DEEP LEARNING

Lecture 15: Special Topics in Deep Learning

Dr. Yang Lu

Department of Computer Science and Technology

luyang@xmu.edu.cn

厦門大學信息学院(特色化示范性软件学院)
School of Informatics Xiamen University (National Characteristic Demonstration Software School)

厦门大学 计算机科学与技术系
Department of Computer Science and Technology, Xiamen University

Outlines

- Knowledge Distillation
- Adversarial Samples
- Model Interpretation
- Fairness
- Privacy

KNOWLEDGE DISTILLATION

Problem

- The large-scale deep models have achieved overwhelming successes.
- However, it requires the huge computational complexity and massive storage.
- It is challenging to deploy them in real-time applications, especially on devices with limited resources, such as video surveillance and autonomous driving cars.
- Can we make small model work comparable with large model?

Solution

- Parameter pruning and sharing: Removing inessential parameters from deep neural networks without any significant effect on the performance.
- Low-rank factorization: Identify redundant parameters of deep neural networks by employing the matrix and tensor decomposition.
- Knowledge distillation (KD): These methods distill the knowledge from a larger deep neural network into a small network.

Knowledge Distillation

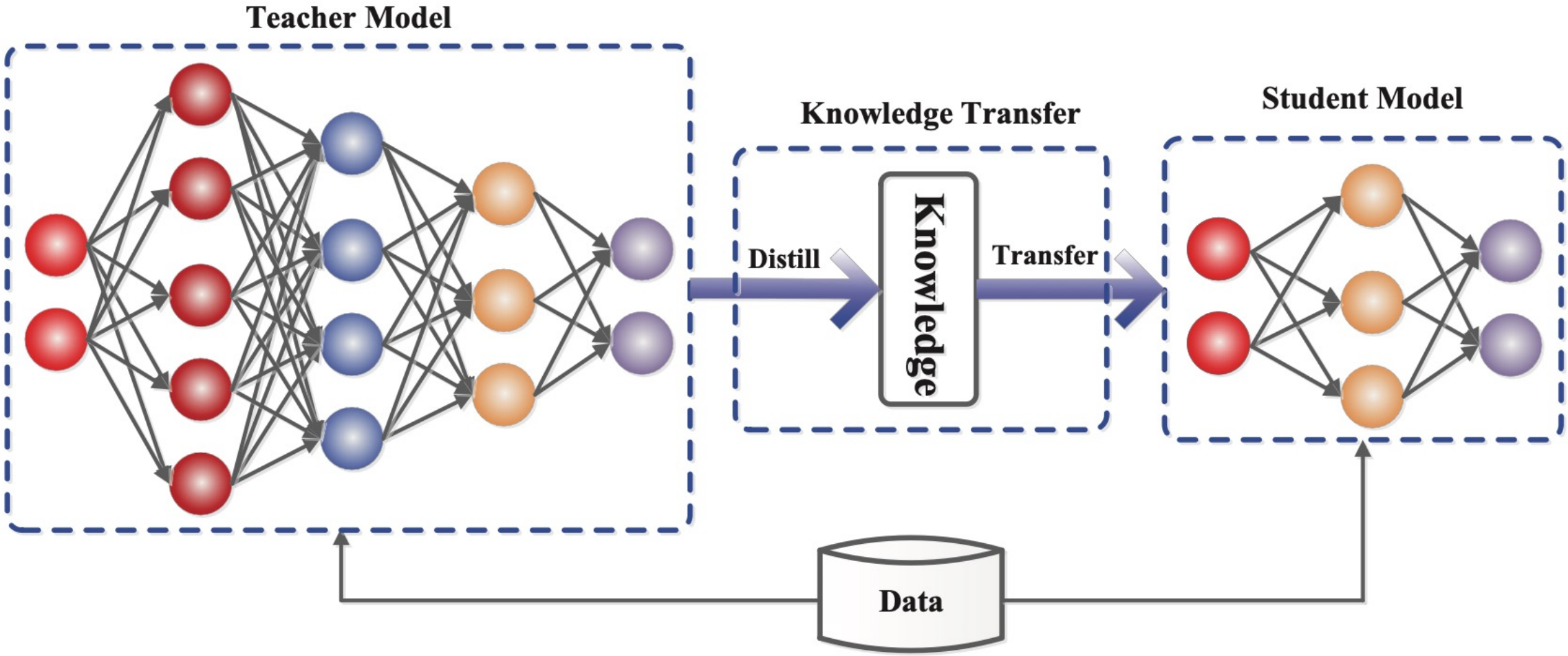


Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

Knowledge Source

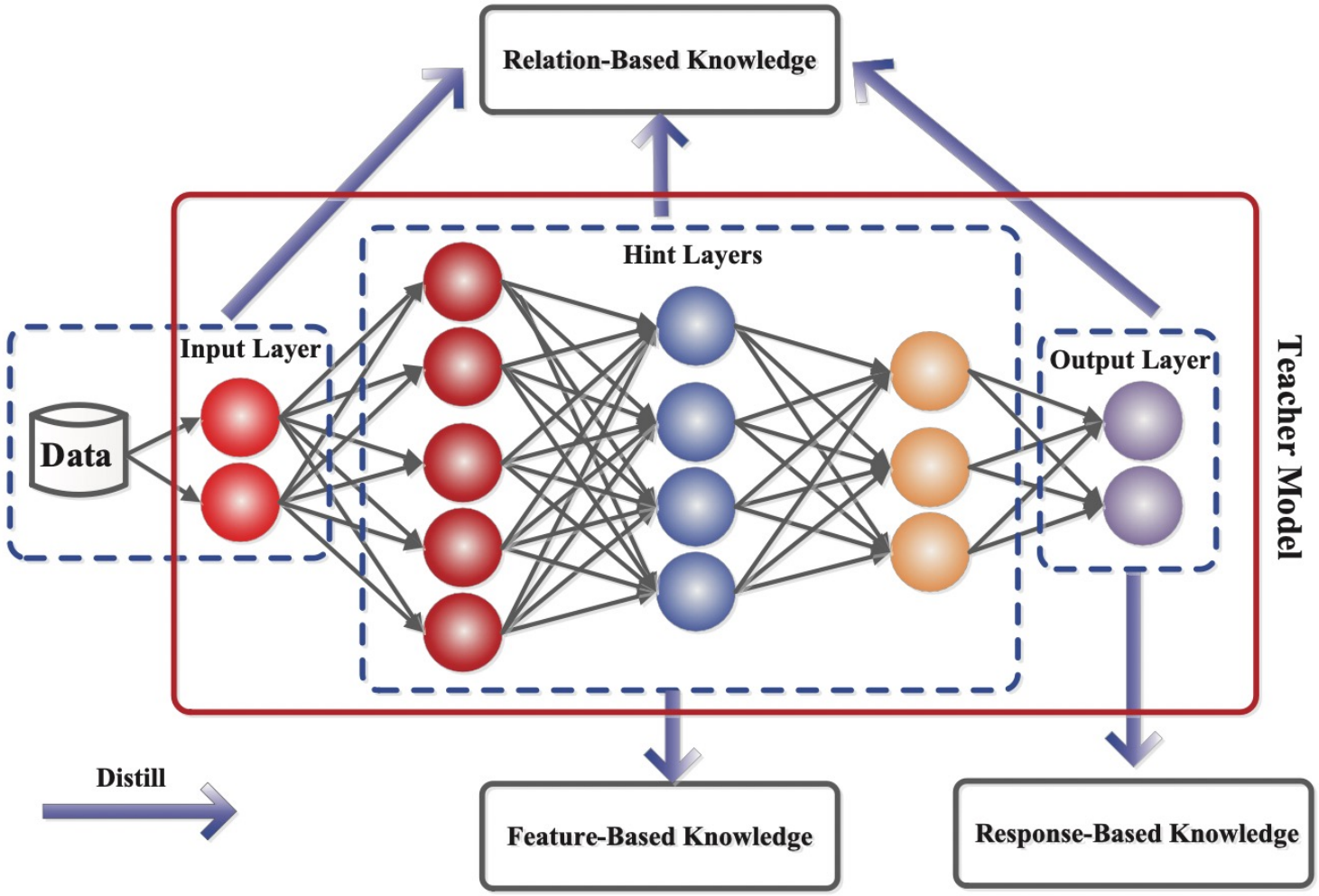


Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

Response-Based Knowledge

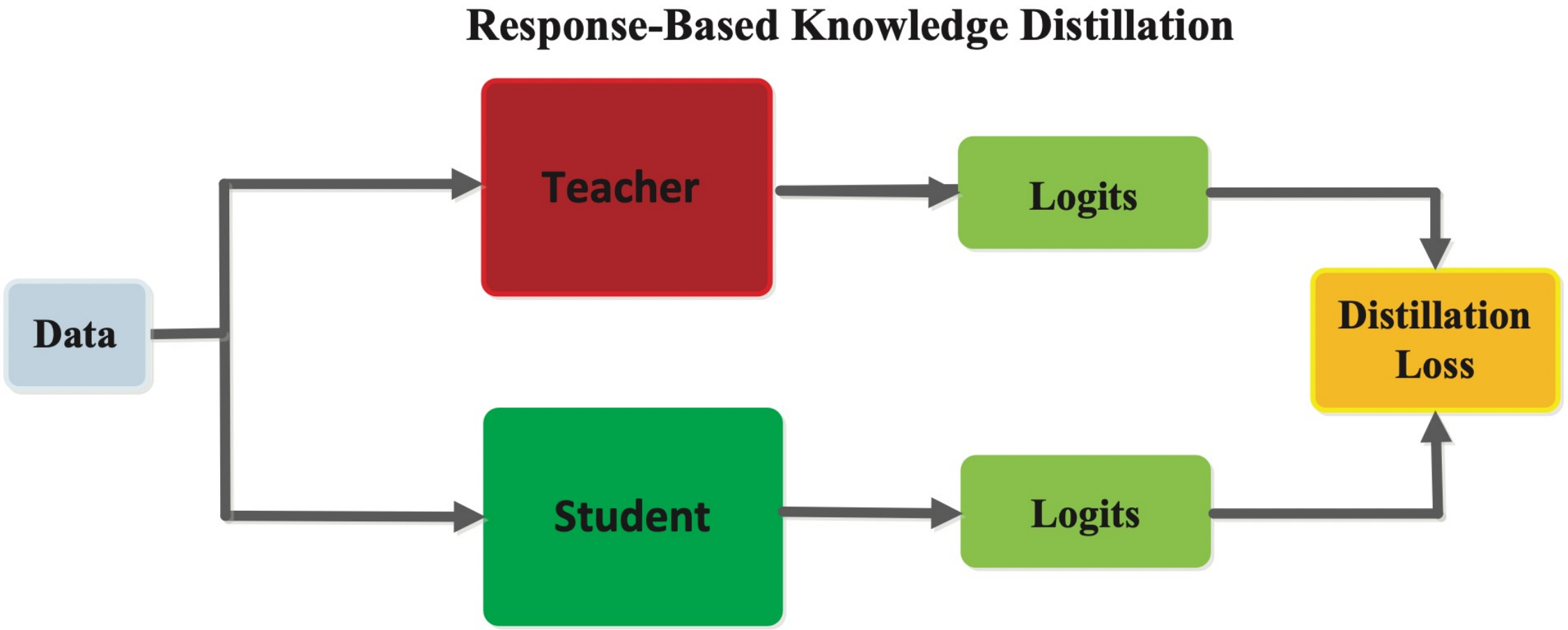


Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

Knowledge Distillation

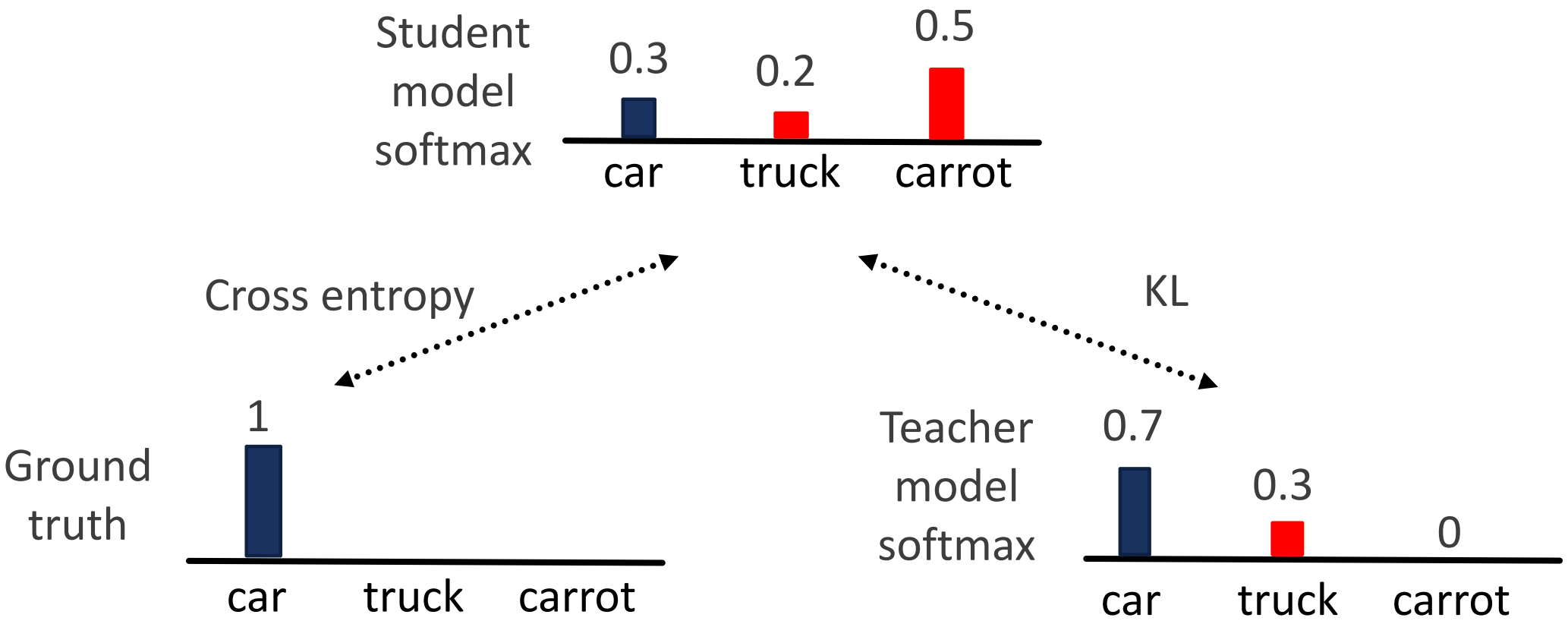
Distilling the **knowledge** in a **neural network**
G Hinton, O Vinyals, J Dean - arXiv preprint arXiv:1503.02531, 2015 - arxiv.org
… In this section, we investigate the effects of ensembling Deep **Neural Network** (DNN) … that the **distillation** strategy that we propose in this paper achieves the desired effect of **distilling** an …
☆ Save 99 Cite Cited by 17032 Related articles All 27 versions »

- Key idea: given a training sample, use the logits of the teacher model to help the student model.
  - Logits can be viewed as soft labels, which include knowledge from the teacher model.
- A common loss of knowledge distillation is:

$$L_{KD} = \frac{1}{|D|}\sum_{x \in D} KL\left(\text{softmax}(f_t(x)), \text{softmax}(f_s(x))\right)$$

$$L = (1-\lambda)L_{CE} + \lambda L_{KD}$$

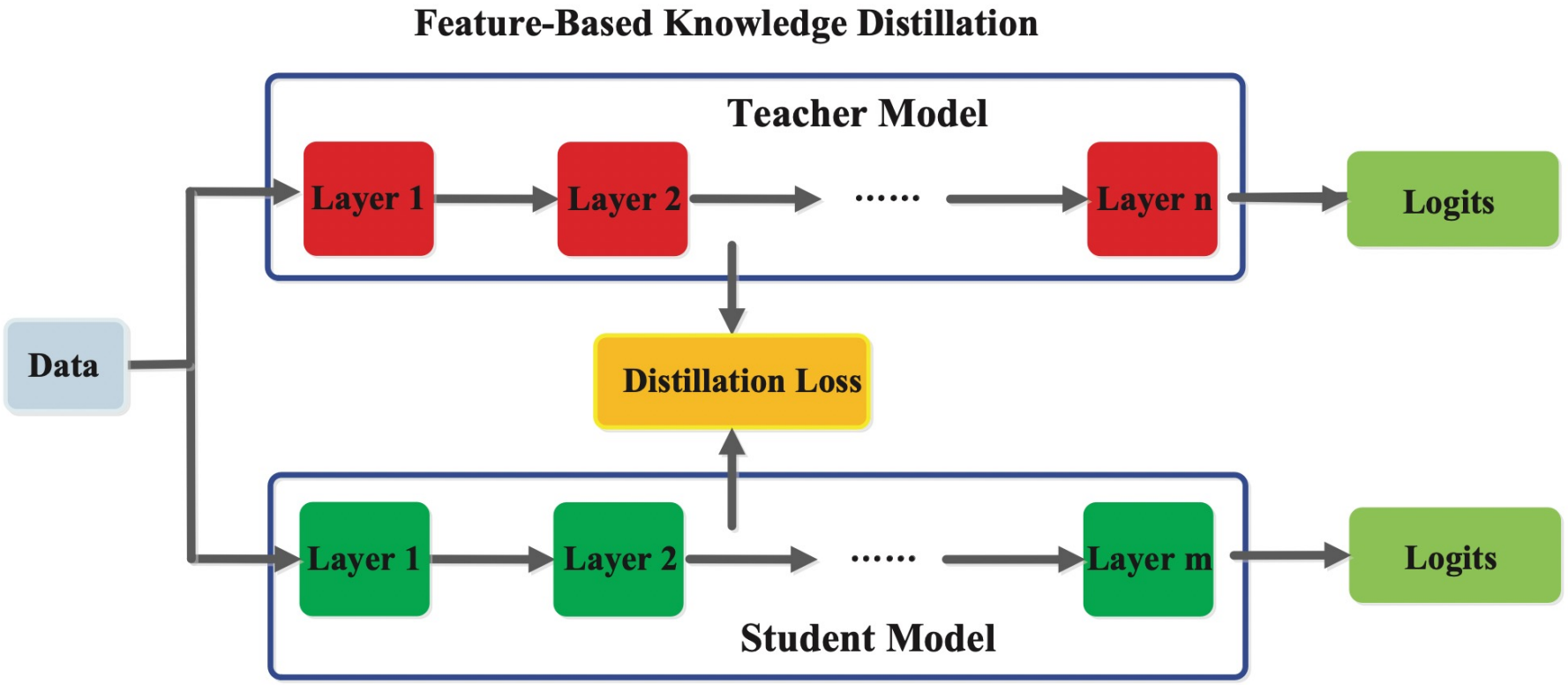
Knowledge Distillation



Common learning

Knowledge distillation

Feature-Based Knowledge



Output of a teacher's hidden layer that supervises the student's learning is called hint.

Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

Feature-Based Knowledge

- Generally, the distillation loss for feature-based knowledge transfer can be formulated as

$$\sum_i L\left(\Phi_t^i\left(f_t(x)\right), \Phi_s^i\left(f_s(x)\right)\right)$$

where Φ_t^i and Φ_s^i are transformation functions for layer i to make the representation between teacher and student comparable.

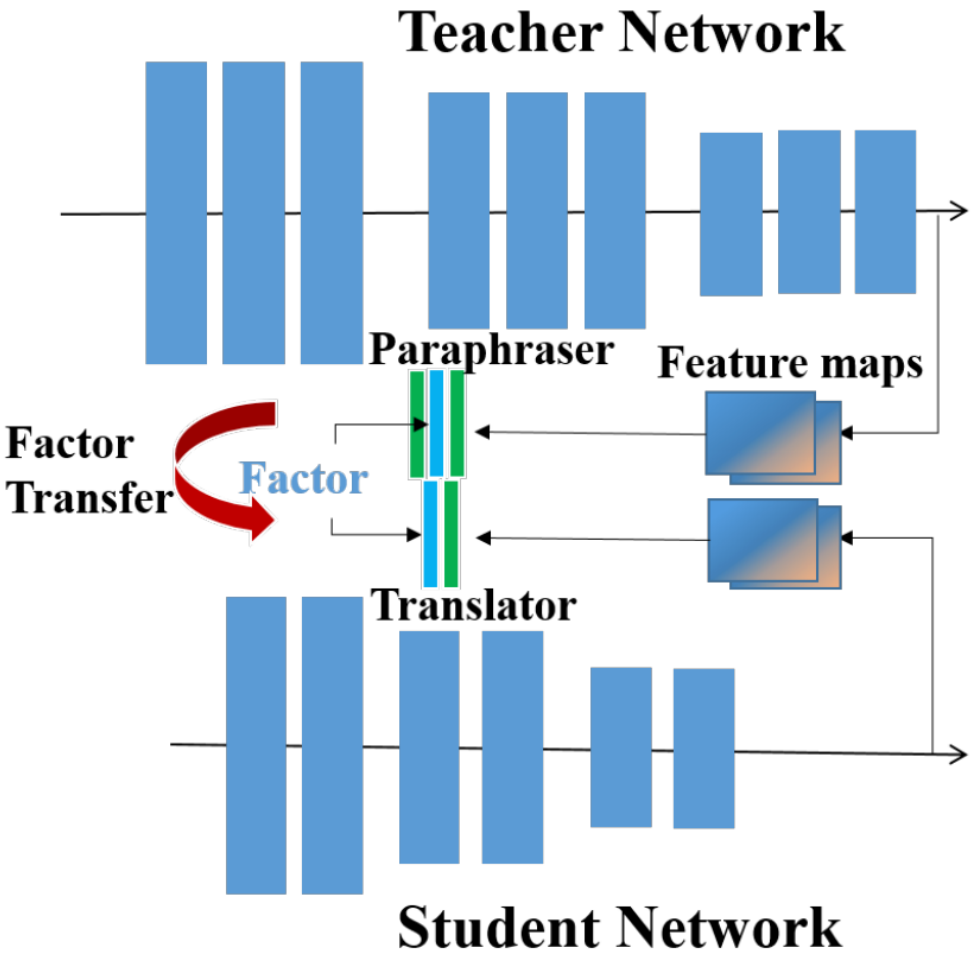


Image source: Kim, Jangho, Seonguk Park, and Nojun Kwak. "Paraphrasing Complex Network: Network Compression via Factor Transfer." In NeurIPS. 2018.

Relation-Based Knowledge

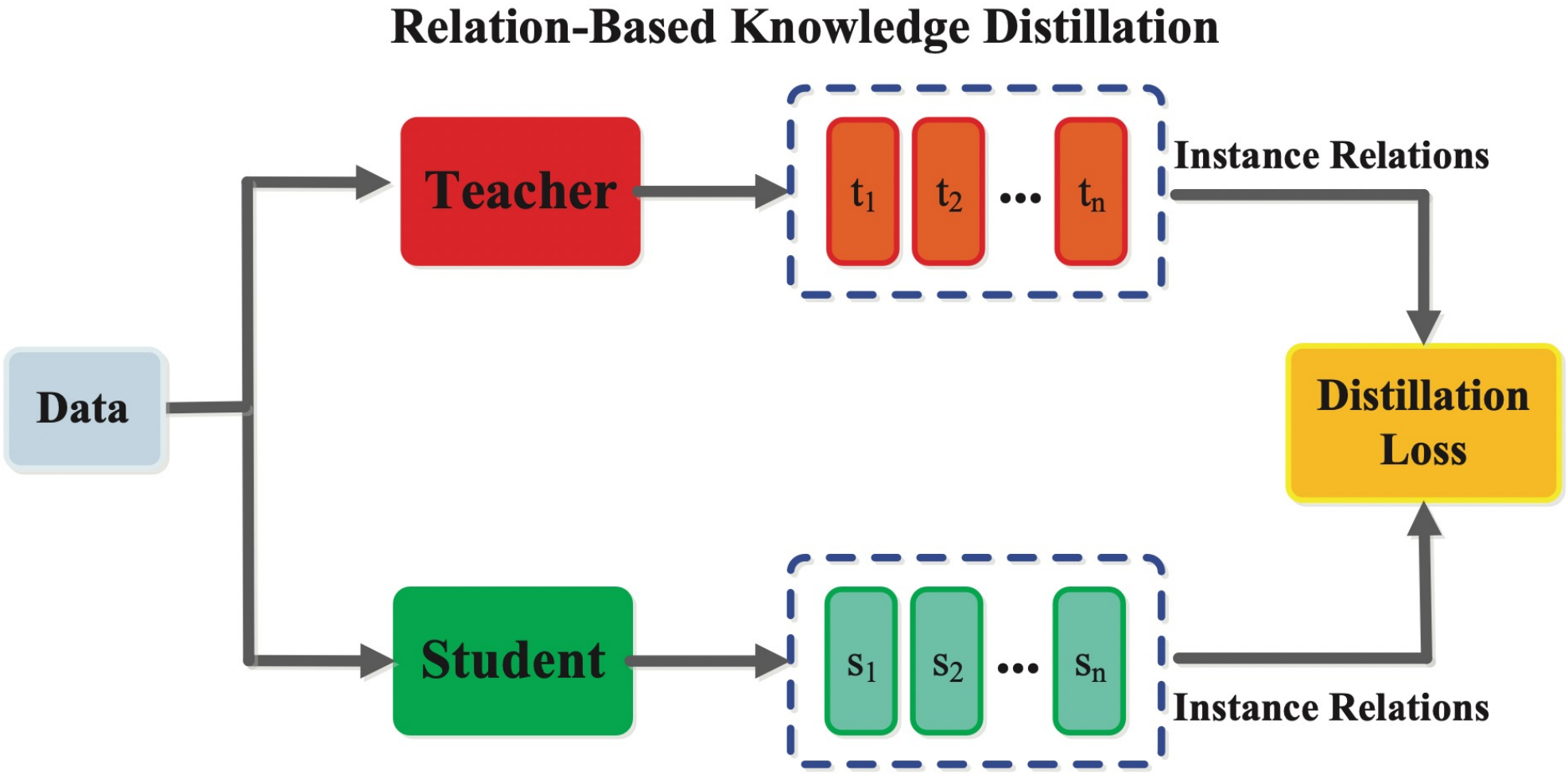


Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

Relation-Based Knowledge

Relational knowledge distillation
W Park, D Kim, Y Lu, M Cho - Proceedings of the IEEE/CVF …, 2019 - openaccess.thecvf.com
… Saussure's concept of the **relational** identity of signs is at the … is that what constitutes the **knowledge** is better presented by … to KD, dubbed **Relational** Knowledge **Distillation** (RKD), that …
☆ Save 99 Cite Cited by 1158 Related articles All 8 versions »

- Instance relation-based knowledge explores the relationships between data samples.
- The distillation loss of relation-based knowledge based on the instance relations can be formulated as

$$L\left(\phi_t(t_i, t_j), \phi_s(s_i, s_j)\right)$$

where ϕ_t and ϕ_s are the similarity functions of sample pairs.

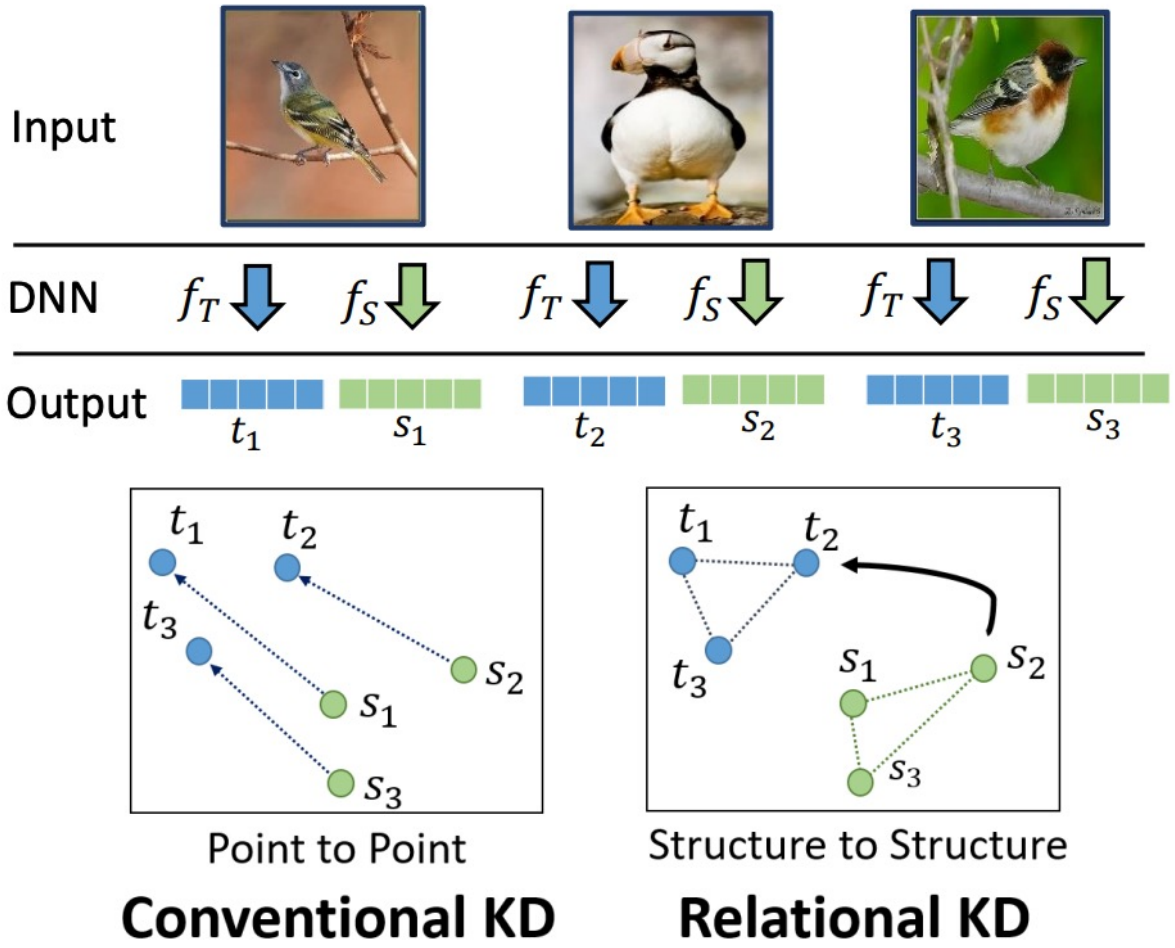


Image source: Park, Wonpyo, Dongju Kim, Yan Lu, and Minsu Cho. "Relational knowledge distillation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3967-3976. 2019.

Distillation Schemes

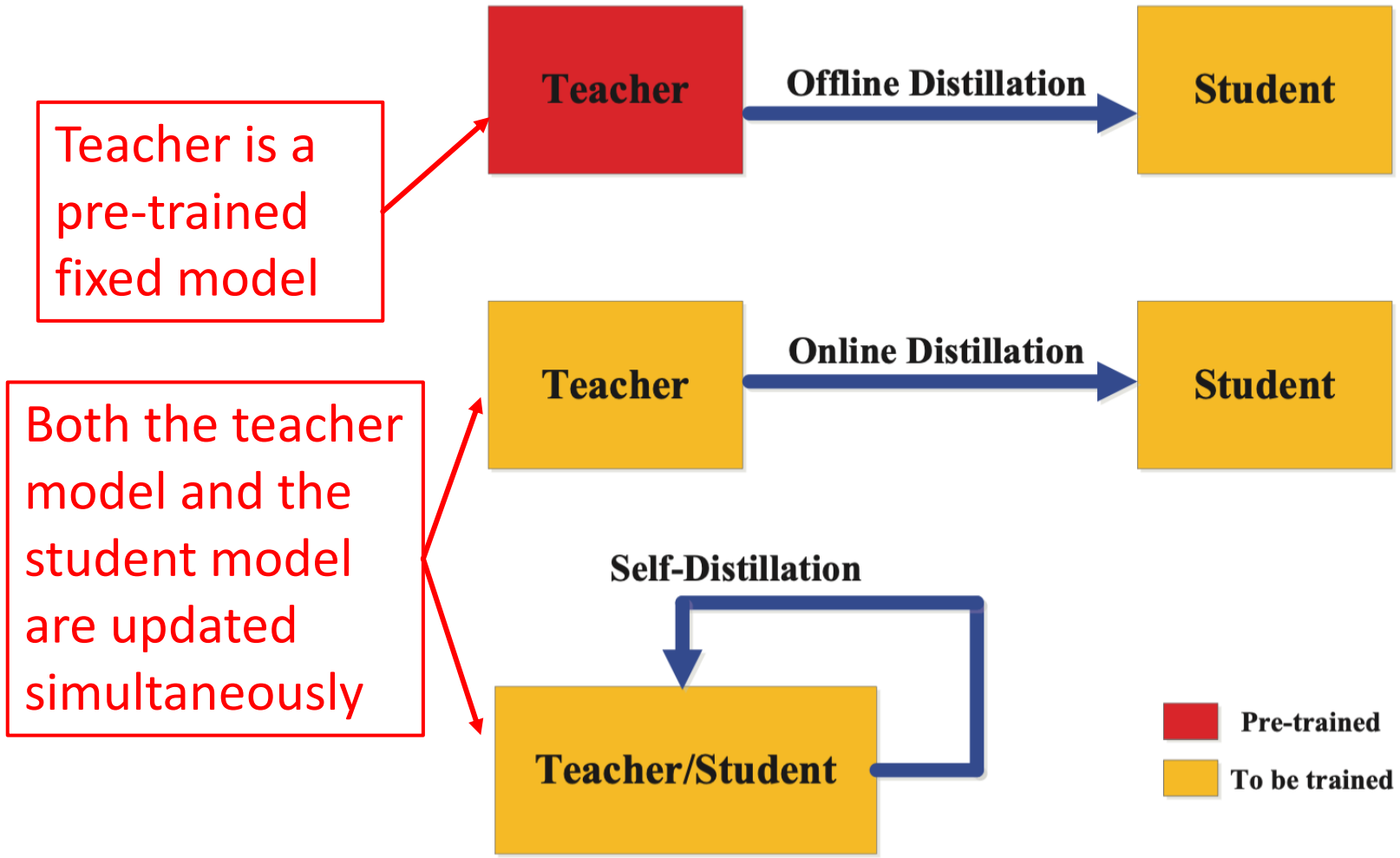
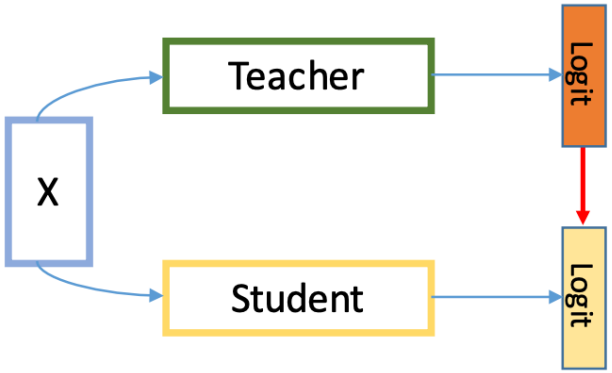
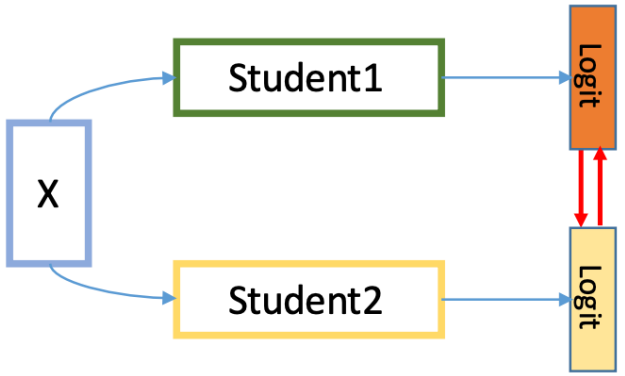


Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

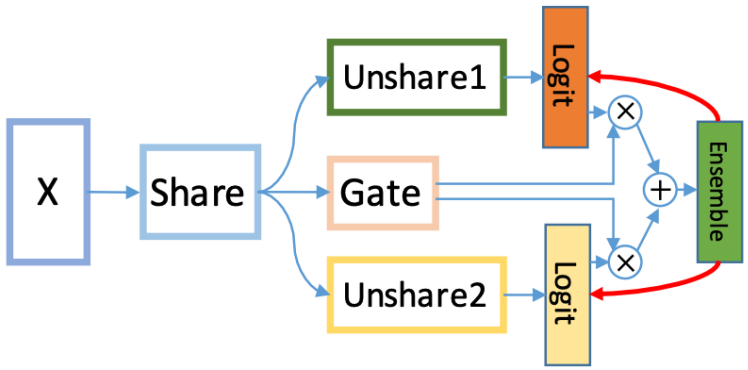
Online Distillation



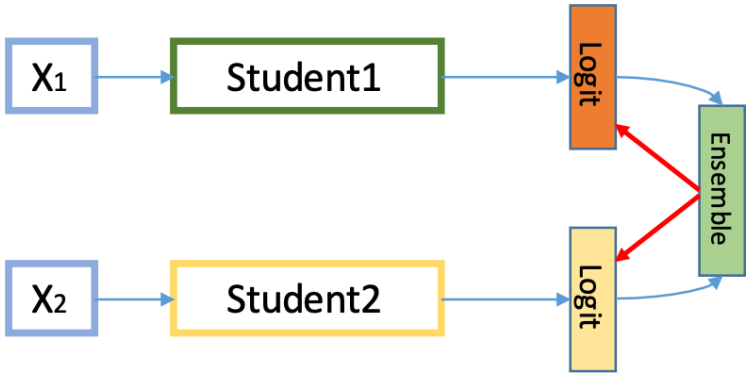
(a) Baseline



(b) DML



(c) ONE



(d) KDCL

Image source: Guo, Qiushan, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. "Online knowledge distillation via collaborative learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11020-11029. 2020.

Self-Distillation

Be **your own teacher**: Improve the performance of convolutional neural networks via self distillation

L Zhang, J Song, A Gao, J Chen… - Proceedings of the …, 2019 - openaccess.thecvf.com

Convolutional neural networks have been widely deployed in various application scenarios. In order to extend the applications' boundaries to some accuracy-crucial domains, …

☆ Save 99 Cite Cited by 694 Related articles All 11 versions »

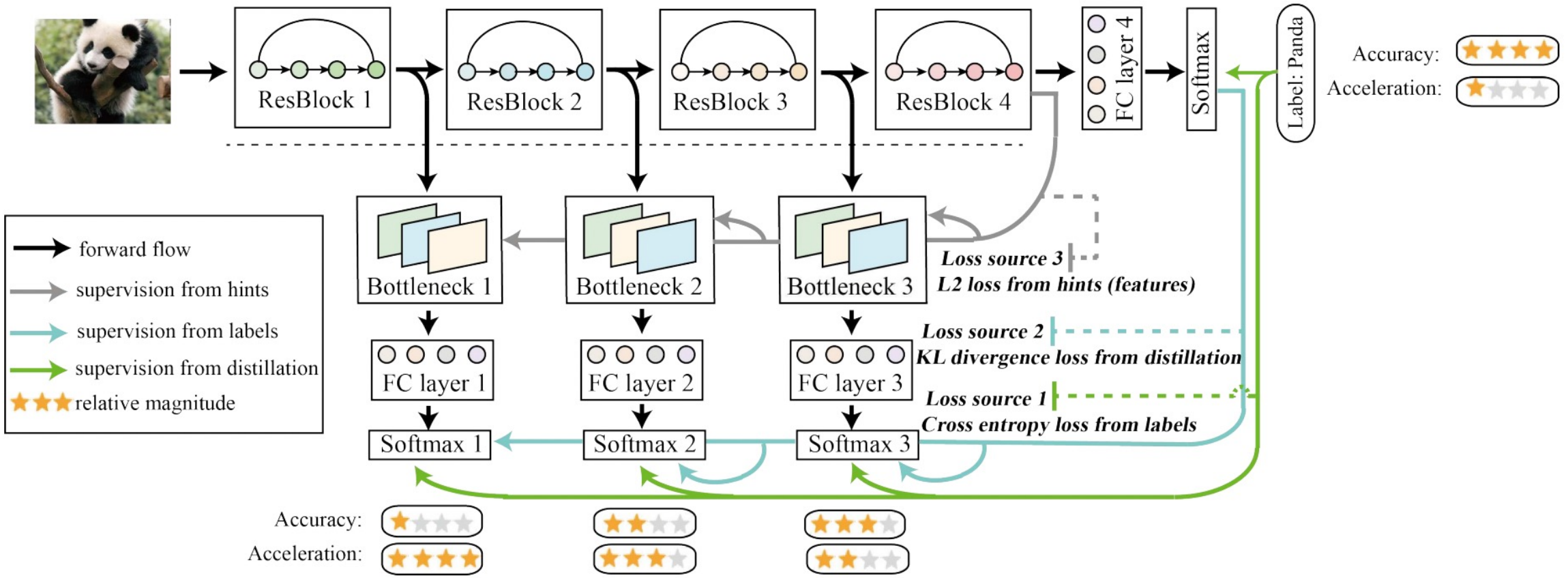
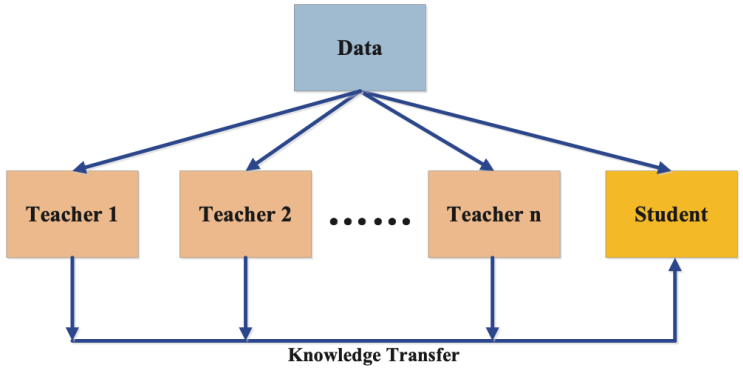
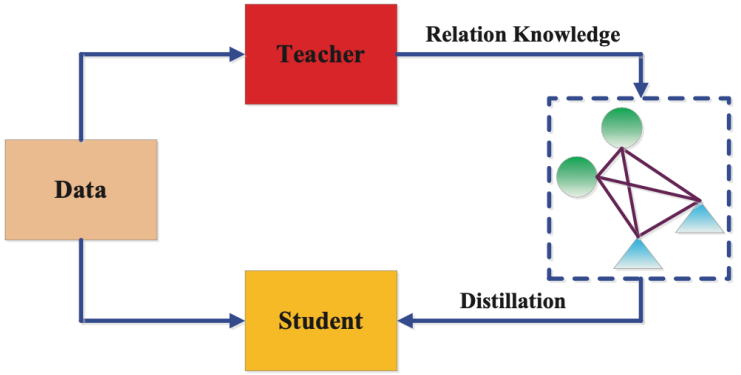


Image source: Zhang, Linfeng, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. "Be your own teacher: Improve the performance of convolutional neural networks via self distillation." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3713-3722. 2019.

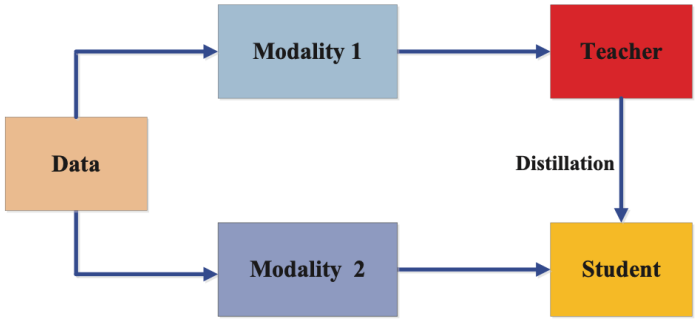
Applications



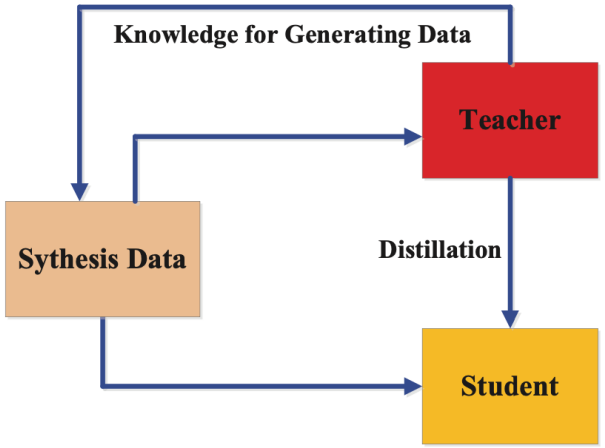
Multi-teacher distillation



Graph-based distillation



Cross-modal distillation



Data-free distillation

Image source: J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.

ADVERSARIAL SAMPLES

Adversarial Samples

- Adversarial samples are the samples $\boldsymbol{x}'$ that can be so similar to $\boldsymbol{x}$ that a human observer cannot tell the difference, but the network can make highly different predictions.
- By adding an imperceptibly small vector to the input, we can change the prediction of the image.

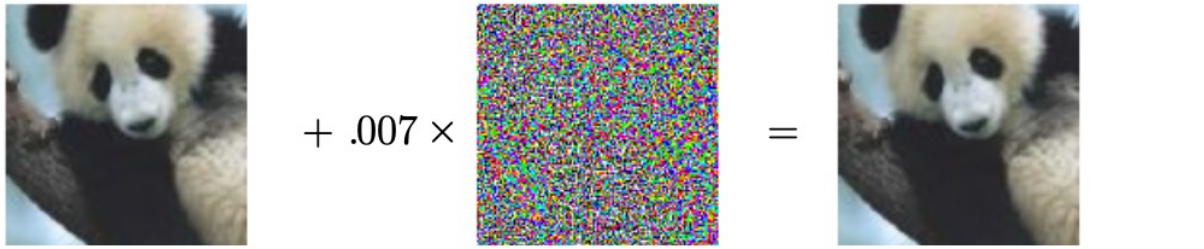


Image source: Figure 7.8, Goodfellow, Bengio, and Courville, Deep Learning, Cambridge: MIT press, 2016.

Adversarial Examples

Explaining and **harnessing** adversarial examples
IJ Goodfellow, J Shlens, C Szegedy - arXiv preprint arXiv:1412.6572, 2014 - arxiv.org
… We start with **explaining** the existence of adversarial … To **explain** why mutiple classifiers assign the same class to … Our hypothesis does not **explain** all of the maxout network's mistakes or …
☆ Save 99 Cite Cited by 18477 Related articles All 17 versions »

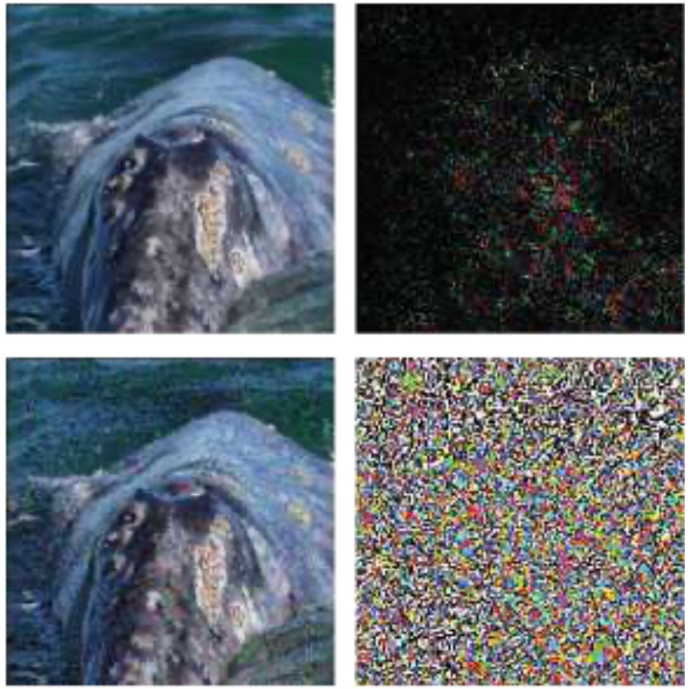
- Usually, we fix training samples and update model parameters. The goal is to decrease the loss.
- To obtain adversarial examples, we fix model parameters and set training samples as a part of model parameters to update. The goal is to increase the loss.
- Fast gradient sign method (FGSM) generates adversarial examples by adding a simple perturbation:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\theta, \boldsymbol{x}, y)\right)$$

Adversarial Examples

- DeepFool creates better adversarial samples with smaller perturbation.

Original image predicted as “whale”



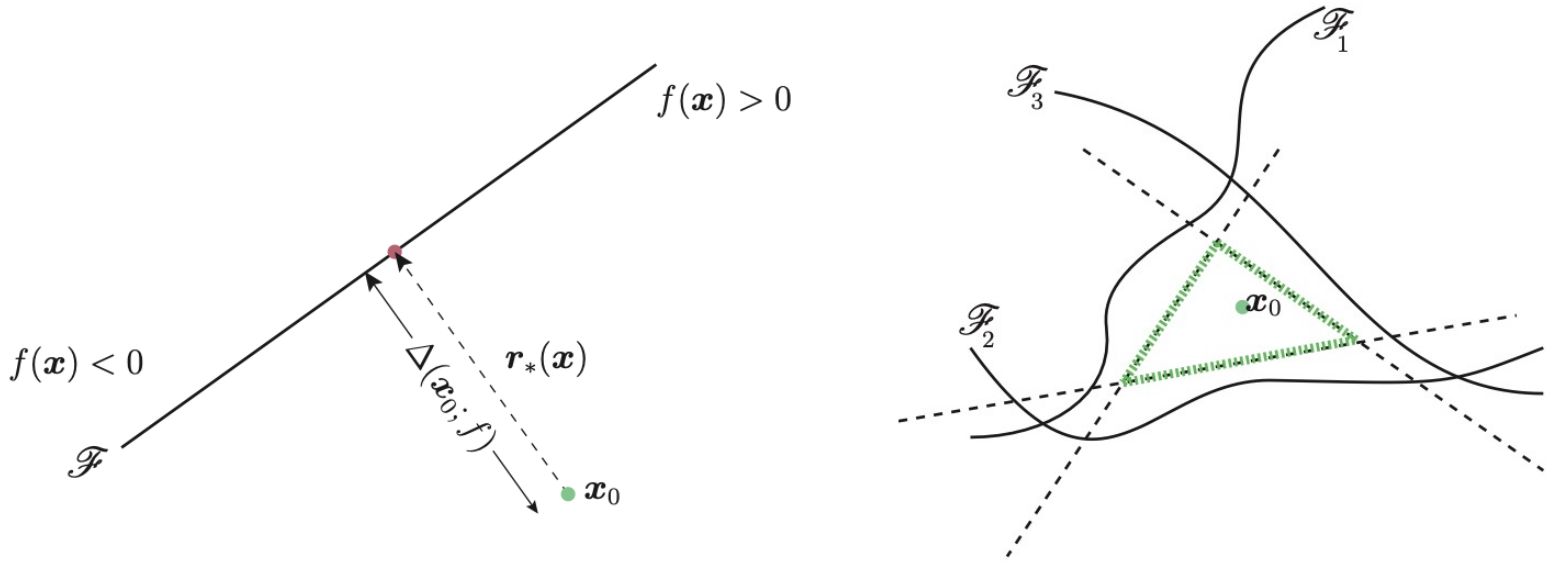
Adversarial image by DeepFool predicted as “turtle”

Adversarial image by FGSM predicted as “turtle”

Image source: Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Adversarial Examples

- DeepFool is an untargeted attack algorithm that aims to find the least perturbation leading to misclassification, by projecting an image to the closest separating hyperplane.



