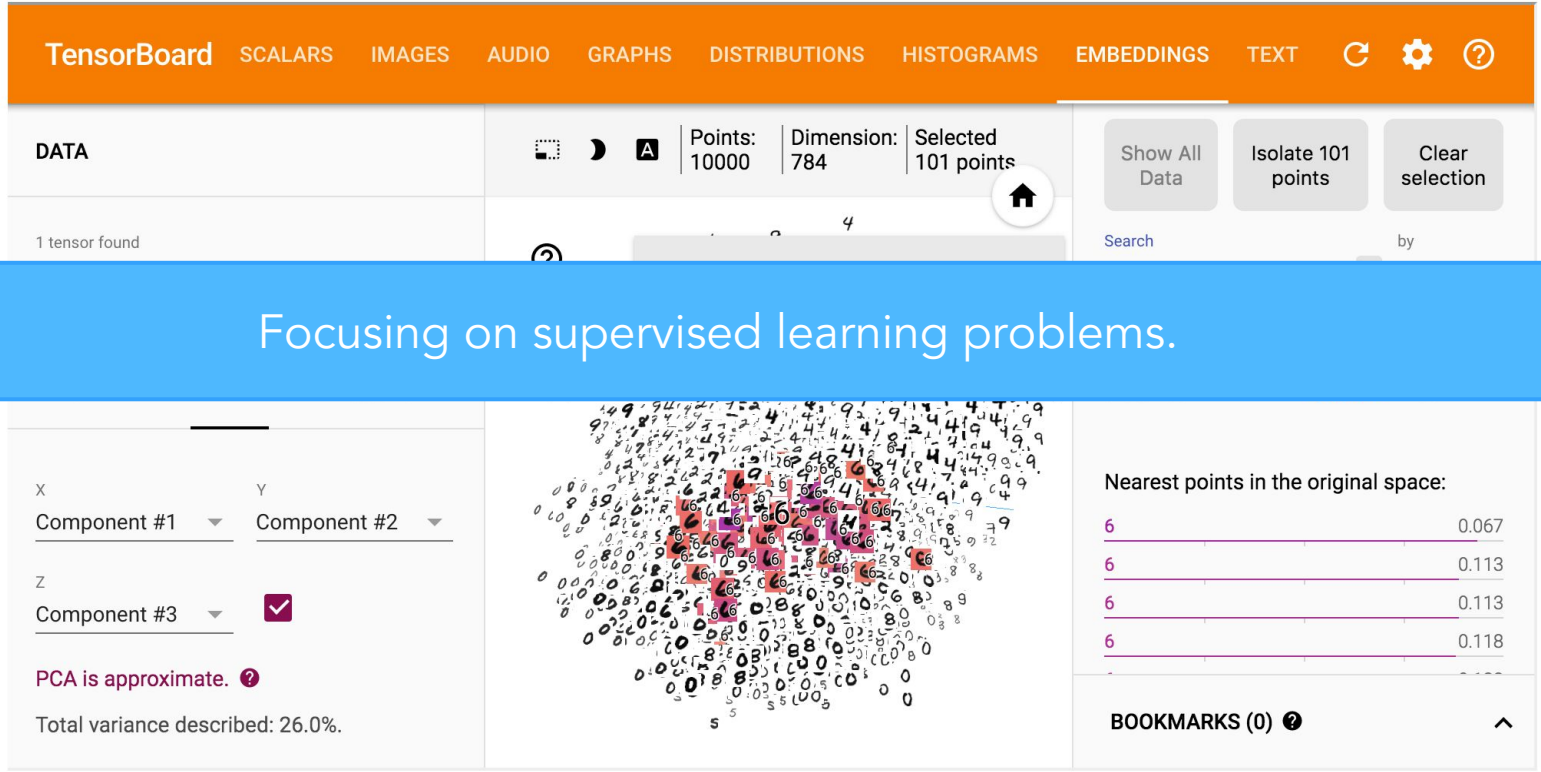
What Do We Want to Achieve?

With model interpretation, we want to answer the following questions:

- **Why** did the model behave in a certain way?
- **What** was the reason for false positives? What are the relevant variables driving a model's outcome, e.g., customer lifetime value, fraud detection, image classification, spam detection?
- **How** can we trust the predictions of a "black box" model? Is the predictive model biased? **How** can we guarantee model's security against adversarial attacks?

In [42]:
```
import IPython
url = 'http://172.31.0.19:6006/'
iframe = '<iframe src=' + url + ' width=1000 height=500></iframe>'
IPython.display.HTML(iframe)
```

Out[42]:



Focusing on supervised learning problems.

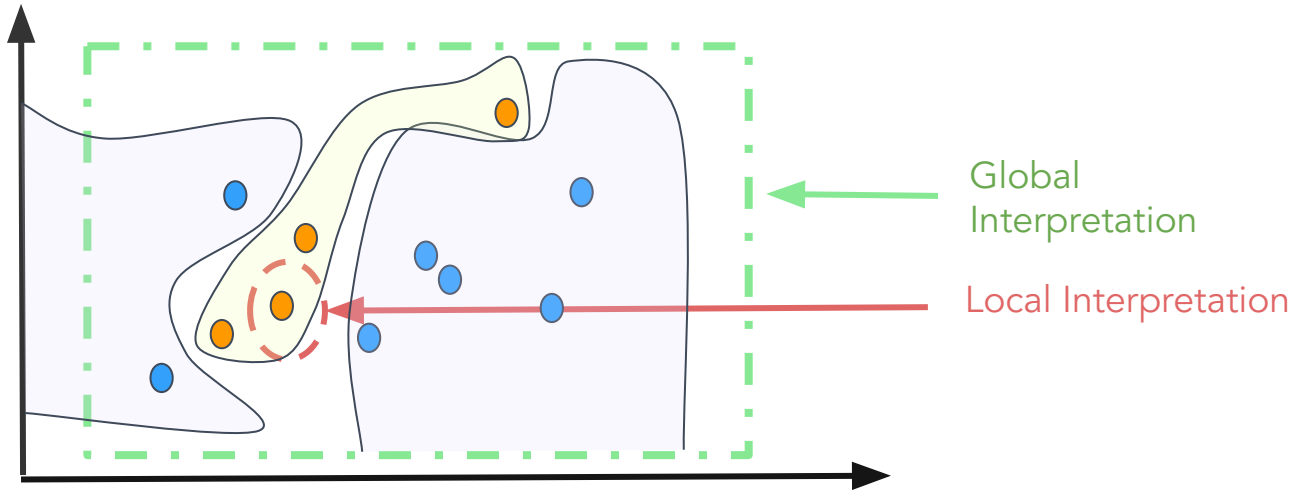
Scope of Interpretation

Global Interpretation

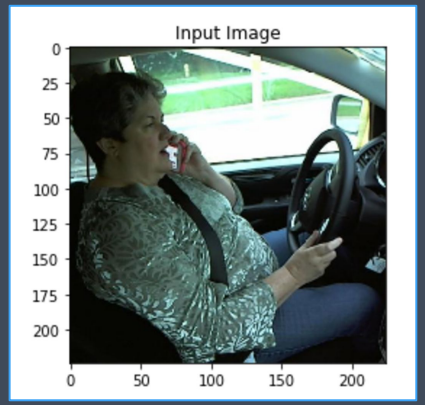
Being able to explain the conditional interaction between dependent (response) variables and independent (predictor or explanatory) variables based on the complete dataset