Select the easiest class to create the adversarial sample

Image source: Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Adversarial Examples

- Carlini & Wagner (C&W) attack formulates targeted adversarial attacks as an optimization problem.

$$\min D(x, x+\delta)$$
$$s.t. \quad f(x+\delta) = t$$
$$x+\delta \in [0,1]^n$$

where D is some distance metric, e.g., L_0, L_2, or L_∞. t is the target class.

- A lot of losses can be adopted to minimize this objective, rather than simply use the gradient of the target class by cross entropy.

Adversarial Goals

- Confidence Reduction. The adversaries attempt to reduce the confidence of prediction for the target model.
  - E.g., the adversarial samples of a “stop” sign is predicted with lower confidence.
- Misclassification. The adversaries attempt to change the output classification of input to any class different from the original class.
  - E.g., an adversarial sample of a “stop” sign is predicted to be any class different from the “stop” sign.
- Targeted Misclassification. The adversaries attempt to change the output classification of the input to a special target class.
  - E.g., any adversarial samples inputted into a classifier is predicted to be a “go” sign.
- Source/Target Misclassification. The adversaries attempt to change the output classification of a special input to a special target class.
  - E.g., a “stop” sign is predicted to be a “go” sign.

White-Box Attack

- If one wants to use FGSM or PGD to attack a model, one pre-requisite is to know the model architecture and learned parameters.

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \epsilon \text{sign}\left(\nabla_{\boldsymbol{x}} J(\theta, \boldsymbol{x}, y)\right)$$

- In white-box attacks, the adversaries have total knowledge about the target about the target model, including algorithm, training data distribution, and model parameters.
  - Everything is known to the attacker.
- If the model in unknown to the attacker, can we still generate adversarial examples to attack this model?

Black-Box Attack

- The attackers have no knowledge about target model in black-box attacks.

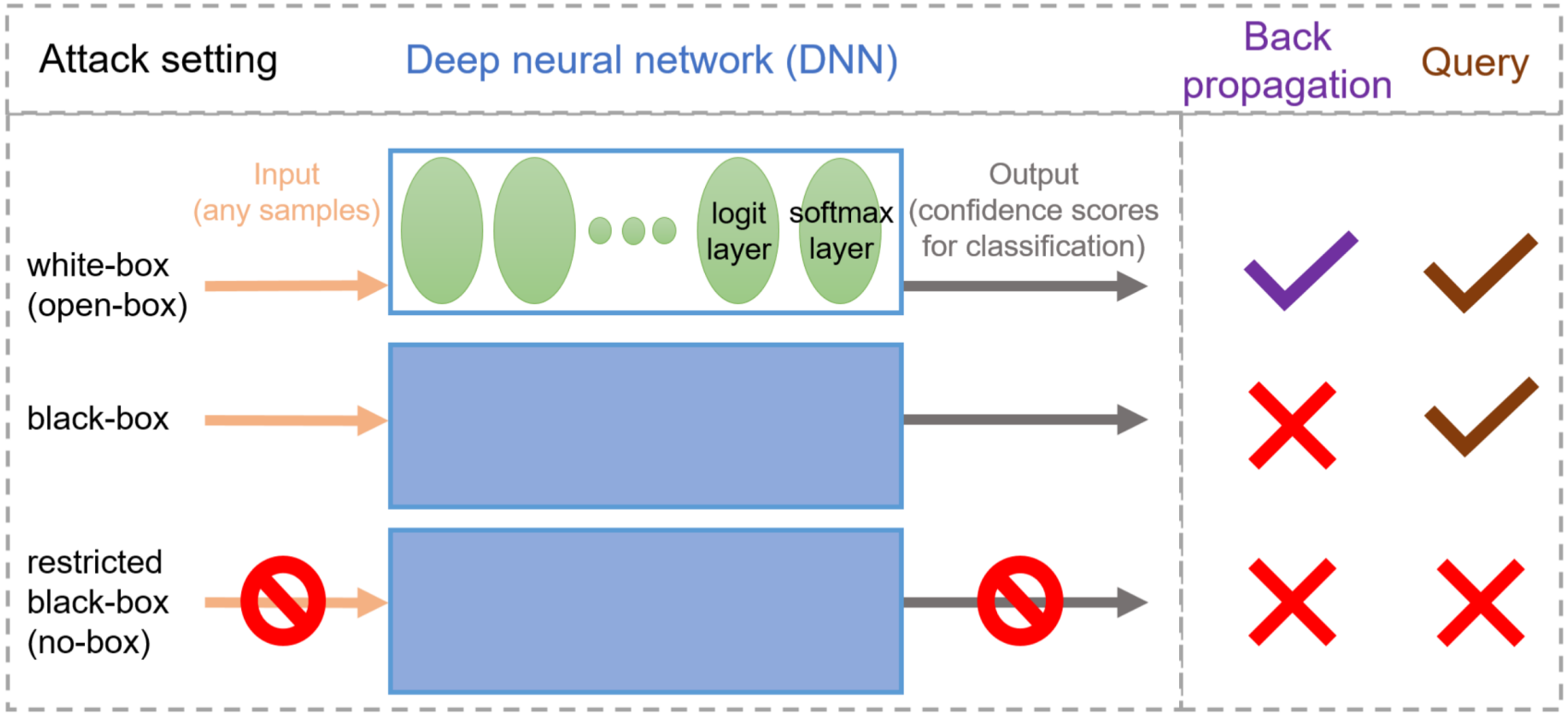


Image source: Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017.

Black-Box Attack

- A most straightforward way to conduct black-box attact is to obtain a substitute model.
  - One can use a lot of samples and query the black-box model and collect their response prediction as a dataset (X, Y).
  - Use (X, Y) to train a substitute model and then apply white-box attach approaches.

Black-Box Attack

- The key difficulty of black-box attack is that we don't know how to generate perturbation.
- Zeroth order optimization (ZOO) uses the symmetric difference quotient to estimate the gradient and Hessian:

$$\hat{g}_i = \frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x - he_i)}{2h}$$

$$\hat{h}_i \frac{\partial^2 f(x)}{\partial x_{ii}^2} \approx \frac{f(x + he_i) - 2f(x) + f(x - he_i)}{h^2}$$

where h is a small constant (0.0001 in the paper) and e_i is a standard basis vector with only the i-th component as 1.

Require: Step size η

while not converged **do**
  Randomly pick a coordinate $i \in \{1, \cdots, p\}$
  Estimate $\hat{g}_i$ and $\hat{h}_i$ using (6) and (7)
  if $\hat{h}_i \leq 0$ **then**
    $\delta^* \leftarrow -\eta \hat{g}_i$
  else
    $\delta^* \leftarrow -\eta \frac{\hat{g}_i}{\hat{h}_i}$
  end if
  Update $\mathbf{x}_i \leftarrow \mathbf{x}_i + \delta^*$
end while

Image source: Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017.

Black-Box Attack

- Use GAN to generate perturbation in the noise space.

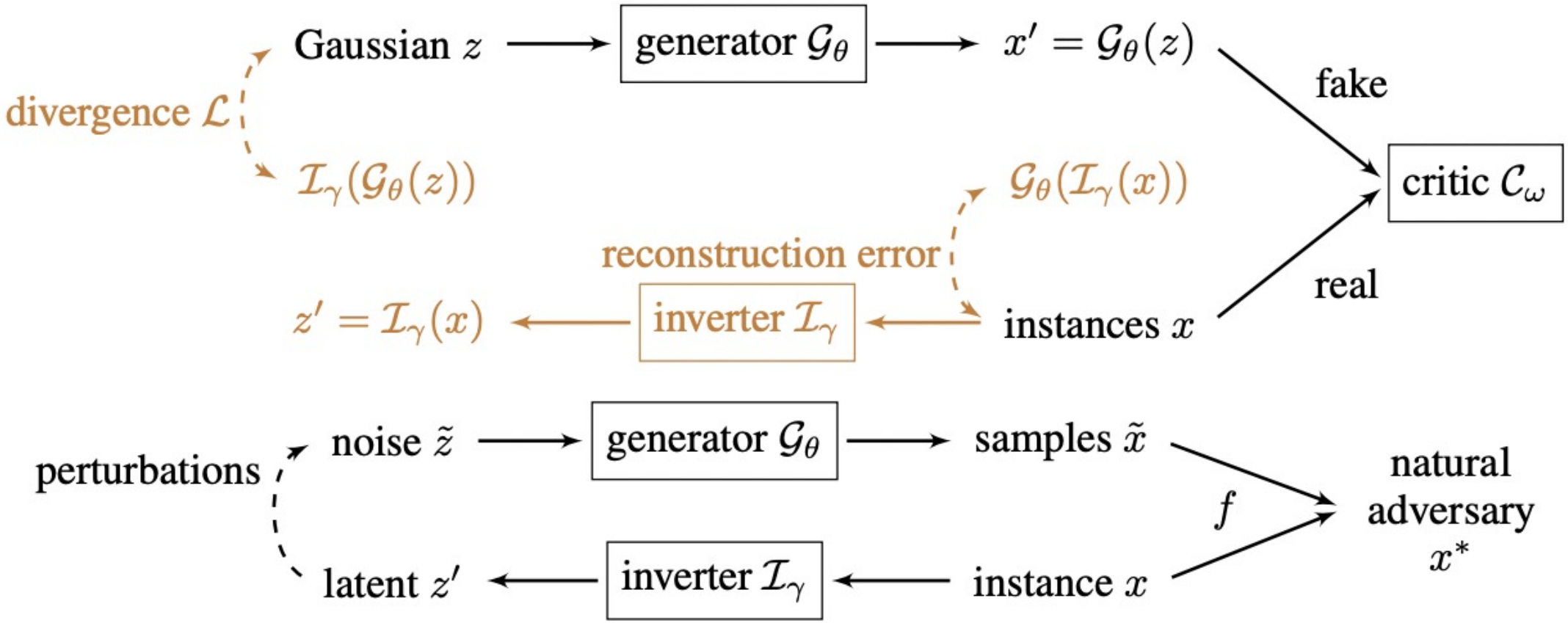


Image source: Zhao, Zhengli, Dheeru Dua, and Sameer Singh. "Generating Natural Adversarial Examples." International Conference on Learning Representations. 2018.

Black-Box Attack

- However, the perturbation in the image space is usually large.
- It is easily filtered out if we use some detection method.

Image source: Zhao, Zhengli, Dheeru Dua, and Sameer Singh. "Generating Natural Adversarial Examples." International Conference on Learning Representations. 2018.

Phisycal Attack

- Infrared light illuminate the attacker's face with some points from the camera angle, but people nearby would not notice.
- With this technique, they successfully attacked the facial authentication system in white-box settings.

Image source: Zhou, Zhe, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu, and Kehuan Zhang. "Invisible mask: Practical attacks on face recognition with infrared." arXiv preprint arXiv:1803.04683 (2018).

Adversarial Defense

- If we know that our model may be attacked. How can we prevent?
- One common observation from the development of security-related research is that attack and defense often come hand-in-hand.
  - One's improvement depends on the other's progress.
- Similarly, in the context of robustness of DNNs, more effective adversarial attacks are often driven by improved defenses, and vice versa.
- The goal of adversarial defense is to disable adversarial attacks while maintaining similar classification performance for the benign examples.

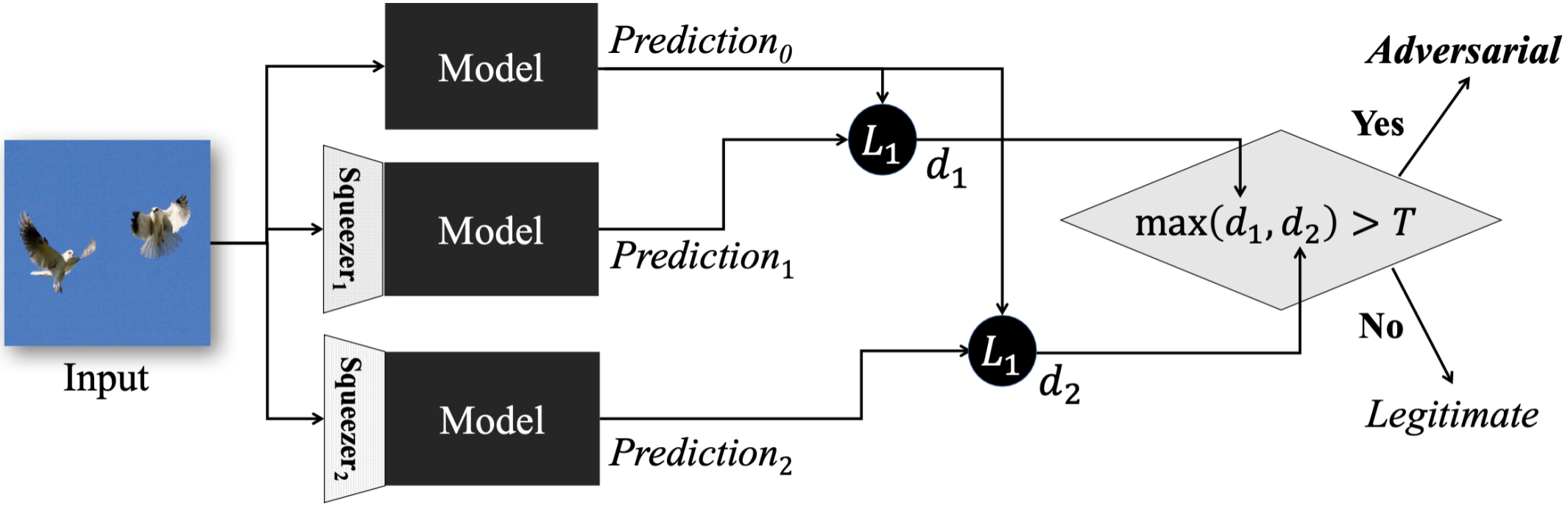
Adversarial Defense

- Adversarial training makes the model insensitive to small changes by encouraging the network to be locally constant in the neighborhood of the training data.
- For example, one can either add adversarial samples to the training set by FGSM or use it as a regularizer:

$$\tilde{J}(\theta, \boldsymbol{x}, y) = \alpha J(\theta, \boldsymbol{x}, y) + (1 - \alpha) J(\theta, \boldsymbol{x} + \text{sign}(\nabla_{\boldsymbol{x}} J(\theta, \boldsymbol{x}, y), y)$$

Adversarial Defense

- Detection-based approaches aim to differentiate an adversarial example from a set of benign examples.



Feature-squeezing framework for detecting adversarial examples

Image source: Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." arXiv preprint arXiv:1704.01155 (2017).

Adversarial Defense

- NULL labeling method adds new NULL label to the dataset and classify adversarial samples to the NULL label class.

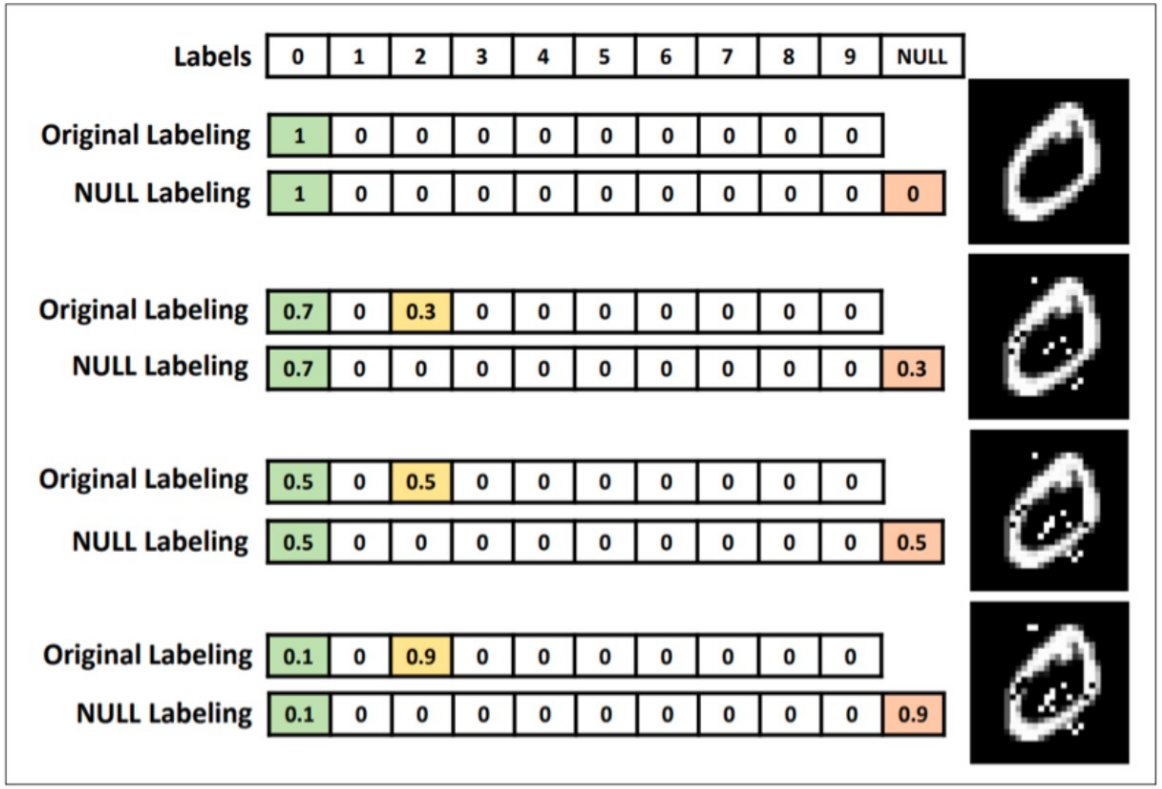


Image source: Hosseini, H.; Chen, Y.; Kannan, S.; Zhang, B.; Poovendran, R. Blocking transferability of adversarial examples in black-box learning systems. arXiv 2017, arXiv:1703.04318.