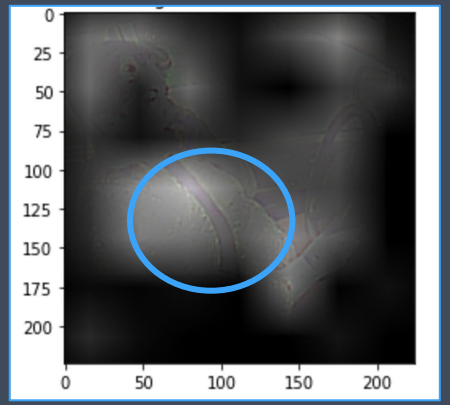
Local Interpretation

Being able to explain the conditional interaction between dependent (response) variables and independent (predictor or explanatory) variables with respect to a single prediction

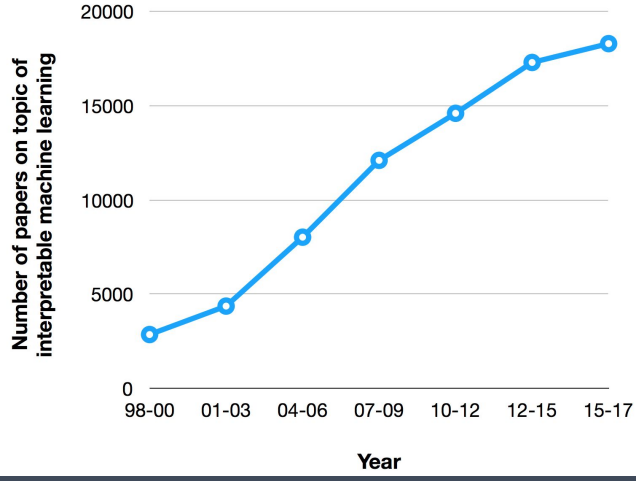


How Do We Enable Model Interpretation?





ML community is responding



Visual question answering (VQA): Is this a case of distracted driving?

Relevance: Insurance fraud

Top predictions include p(seat-belt)=0.75, p(limousine)=0.051, and p(golf cart) = 0.017

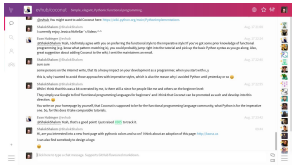
Reference: Kim, Been & Doshi-Velez, Finale. (ICML 2017). Interpretable Machine Learning:
The fuss, the concrete and the questions
(http://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf)

Introducing Skater

Unstar 467 | Fork 52

If you like the idea, follow our journey!

GitHub https://github.com/datascienceinc/Skater



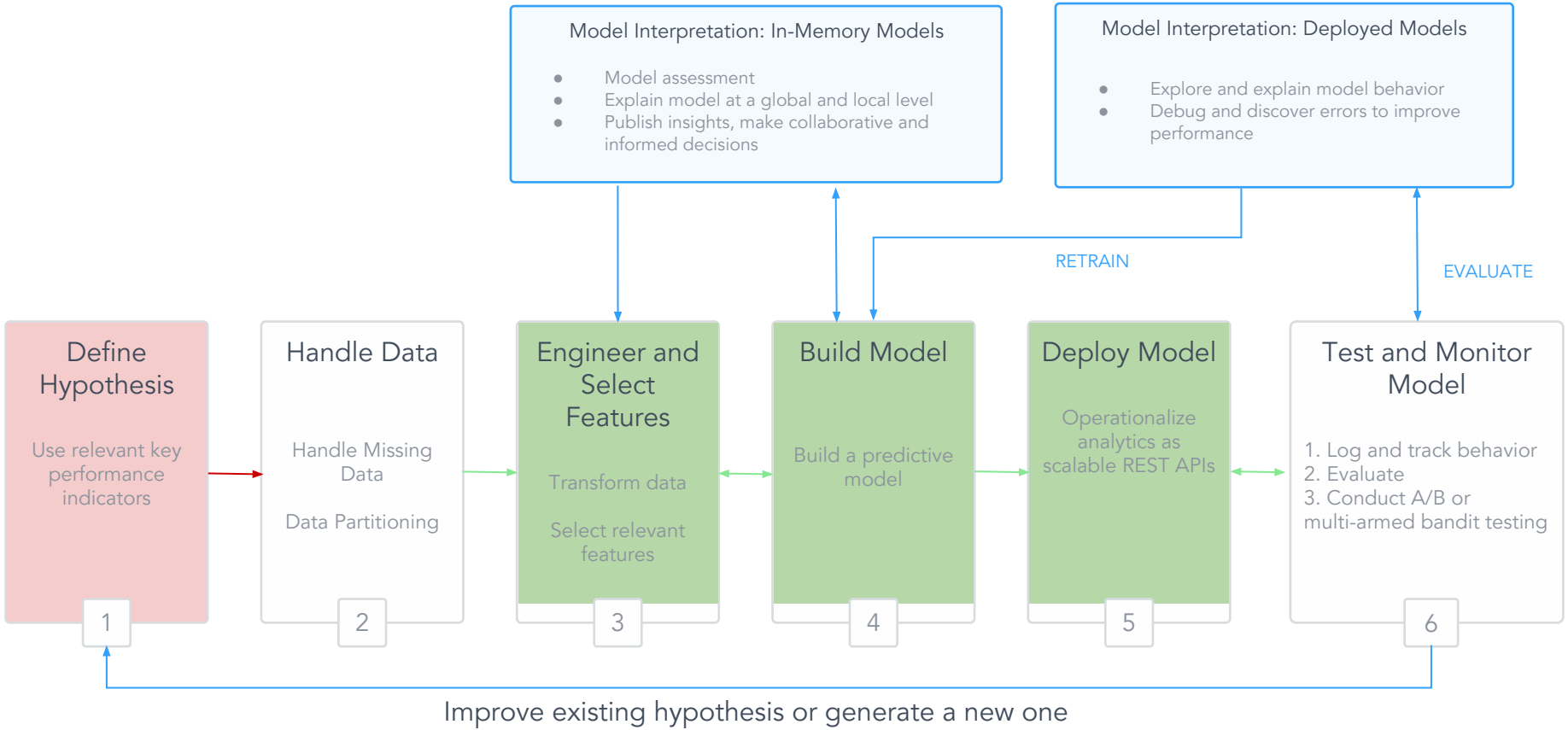
Gitter Channel (join us here):
https://gitter.im/datascienceinc-skater/Lobby

DATASCIENCE.COM

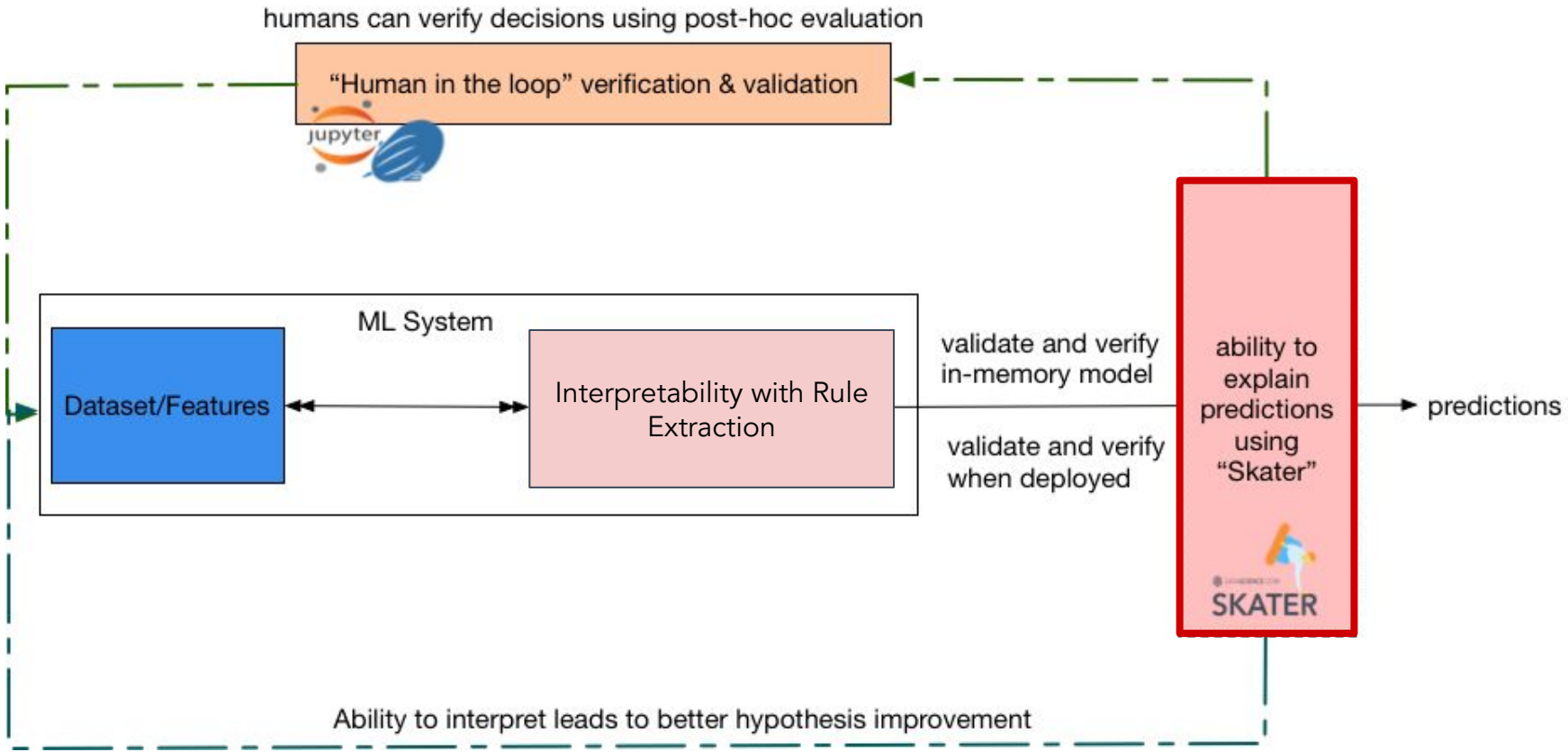
SKATER

A unified framework to enable Model Interpretation both globally (on the basis of a **complete data set**) and locally (in-regards to **an individual prediction**)