Adversarial Defense

- Gradient masking hides the gradient information while training a DNN.

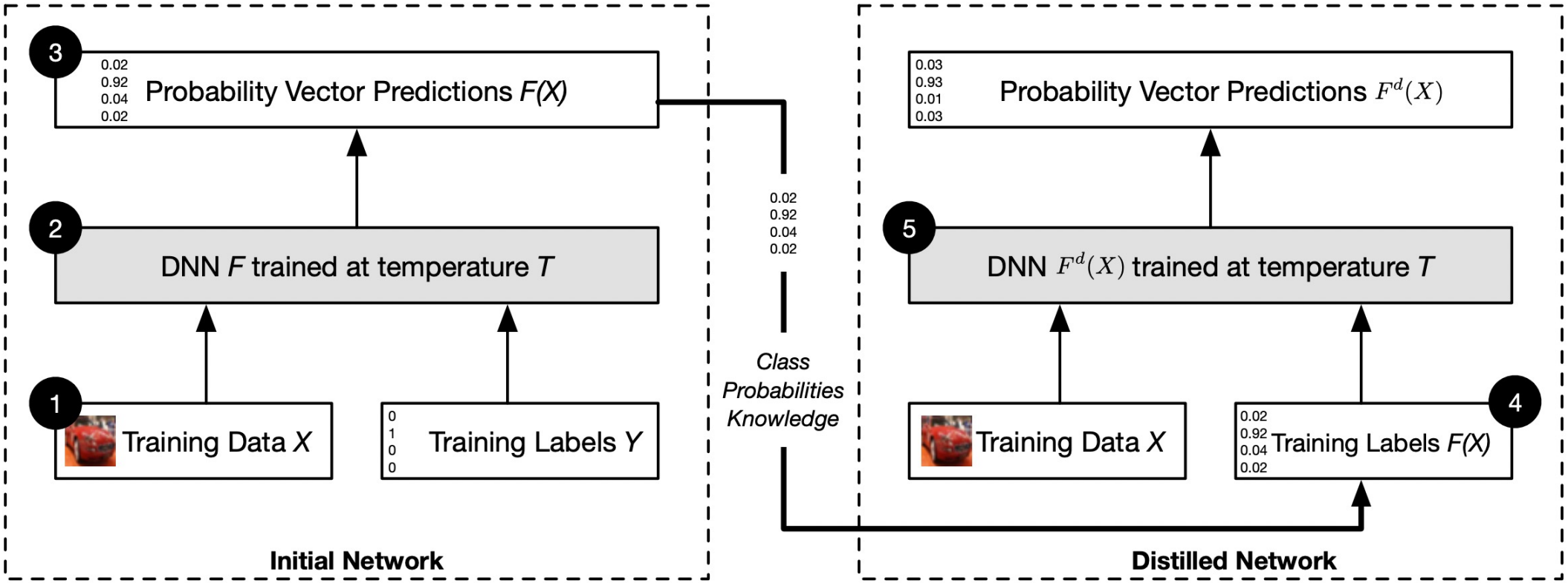


Image source: Papernot, Nicolas, et al. "Distillation as a defense to adversarial perturbations against deep neural networks." 2016 IEEE symposium on security and privacy (SP). IEEE, 2016.

Adversarial Defense

- Defense-GAN projects input images onto the range of the generator G by minimizing the reconstruction error.
- Use the optimized noise z^* to generate a corresponding $\hat{x}$ to feed the classifier.

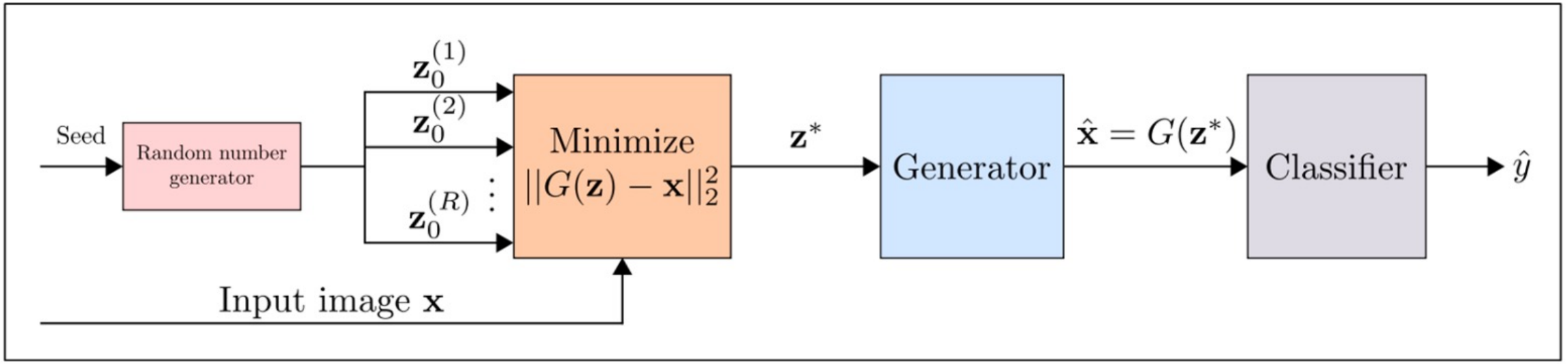


Image source: Samangouei, Pouya, Maya Kabkab, and Rama Chellappa. "Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models." International Conference on Learning Representations. 2018.

MODEL INTERPRETATION

Model Interpretation

- Although deep learning has shown its power, it is usually treated as a block-box predictor.
- We want to know the reason behind its inference, such that we can trust it.
  - If the model makes a correct prediction, it tells you why. And if the model makes a incorrect prediction, it also tells you why.
  - For example, if a diagnosis model successfully predicts a disease, the factors that make it correct are more interested to the doctors.
- Interpretation is the process of giving explanations to Human.

Model Interpretation

Why do we need model interpretation?

- Safety: can help expose safety issues.
- Mismatched objectives and multi-objective trade-offs: what you optimize is not what you meant to optimize.
- Debugging: understand why the system doesn’t work, and fix it.
- Sensitive domain: decisions in medicine, criminal justice, etc.
- Legal/Ethics: legally required to provide an explanation (e.g. GDPR) and/or we don't want to discriminate against particular groups.
- …

Model Interpretation

For example, we want a model to tell us the probability to date a guy:

- If the probability is close to 0 and the reason is that you are a computer science PhD student with no hair, it is acceptable.
- If the probability is close to 0 and the reason is

$$\frac{\exp(4y+2)}{\exp(5x+1)} + \exp(-xy-5) + 1 < 0.5$$

where x and y are part of your profile features, it is ?????

Interpretable Model

Which model is naturally interpretable?

Linear model

$$f(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

- The learned weight w_i directly implies its importance to feature x_i.
- E.g. linear regression and logistic regression.

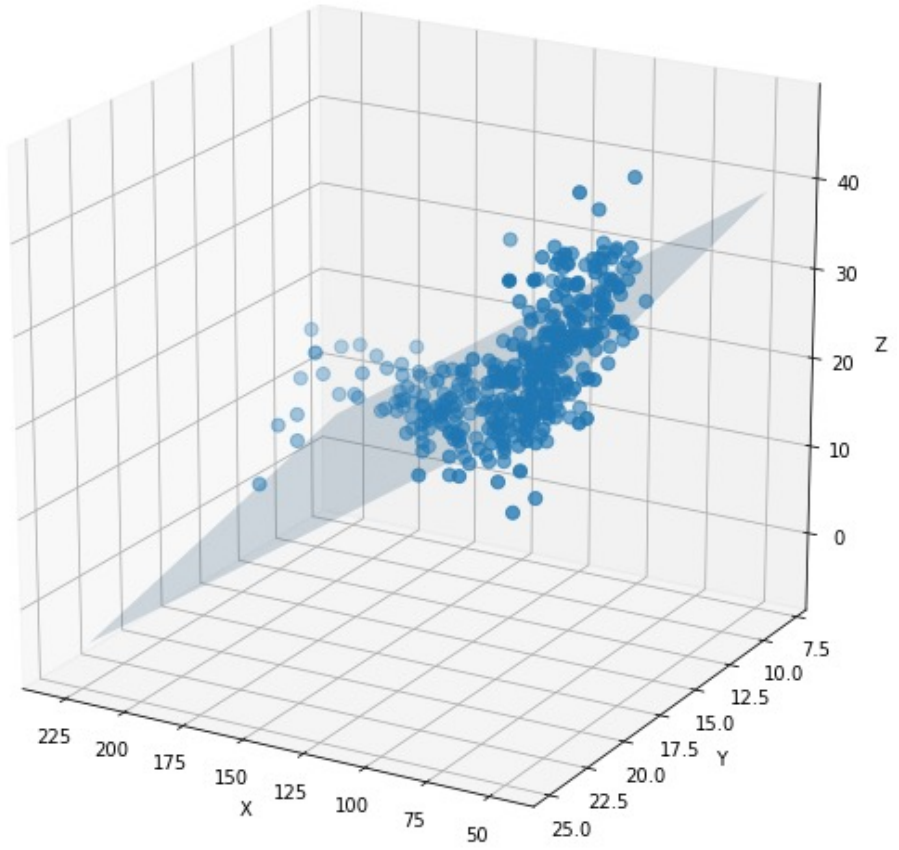
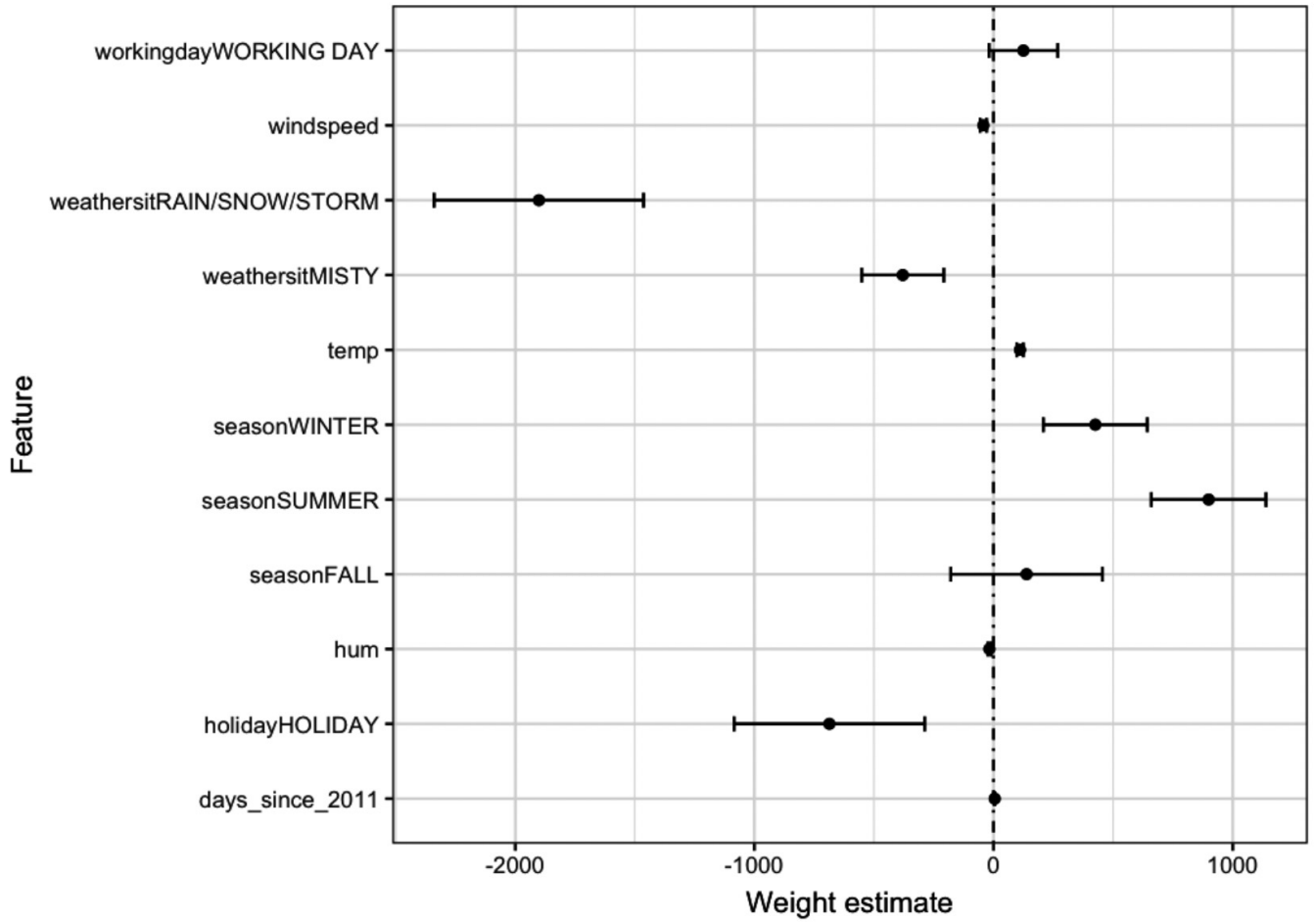


Image source: https://stackoverflow.com/questions/66098443/how-to-plot-3d-linear-regression

Interpretable Model



Visualization of feature importance of bike rental number prediction

Image source: Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

Interpretable Model

- Linear model provides promising interpretable results.
- However, its major problem is that it can only handle linear hand-crafted data.
- When the data is nonlinear, it cannot provide accurate prediction, on which the interpretability makes no sense.

Interpretable Model

Which model is naturally interpretable?

Decision tree

- Handle numerical and categorical data at the same time.
- The interpretation can be made by a series of “AND” statements.

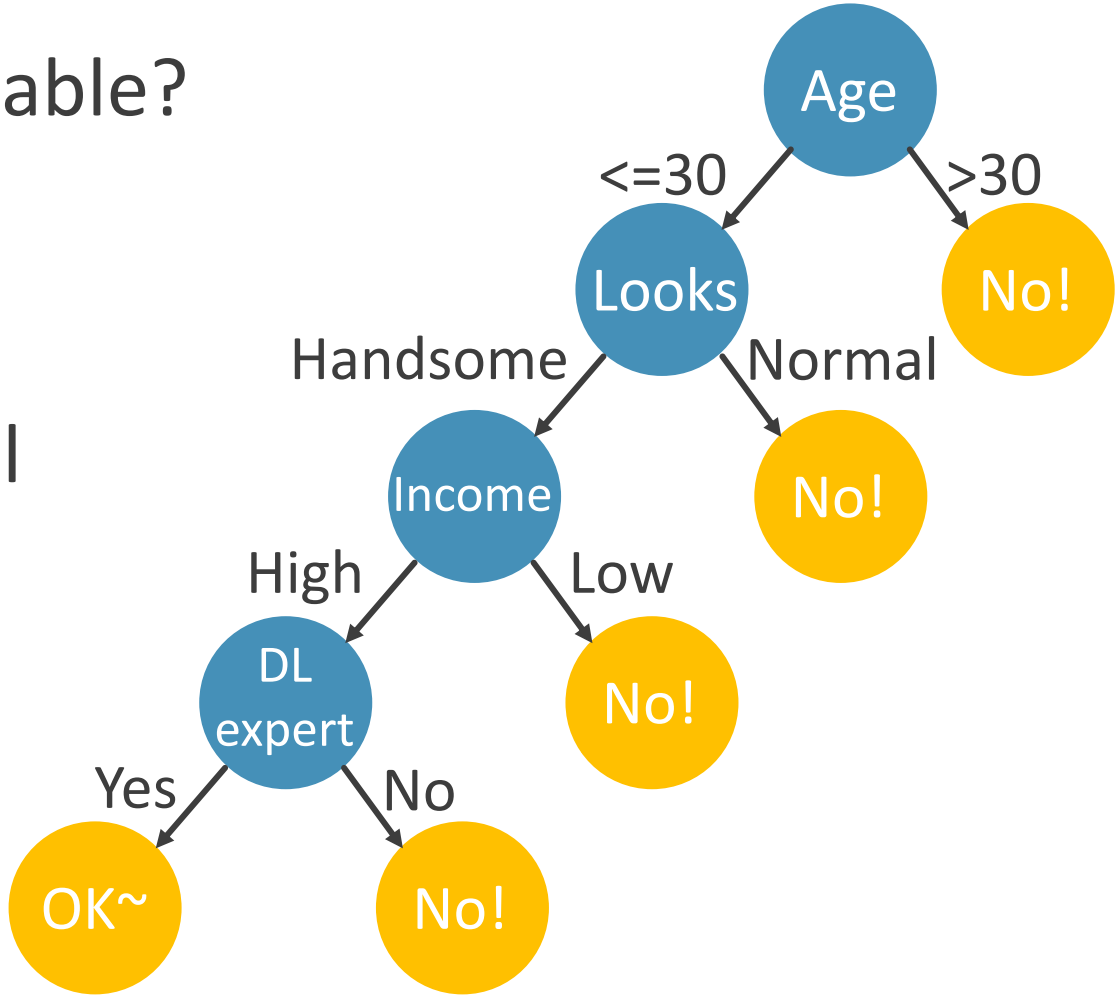


Image source: https://stackoverflow.com/questions/66098443/how-to-plot-3d-linear-regression

Model-Agnostic Methods

- If a model itself is not obviously interpretable, can we still make it interpretable?
- We can separate the explanations from the machine learning model.
  - We may design an interpretability method on some black-box models to explain it.
  - In this way, the machine learning developer is free to use any machine learning model, when the interpretability methods can be applied to any model.