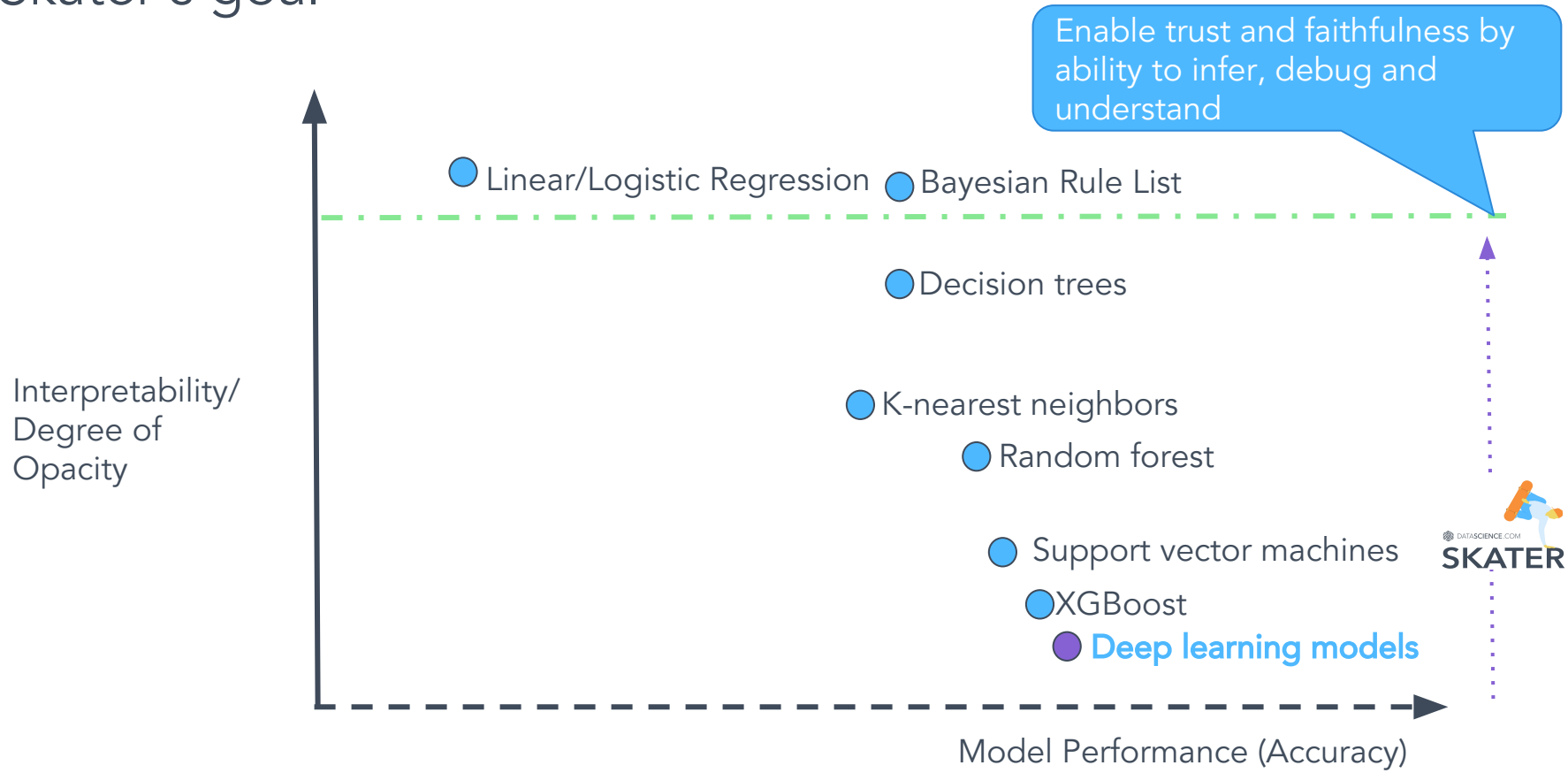
Machine Learning Workflow



An Interpretable Machine Learning System

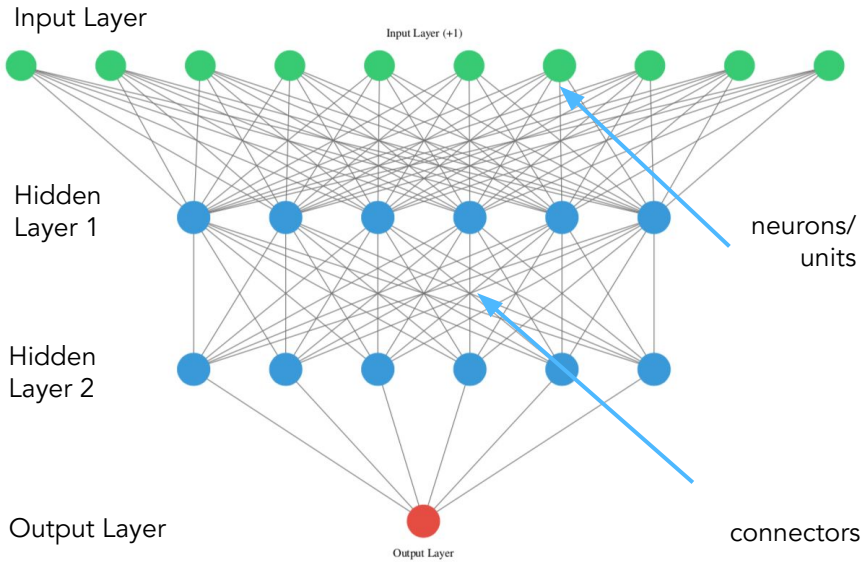


Skater's goal



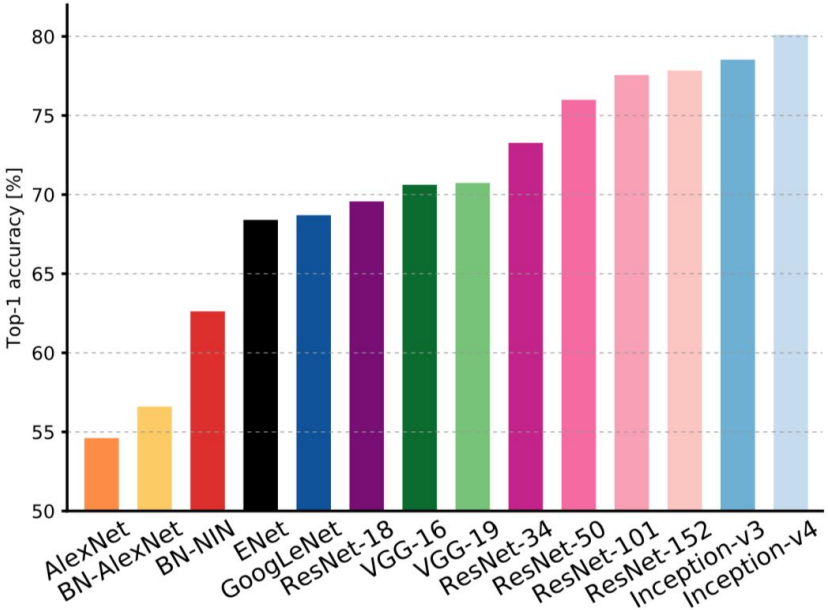
Note: The purpose of the chart is not to mirror any benchmark on model performance, but to articulate the opacity of predictive models

Deep Neural Networks



- Helps build expressive and flexible models by learning arbitrary non-linear and non-convex functions easily
- Can be expressed in different architectures; optimizing for accuracy and computation efficiency often leads to complex designs
- Need for manual feature engineering is less; lower layers can extract complex features
- With advancement in software - Keras/Tensorflow/MXNet and hardware-better integration with GPU, it's easier to train DNN's over billions of data points optimizing over large number of parameters.
- But, models are often perceived as black boxes because of lack of tools to infer them

Modern DNNs with complex designs



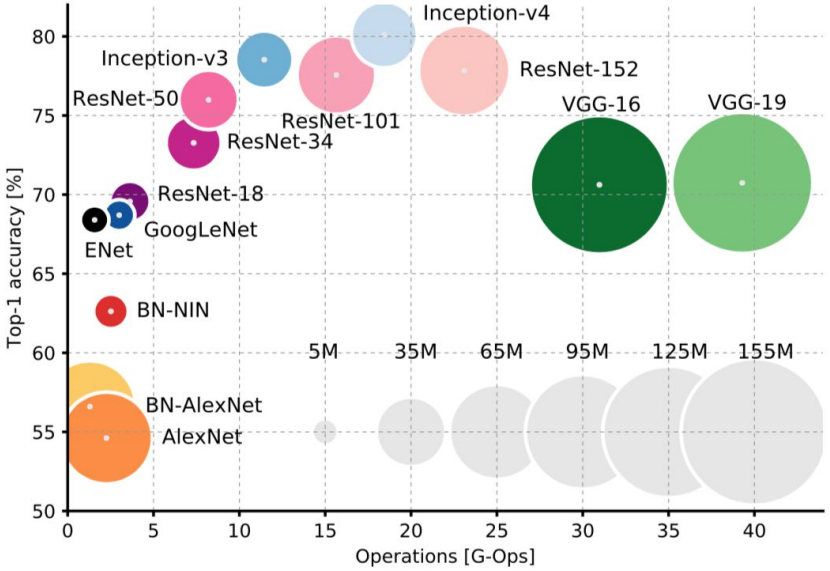


Image Source: Canziani, Alfredo, Paszke, Adam & Culurciello, Eugenio (2016). An Analysis of Deep Neural Network Models for Practical Applications

Optimizing on accuracy, has made Modern DNN architectures complex and often difficult to interpret and understand as humans

Layer-wise Relevance Propagation (LRP)

- Decomposes the predictions of a deep neural network to pixel level relevance scores using first order approximation
- Initially proposed by Bach S. et. al (2015) https://doi.org/10.1371/journal.pone.0130140
- Computed as a backward pass using a modified gradient from the output layer to the input layer
- Skater supports e-LRP (a version of LRP) as proposed by Ancona M., Ceolini E., Cengiz Ö., Gross M. (2018) in Towards better understanding of gradient-based attribution methods for Deep Neural Networks using chain rule with a modified gradient

$$r_i^{(l)} = \sum_j \frac{z_{ji}}{\sum_{i'}(z_{ji'} + b_j) + \epsilon \cdot sign(\sum_{i'}(z_{ji'} + b_j))} r_j^{(l+1)}$$

$$x_i * \frac{\partial^g S_c}{\partial x_i}, \quad g = \frac{f(z)}{z}$$

- **Scope of Interpretation**: Local Interpretation
- Framework supported by Skater:

Integrated Gradient

- Computes relevance score for Deep Networks for Image and Text using first order approximation
- Proposed by *Sundararajan, Mukund, Taly, Ankur & Yan, Qibi (2017)* in Axiomatic Attribution for Deep Networks
- Implementation adopted as suggested by Ancona M., Ceolini E., Cengiz Ö., Gross M. (2018) in Towards better understanding of gradient-based attribution methods for Deep Neural Networks
- Determines relevance (contribution) of an input $X = \{x_1, x_2, \ldots x_n\} \in R^n$ relative to baseline input X'
- Compute the average gradient while the input varies along a linear path from a baseline x' to x

$$IG(x) = (x_i - x_i') * \sum_{k=1}^{m} \frac{\partial}{\partial x_i} F(x' + \frac{k}{m} * (x - x')) * \frac{1}{m}$$

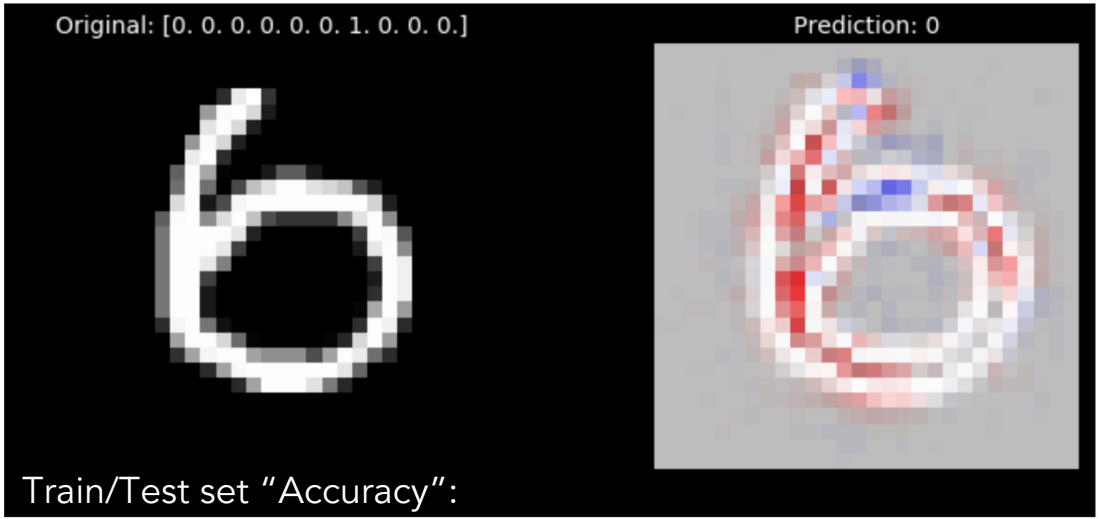
- Baseline x': for Image: [redacted]; for Text: zero embedding vector
- Statisfies sensitivity and implementation Invariance
- **Scope of Interpretation**: Local Interpretation
- Framework supported by **Skater**:

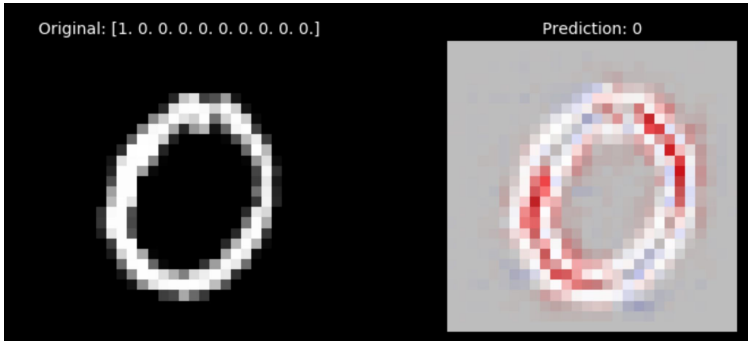
Demo

Evaluating Deep Neural Networks with Skater

1. CNN on MNIST dataset
2. Imagenet with pre-trained Inception-V3
3. CNN/LSTM sentiment analysis with IMDB dataset

Evaluate Model Stability





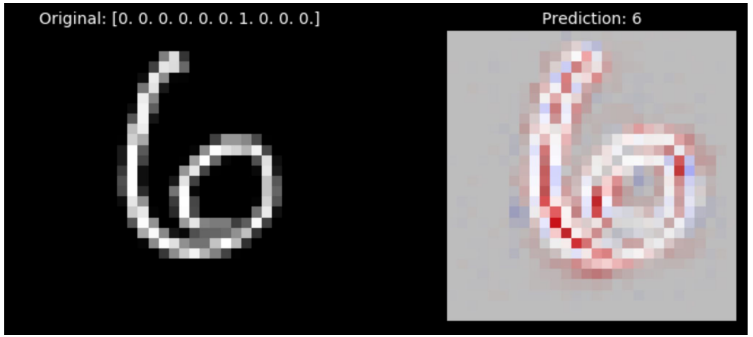
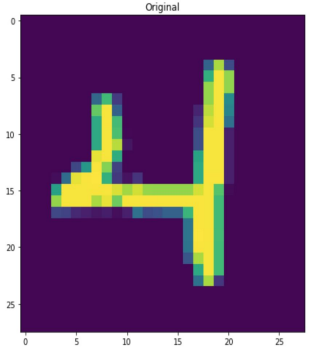
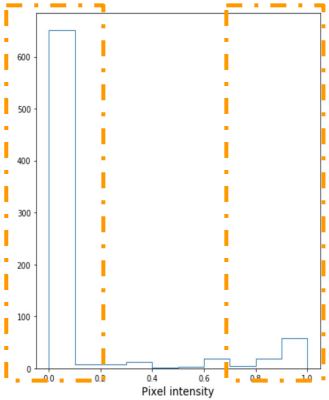


Figure: An MNIST experiment with CNN model with 98.8% train and 98.6 test ‘Accuracy’. Interpretation CNN model with ‘ReLU activation using e-LRP. Image-6 on the left is in-correctly classified as 0. Skater provides the ability to infer the cause of mis-classification (Pixels colored in Red have a positive influence and Blue negative influence). Images share a semantic properties globally. In the above example we can see 6 and 0 sharing semantic properties around the lower curvy round ‘O’, probably the reason for misclassification.

Identifying blind spots

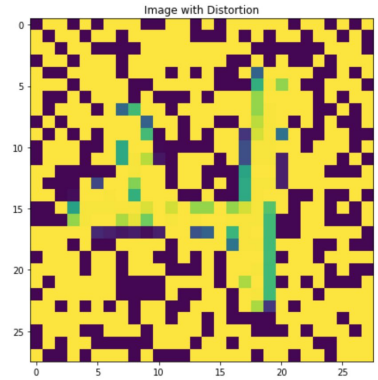
$X \in R^m$, where R is a image vector, mapped to a discrete label set, $L \in \{1 \ldots k\}$. Interpreting CNN model on MNIST dataset.

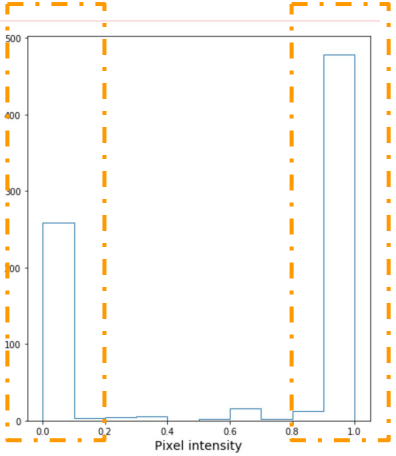




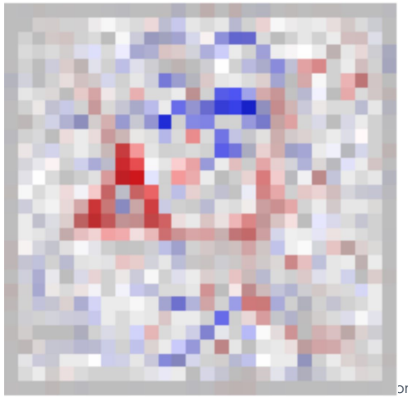
Distortion(D)