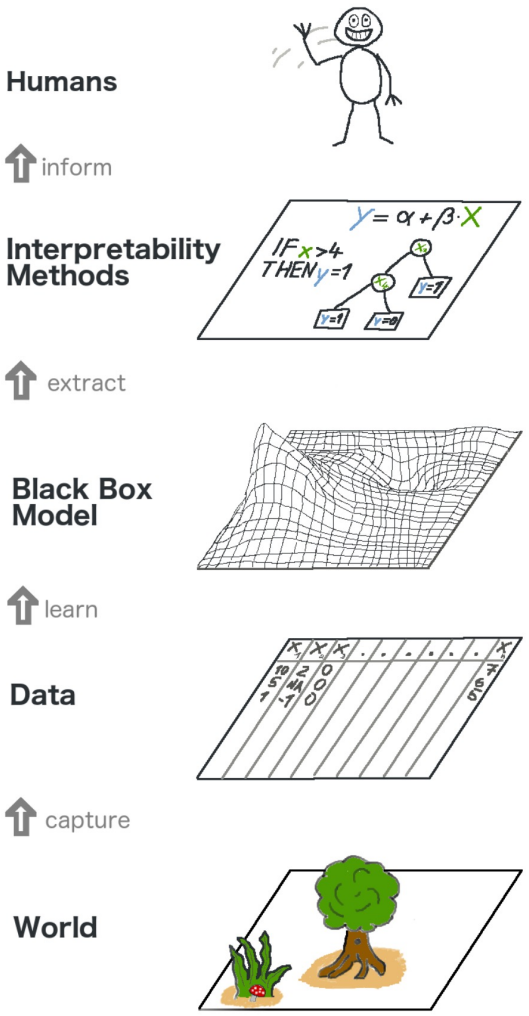


Image source: Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

LIME

" Why **should** i **trust you**?" Explaining the predictions of any classifier
MT Ribeiro, S Singh, C Guestrin - Proceedings of the 22nd ACM …, 2016 - dl.acm.org
… that require **trust**: deciding if one **should trust** a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier **should** not be **trusted**. …
☆ Save 99 Cite Cited by 15355 Related articles All 34 versions

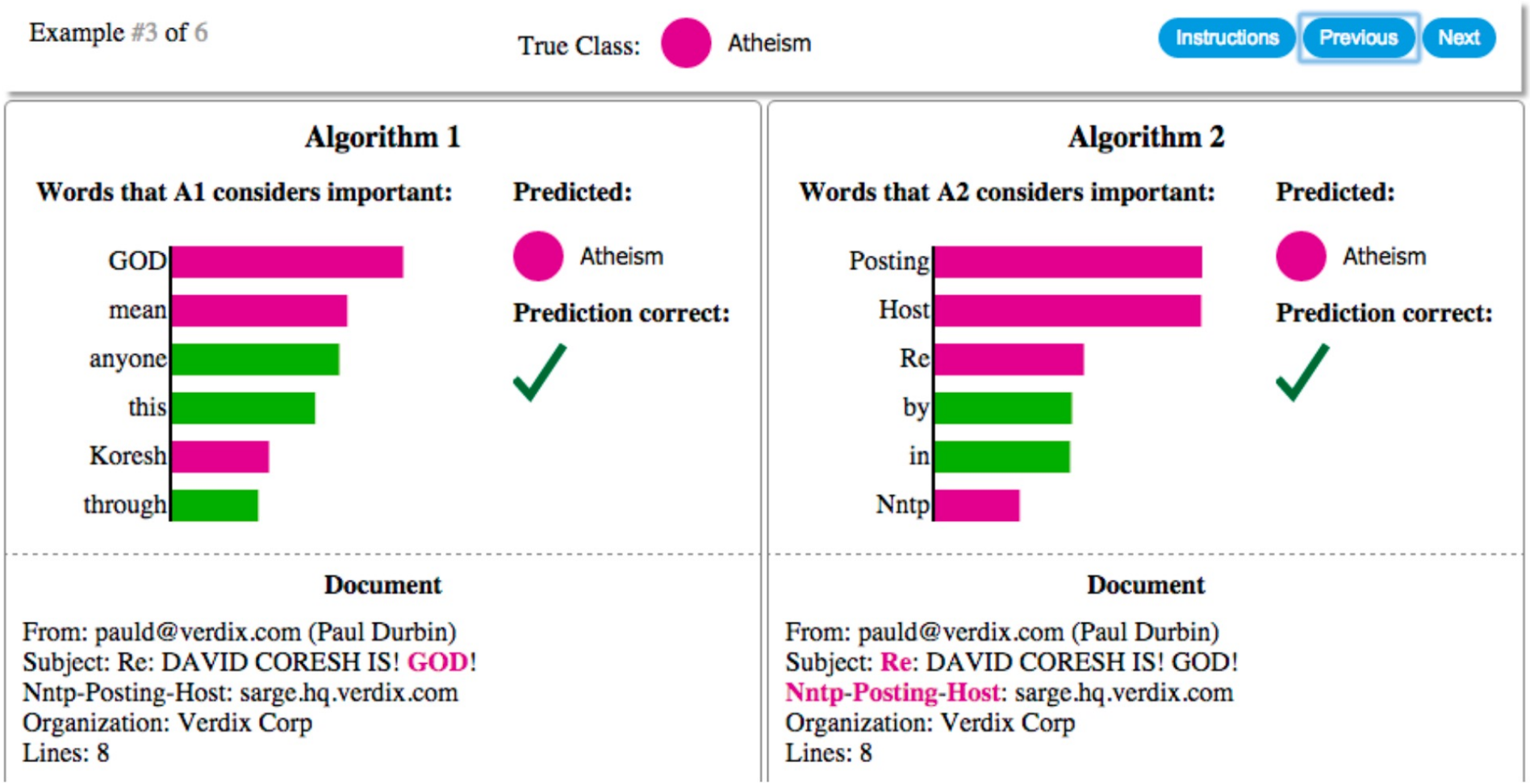
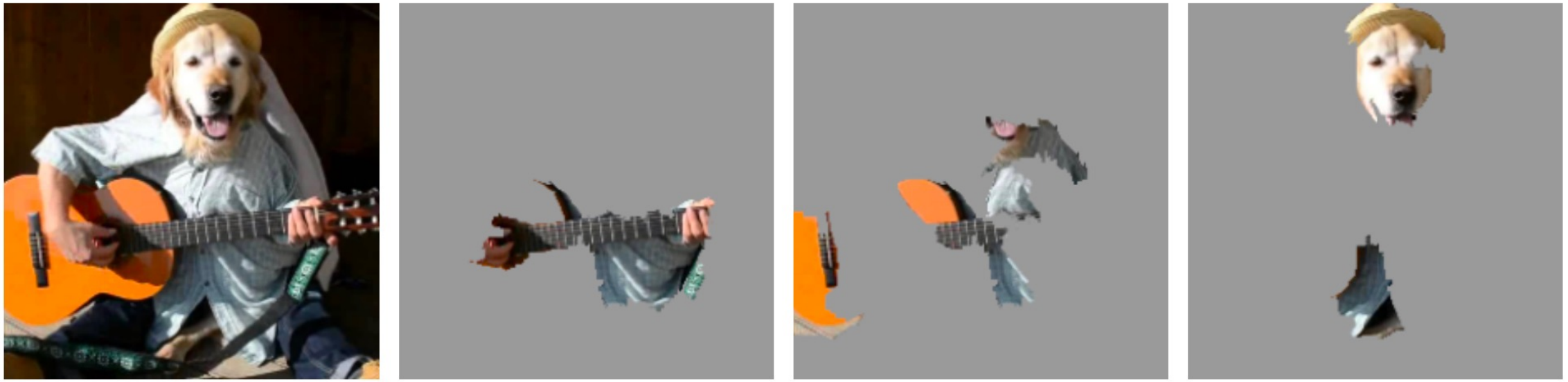


Image source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016.

LIME



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" (p = 0.32), "Acoustic guitar" (p = 0.24) and "Labrador" (p = 0.21)

Image source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016.

LIME

- Local Interpretable Model-agnostic Explanations (LIME) aims to identify an interpretable model (explainer) over the interpretable representation that is locally faithful to the classifier.
  - locally faithful: the explainer must correspond to how the model behaves in the neighborhood of the instance being predicted.
  - interpretable representation: the explained features should be in the form that human can understand.

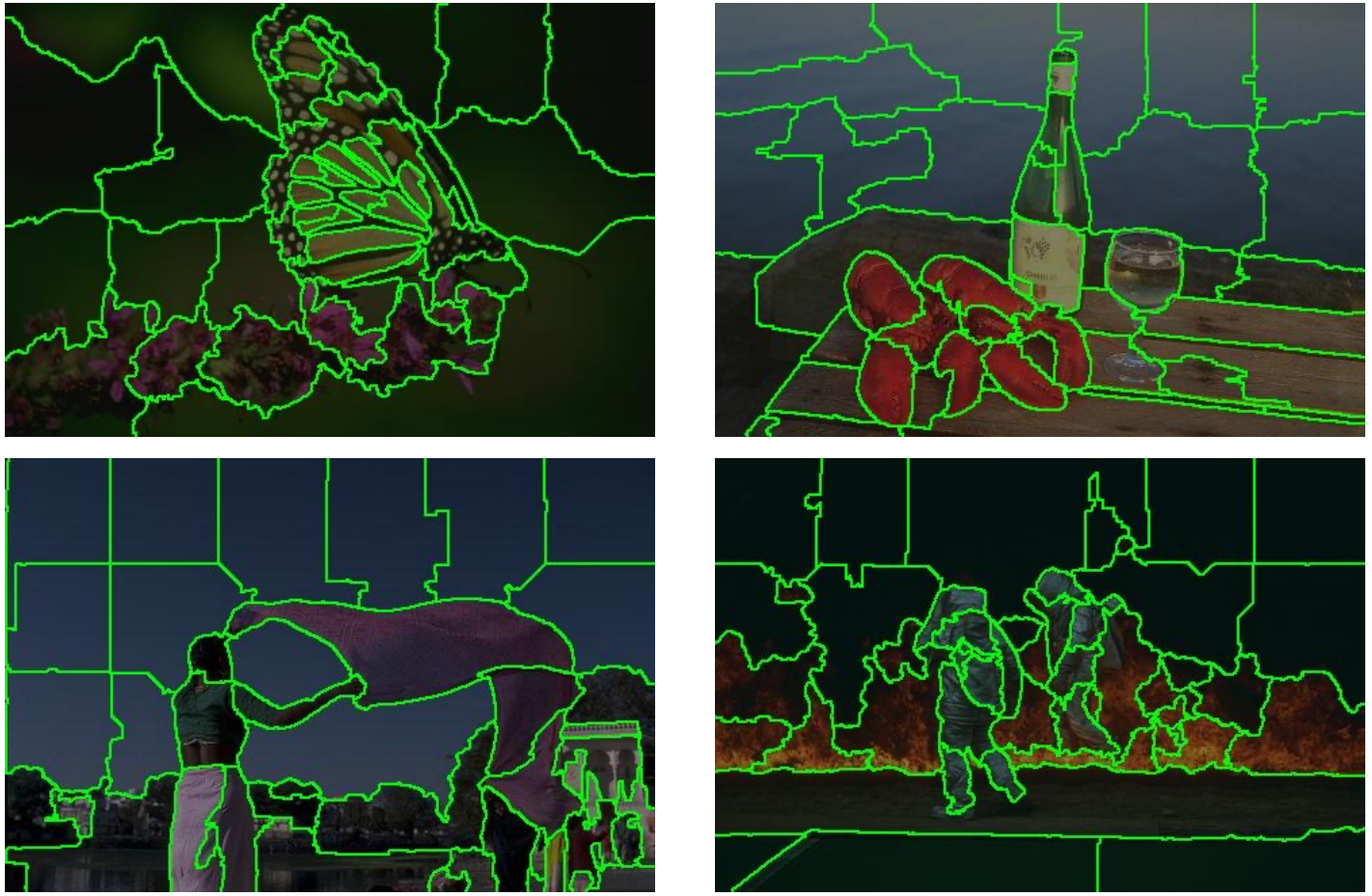
Local Fidelity

- The explainer can't behave like the model globally.
  - Otherwise, the explainer itself is the model.
- However, the explainer can behave like the model locally.
  - Global nonlinear hyperplane can be constructed by many linear local hyperplanes.
- Given a sample, the explainer is built for this sample locally. It can tell you which features are the most important ones to distinguish it from its neighbors.

Interpretable Representation

- The features that are meaningful to humans are usually not embeddings.
  - If the explainer tells you the 91's dimension of the last feature map is important to classify dog, so what?
- Interpretable representation is necessary.
  - For text classification, absence or presence of a particular word.
  - For image classification, absence or presence of a patch of pixels.
- We denote $x \in \mathbb{R}^d$ be the original representation of an instance being explained, and we use $x' \in \{0,1\}^{d'}$ to denote a binary vector for its interpretable representation.

Interpretable Representation



Superpixel segmentation is used as interpretable representation for images.

Image source: https://cvlab.postech.ac.kr/research/mmlp/

LIME

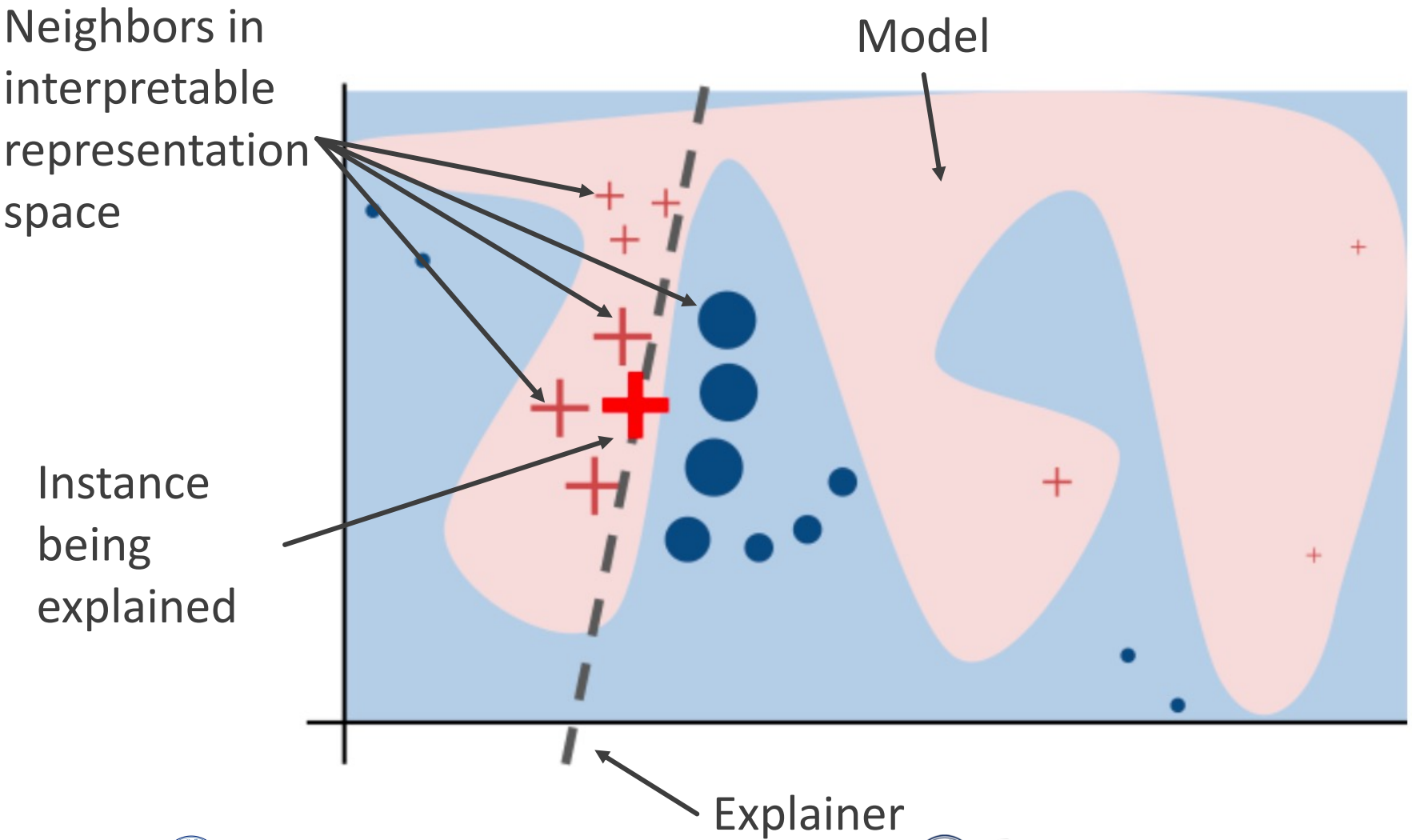


Image source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016.