$X + r \in [0, 1]^m$

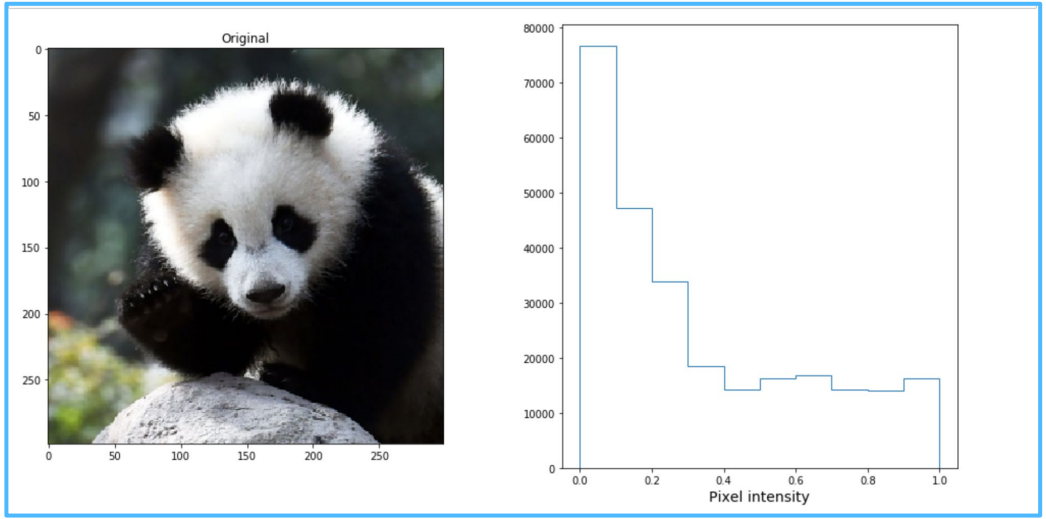




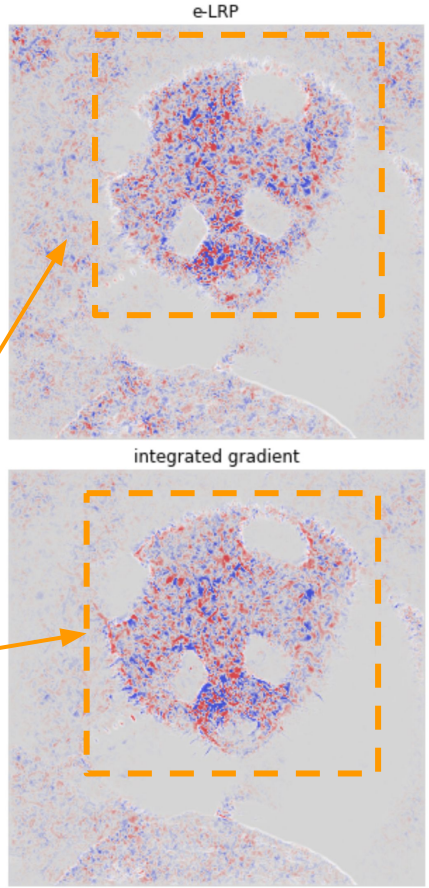
Relevant Image pixels are retained and correctly identified



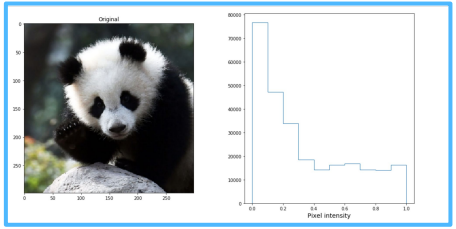
Creating Adversarial Examples



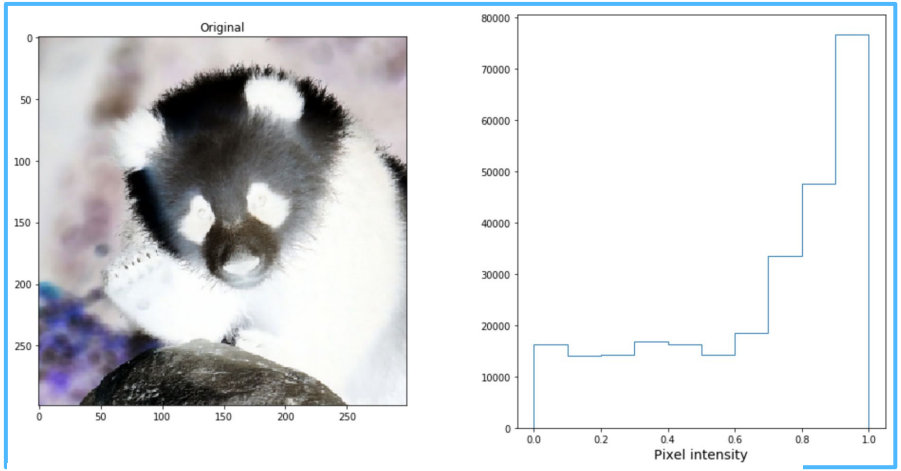
	Class Labels	Scores
0	giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca	0.934644
1	soccer ball	0.001220
2	space shuttle	0.000602



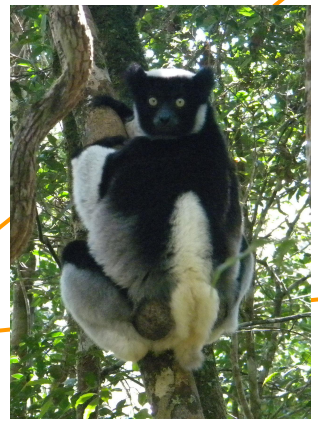
Conditional adversarial tested against pre-trained Inception-V3



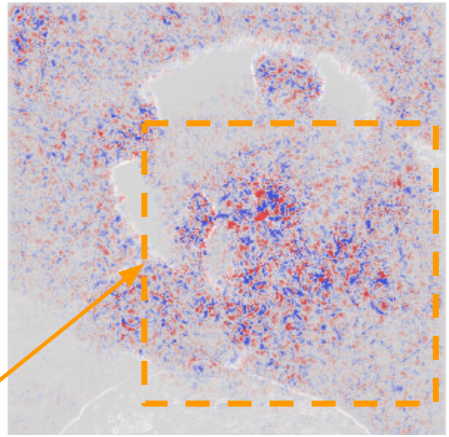
	Class Labels	Scores
0	giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca	0.934644
1	soccer ball	0.001220
2	space shuttle	0.000602



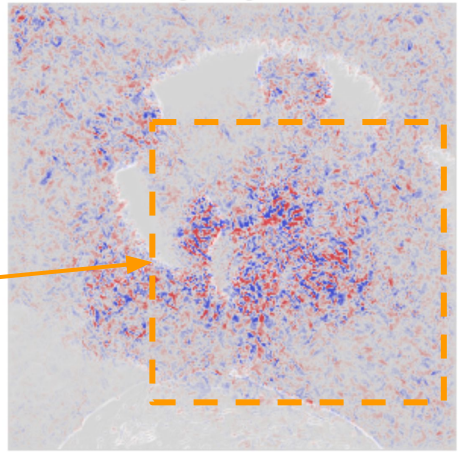
	Class Labels	Scores
0	indri, indris, Indri indri, Indri brevicaudatus	0.266485
1	guenon, guenon monkey	0.037744
2	colobus, colobus monkey	0.035104