LIME

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

Image source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016.

Fidelity-Interpretability Trade-off

- The objective to optimize the explainer g:

$$\underset{g}{\operatorname{argmin}}\, L(f, g, \pi_x) + \Omega(g)$$

Make g predict like f based on π_x.

$$L(f, g, \pi_x) = \sum_{z \in \mathrm{N}(x)} \pi_x(z)\big(f(z) - g(z')\big)^2$$

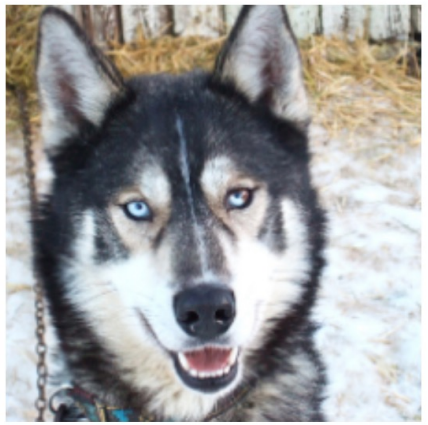
  - $L(f, g, \pi_x)$ measures how unfaithful g is in approximating f in the locality defined by π_x.
  - $\Omega(g)$ is the complexity of g. More complex, less interpretable.

Explanations Lead to Insights

Experiment

- Train a binary classifier with
  - wolve with snow background;
  - husky without snow background.
- Test a husky with snow background.
- Survey on students in a machine learning course.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27



(a) Husky classified as wolf

(b) Explanation

Image source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016.

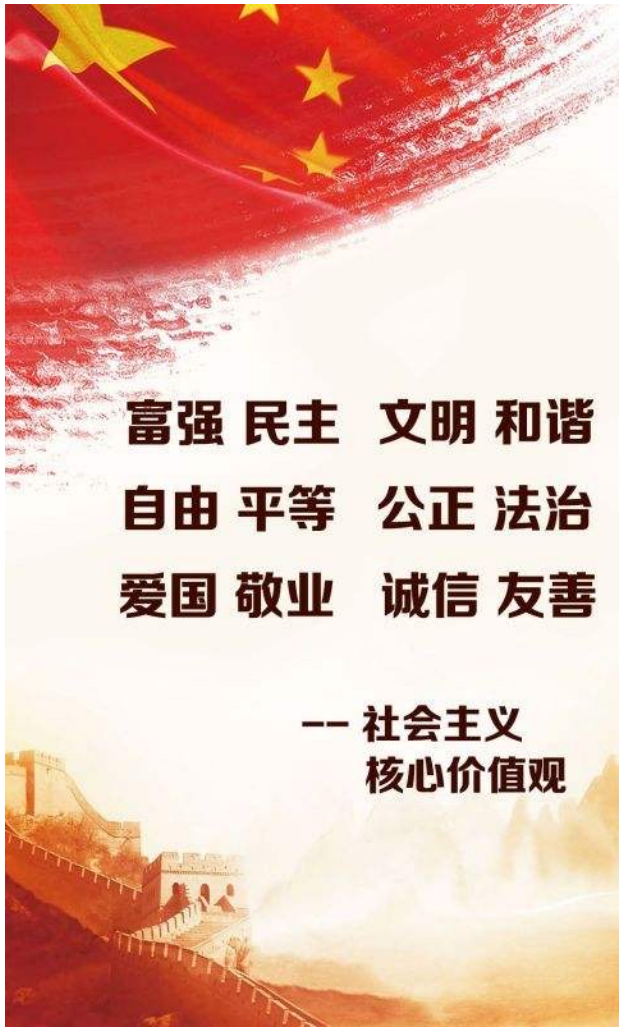
Model Interpretation

- Although most of the deep learning models are still not interpretable, studying on them is still meaningful.
- However, studying model interpretation makes it more clear.
- If a model with both high accuracy and high interpretability, it will be the king of models.

FAIRNESS

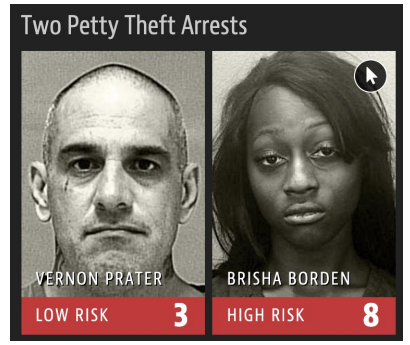
Fairness

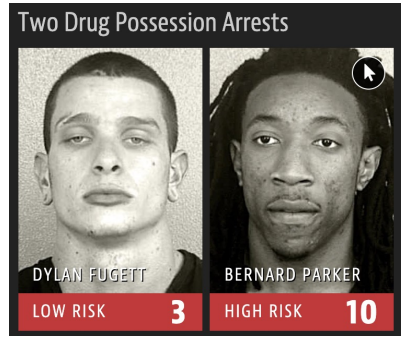
- There are biases when people make decisions.
  - For example, the recruitment decision should mainly reflect the candidates' ability, rather than race, gender, age, marriage, birthplace, and so on.
- In the context of decision-making, fairness is the *absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics*.
  - Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people.



Fairness

- The software, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), measures the risk of a person to recommit another crime.
  - Judges use COMPAS to decide whether to release an offender, or to keep him or her in prison.
- An investigation into the software found: COMPAS is more likely to assign a higher risk score to African-American offenders than to Caucasians with the same profile.





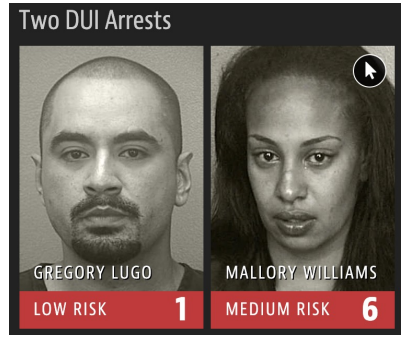


Image source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bias in Data

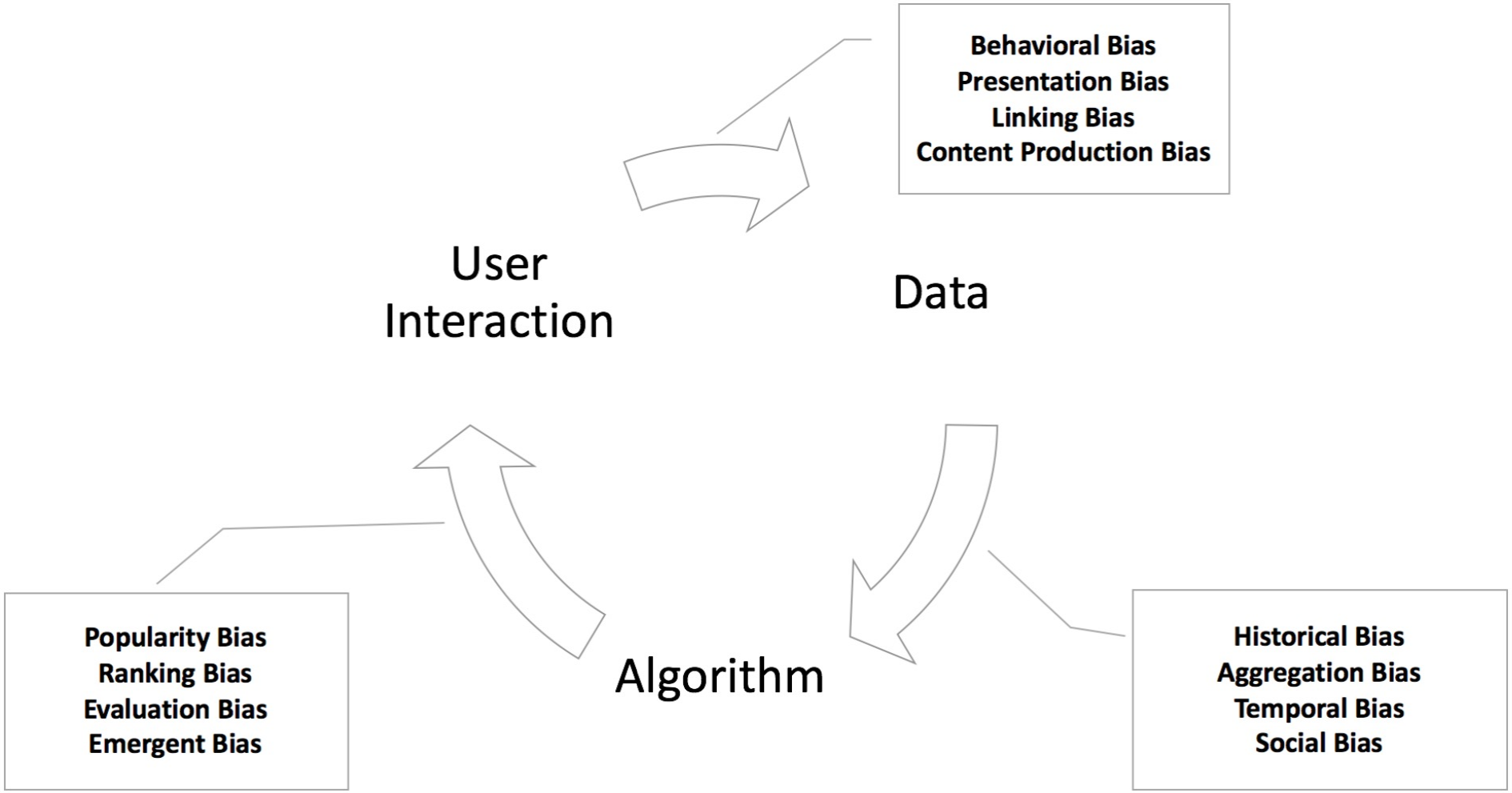


Image source: Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." ACM Computing Surveys (CSUR) 54, no. 6 (2021): 1-35.

Bias in Data

- From data to algorithm:
  - Historical bias: the already existing bias.
  - Aggregation bias: false conclusions are drawn for a subgroup based on observing other different subgroups.
  - Temporal bias: differences in populations and behaviors over time.
  - Social bias: other people's actions or content coming from them affect our judgment.

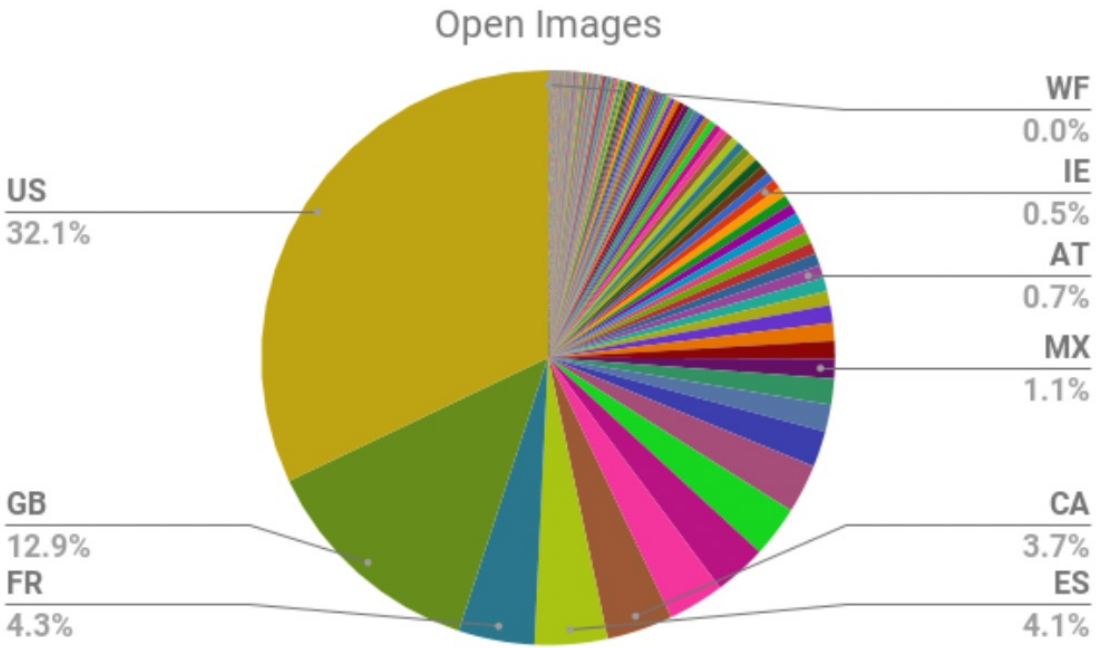
Bias in Data

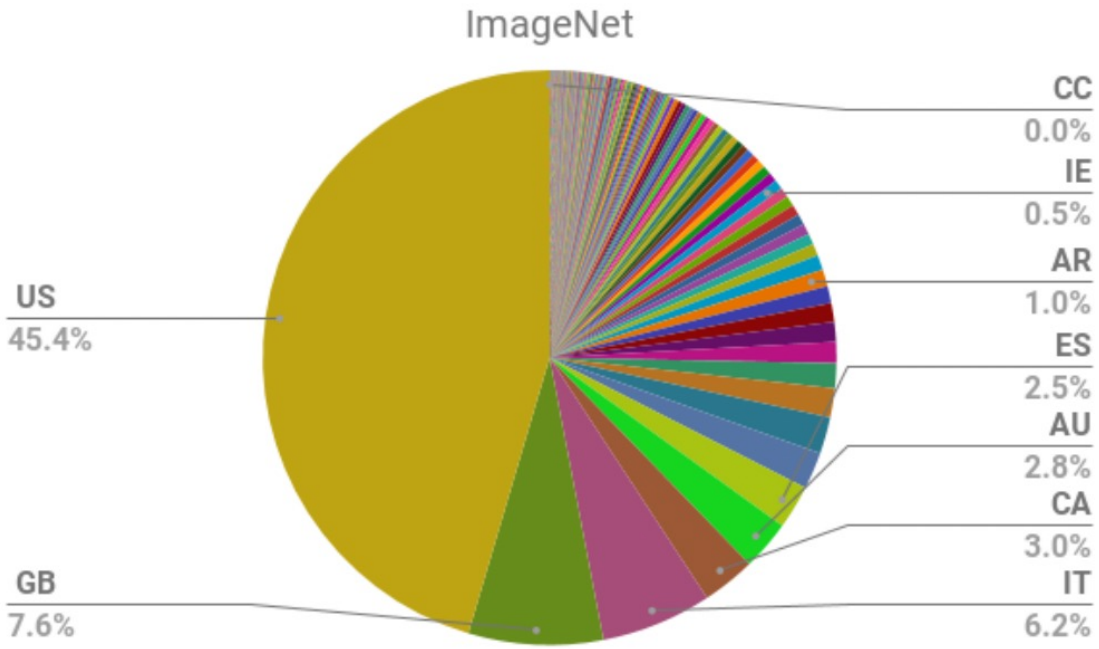
- From algorithm to user:
  - Population bias: statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset from the original target population.
  - Ranking bias: top-ranked results are the most relevant and important will result in attraction of more clicks than others.
  - Evaluation bias: use of inappropriate and disproportionate benchmarks for evaluation of applications.
  - Emergent bias: change in population, cultural values, or societal knowledge usually some time after the completion of design.

Bias in Data

- From user to data:
  - Behavioral bias: different user behavior across platforms, contexts, or different datasets.
  - Presentation bias: information presented in different ways draws different attention.
  - Linking bias: network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users.
  - Content production bias: different structural, lexical, semantic, and syntactic differences in the contents generated by users.

Bias in Data





The distribution of ImageNet differs from the open images.

Image source: Shankar, Shreya, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. "No classification without representation: Assessing geodiversity issues in open data sets for the developing world." arXiv preprint arXiv:1711.08536 (2017).

Definitions of Fairness

- Group Fairness: Treat different groups equally.
- Individual Fairness: Give similar predictions to similar individuals.
- We call the attribute that may produce bias the protected attribute or protected variable.

Definitions of Fairness

Definition (Demographic Parity)

A predictor $\hat{Y}$ satisfies demographic parity if:

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

- The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group.
  - E.g. for a job position, the probability that the model predicts one can obtain an offer is equal for male and female.

Definitions of Fairness

Definition (Equalized Odds)

A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y:

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

- The protected and unprotected groups should have equal true positives and false positives.
  - E.g. for a job position, the probability that the predictor makes mistakes is same for both groups.