elrp



integrated gradient

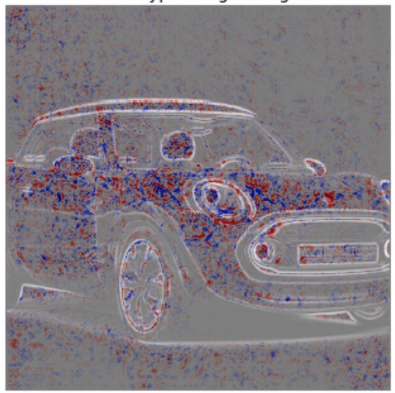


More Examples

Input Image

sports car: 0.54%

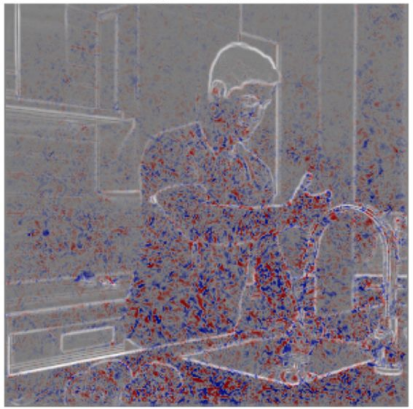
Relevance Type integrated gradient



Input Image

washbasin: 0.43%

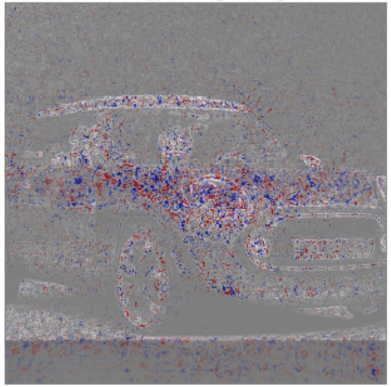
Relevance Type: integrated gradient



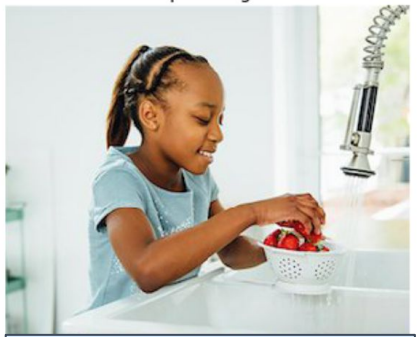
Input Image

grille: 0.21%

Relevance Type integrated gradient

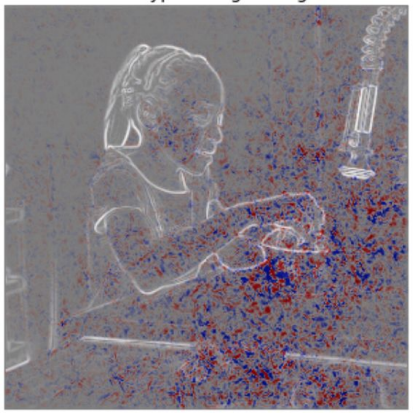


Input Image



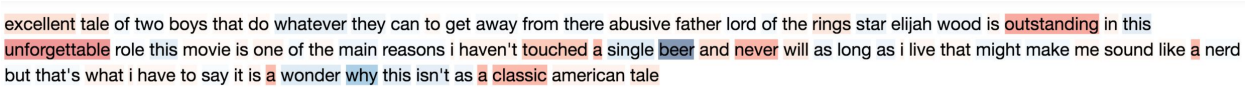
ping-pong ball: 0.14%

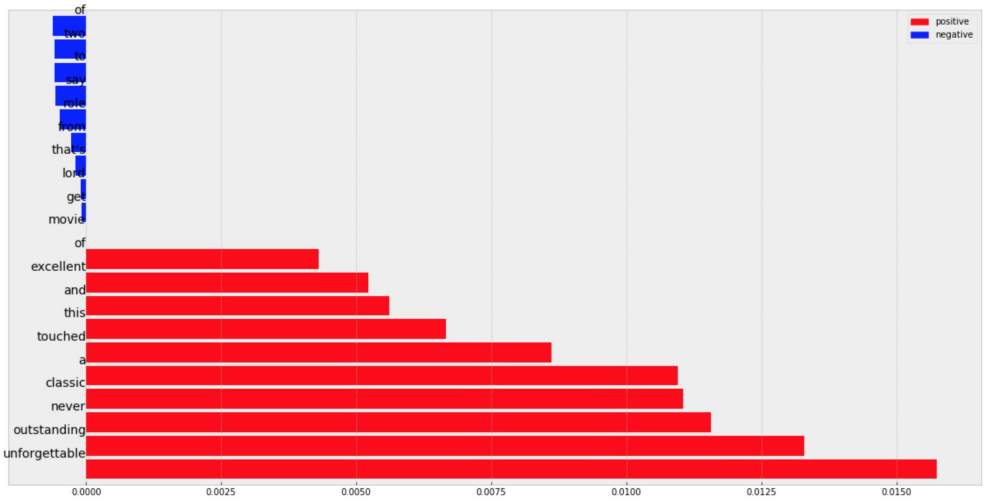
Relevance Type: integrated gradient



What about text classification?

X = excellent tale of two boys that do whatever they can to get away from there abusive father lord of the rings star elijah wood is outstanding in this unforgettable role this movie is one of the main reasons i haven't touched a single beer and never will as long as i live that might make me sound like a nerd but that's what i have to say it is a wonder why this isn't as a classic american tale





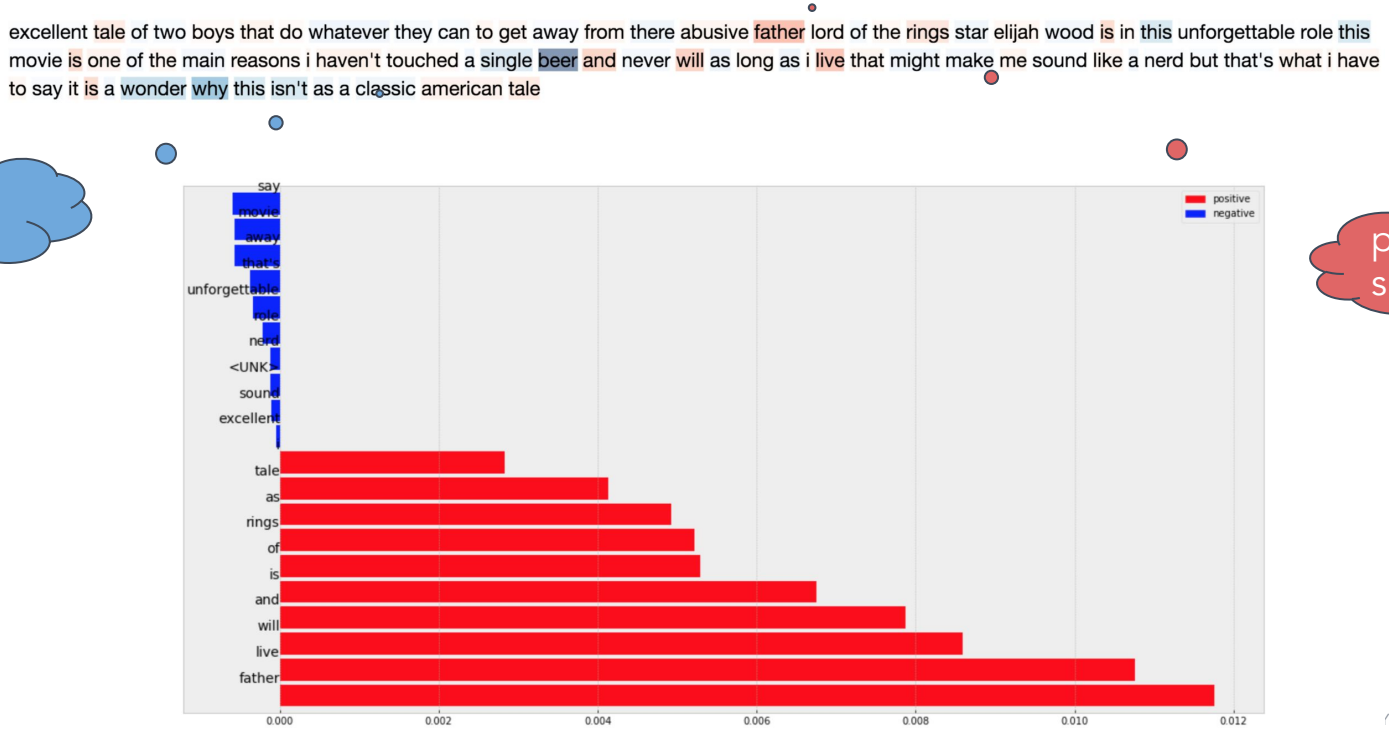
Trained Model: CNN for sentiment analysis
Dataset: IMDB

For a trained DNN classifier model(F),

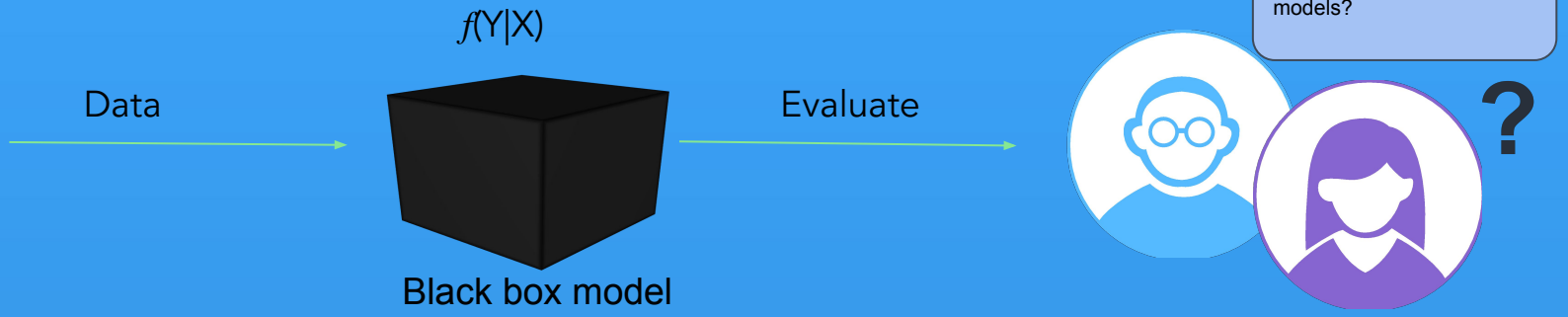
- Craft an adversarial attack by adding perturbation ΔX
  $X^* = X + \Delta X$
- ΔX = <Insert, delete, replace>
- $F(X) \neq F(X^*)$

- Replace [“outstanding”, “excellent”, “classic”, “unforgettable”, “touched”, “never”] with “<UNK>”

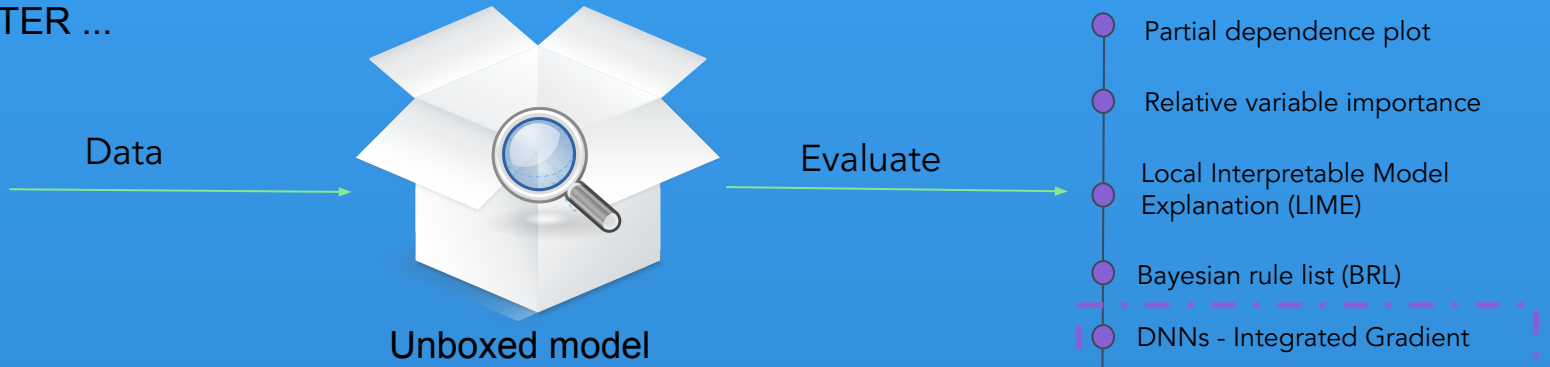
<UNK> tale of two boys that do whatever they can to get away from there abusive father lord of the rings star elijah wood is <UNK> in this <UNK> role this movie is one of the main reasons i haven't <UNK> a single beer and <UNK> will as long as i live that might make me sound like a nerd but that's what i have to say it is a wonder why this isn't as a <UNK> american tale



WITHOUT INTERPRETATION ...



WITH SKATER ...



R or Python model (linear, nonlinear, ensemble, neural networks)

Scikit-learn, caret and rpart packages for CRAN

H20.ai, Algorithmia, etc.

Special Thanks

- Marco Ancona, Researcher at ETH Zurich-Department of Computer Science, for helping us in enabling the journey of supporting interpretation for DNNs in Skater

- O'Reilly and AI Conference for allowing us to share our thoughts with all you guys

Q&A

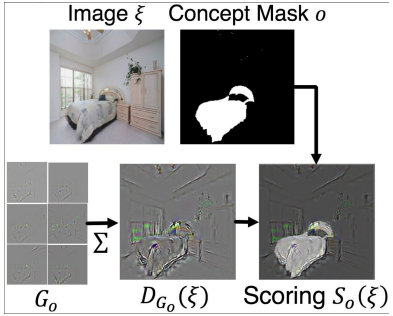
info@datascience.com


@DataScienceInc


@MaverickPramit

Future Work and Improvement

- Add better support for generating adversarial examples
- Define and support ways to quantitatively evaluate Model Interpretation
- Implement feature perturbation using Occlusion.
- Add support for other frameworks - PyTorch/MXNet/CNTK



- Explore concept recognition. Xie, Sarker, Doran, Hitzler, & Raymer. (2017). Relating Input Concepts to Convolutional Neural Network Decisions. arXiv:1711.08006
- More experimentation and notebook examples on inferring DNNs

References

- Szegedy C. et al. (2014). Intriguing properties of neural networks. arXiv:1312.6199
- Basch S. et al. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PloS one, 10(7):e0130140
- Canziani A., Paszke A., Culurciello E. (2016). An Analysis of Deep Neural Network Models for Practical Applications. arXiv:1605.07678
- Jain S., Ghosh R. (2017). Visualizing Deep Learning Networks. arXiv:1605.07678
- Liang B. et al. (2017) Deep Text Classification Can be Fooled. arXiv:1704.08006
- Sundararajan M., Taly A., Yan Q. (2017) Axiomatic Attribution for Deep Networks. arXiv:1703.01365
- Ancona M., Ceolini E., Cengiz Ö., Gross M. (2018). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. arXiv: 1711.06104

- Olah, C. et al. (2018). The Building Blocks of Interpretability.

END