Definitions of Fairness

Definition (Fairness Through Awareness)

An algorithm is fair if it gives similar predictions to similar individuals.

Definition (Fairness Through Unawareness)

An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process.

Methods for Fair Machine Learning

- Generally, methods that target biases in the algorithms fall under three categories:
  - Pre-processing techniques try to transform the data so that the underlying discrimination is removed.
  - In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process.
  - Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model.

Fair Classification

- For a classification problem, to make the prediction fair for equalized odds on an attribute A, one can formulate it as an optimization problem:

$$\begin{aligned} \min_{p} \quad & \mathbb{E}l(\hat{Y}_p, Y) \\ s.t. \quad & \gamma_0(\hat{Y}_p) = \gamma_1(\hat{Y}_p) \\ & \forall_{y,a} 0 \le p_{ya} \le 1 \end{aligned}$$

where $\gamma_a(\tilde{Y})$ is a pair of true positive and false positive.

$$\gamma_a(\tilde{Y}) = \left(P(\hat{Y} = 1 | A = a, Y = 0), P(\hat{Y} = 1 | A = a, Y = 1)\right)$$

Reference: Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016): 3315-3323.

Adversarial Debiasing

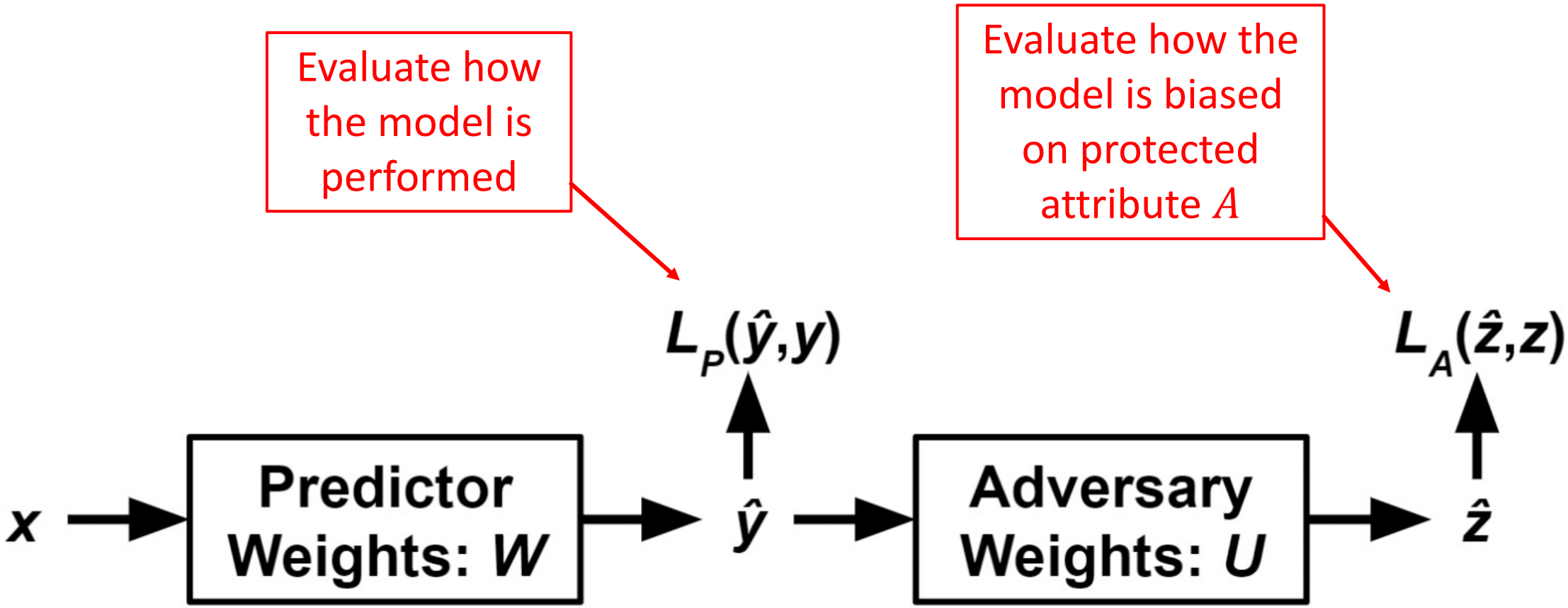


Image source: [1] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," AIES 2018 - Proc. 2018 AAAI/ACM Conf. AI, Ethics, Soc., pp. 335–340, 2018.

PRIVACY

Privacy in Deep Learning

- In the process of deep learning, some part of information may be sensitive.
  - This information can be the training data, inference queries or model parameters or hyperparameters.
- If we assume that information cannot be attained directly, there is still the threat of information exposure through inference, indirectly.
  - Black-box model is not safe either.

Threats: Membership Inference

- A membership inference attack speculates whether or not the given data sample x has contributed to the training step of the target model.
- Deep learning models have memorization effect, such that a trained sample will be correctly classified with higher probability.
- One simple way to attack:
  - Generate dataset D from the black-box model.
  - Create two models f and f', trained on D and $D \cup \{x\}$.
  - Compare the confidence score of $f(x)$ and $f'(x)$.

Threats: Attribute Inference

- Attribute inference attacks are against attribute privacy
- An attacker tries to infer sensitive attributes of given data instances from a released model and the instance's non-sensitive attributes.
  - For example, the attackers can invert the linear regression model of a medicine dosage prediction task.
  - They recover genomic information about the patient, based on the model output and several other non-sensitive attributes (e.g., height, age, weight).

Threats: Model Inversion

- Given white-box access to a neural network, model inversion could extract instances of training data, from observed model predictions.

The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score

Threats: Model Stealing

- A model stealing attack is meant to recover the model parameters via black-box access to the target model.
  - Their attack tries to find parameters of the model through equation solving, based on pairs of input-outputs.
- A Hyperparameter stealing attack tries to find the hyperparameters used during the model training, such as the regularization coefficient or model architecture.

Privacy-Preserving Mechanisms

- Data aggregation: collect data and form datasets, while preserving the privacy of the contributors.
- Training phase: make the training process of models private so that sensitive information about the participants of the training dataset would not be exposed.
- Inference phase: protect the privacy of users of deployed models, who send their data to a trained model for having a given inference service carried out.

Differential Privacy

- If we can't query a single record x, but we can query a group of records D and D', where $D = D' \cup \{x\}$, we can use the information difference to decide if x is in the database.
- This is called differential attack that aims at membership inference.

Definition ε-Differential Privacy (ε-DP).

For $\varepsilon \geq 0$, an algorithm A satisfies ε-DP if and only if for any pair of datasets D and D' that differ in only one element:

$$P(A(D) = t) \leq e^{\epsilon} P[A(D') = t]$$

- $P(A(D) = t)$ denotes the probability that the algorithm A outputs t.

Differential Privacy

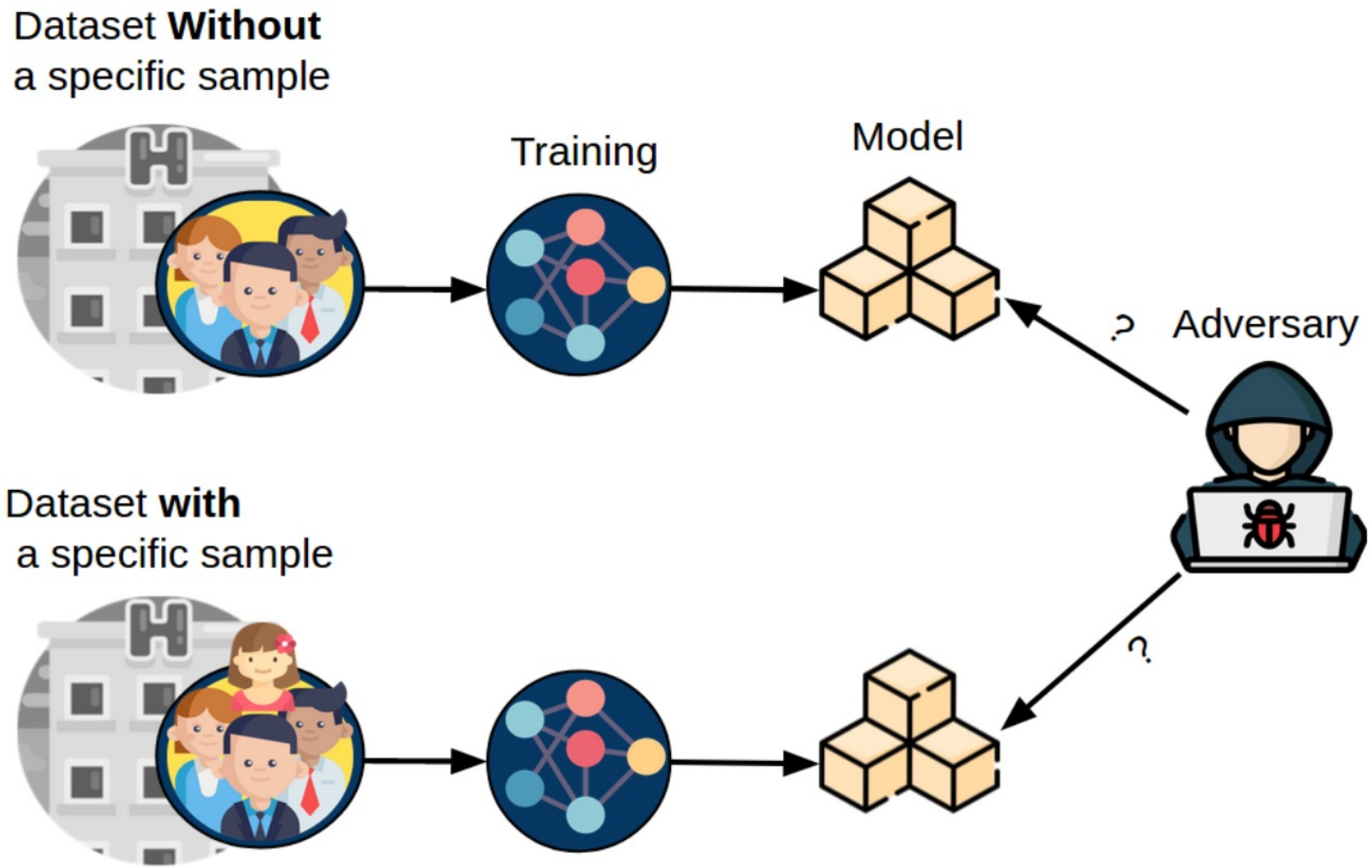
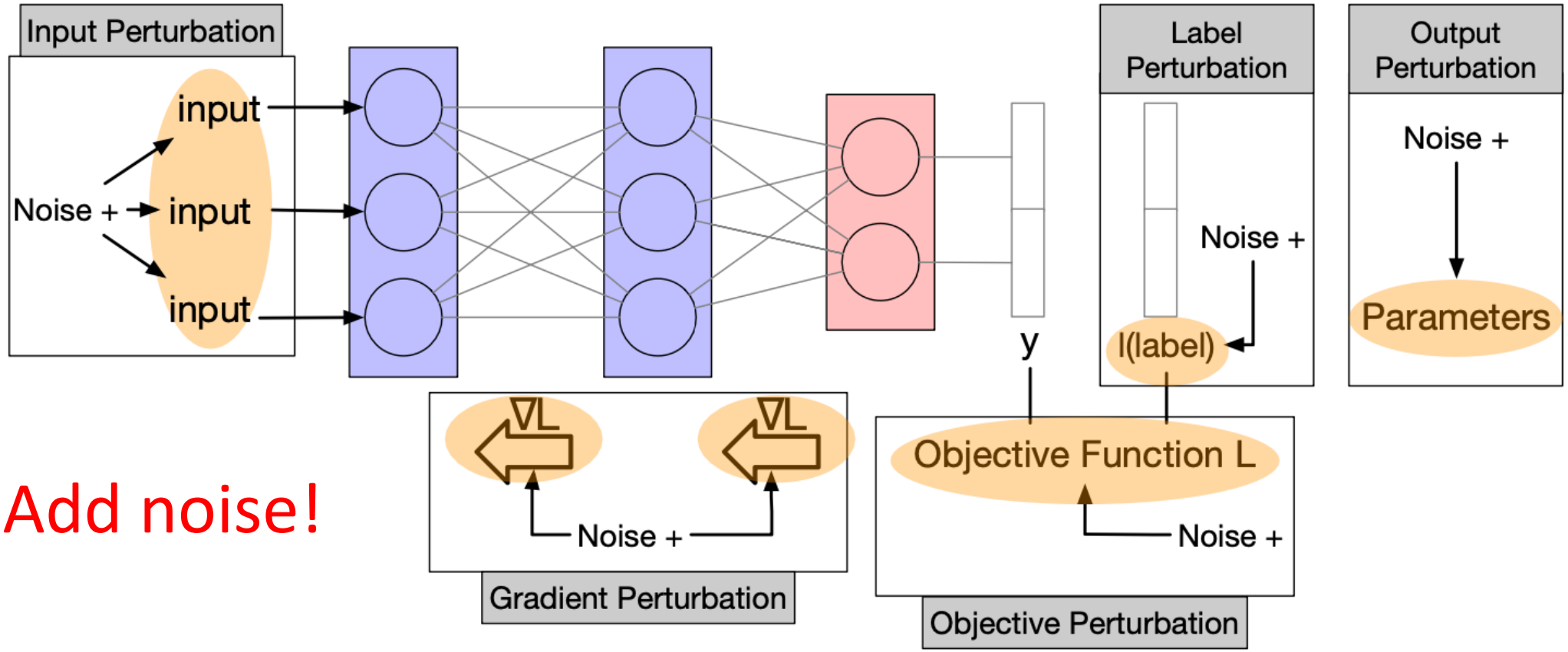


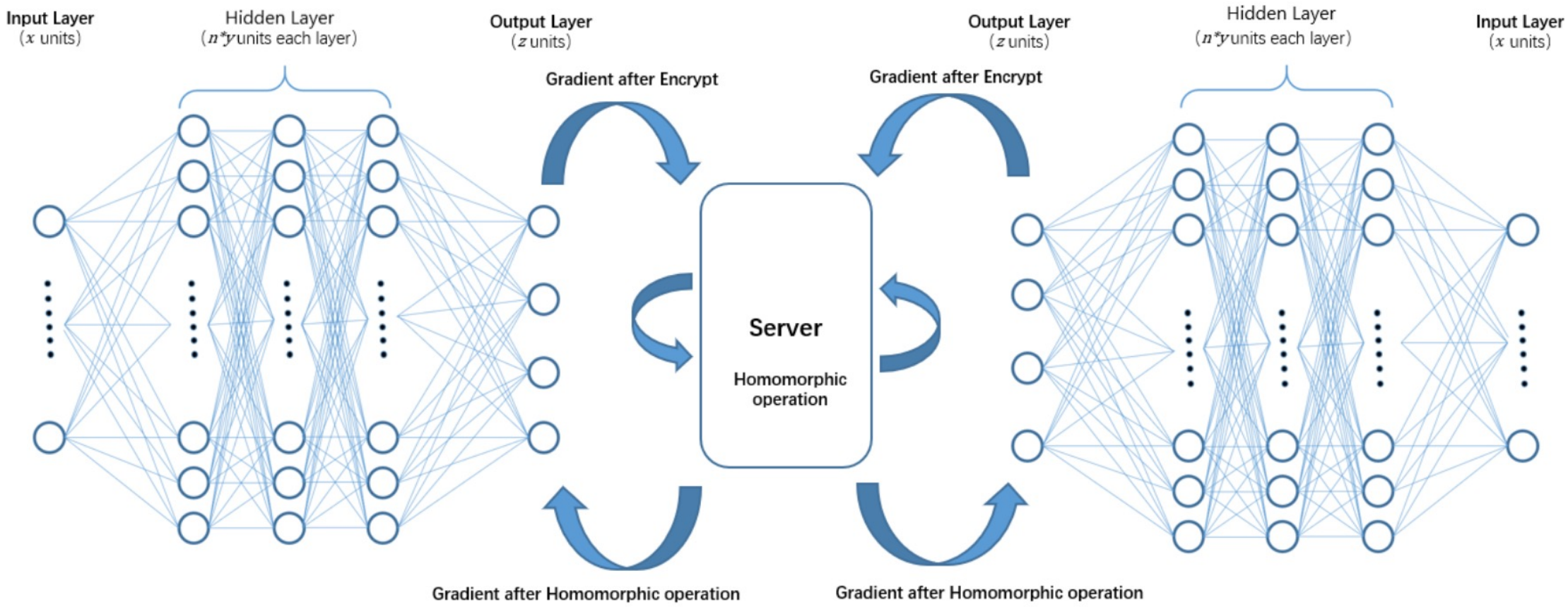
Image source: Torkzadehmahani, Reihaneh, Reza Nasirigerdeh, David B. Blumenthal, Tim Kacprowski, Markus List, Julian Matschinske, Julian Späth et al. "Privacy-preserving Artificial Intelligence Techniques in Biomedicine." arXiv preprint arXiv:2007.11621 (2020).

Differential Privacy

- How to make a model ϵ-DP?



Homomorphic Encryption



Homomorphic encrypted federated learning

Image source: Fang, Haokun, and Quan Qian. "Privacy Preserving Machine Learning with Homomorphic Encryption and Federated Learning." Future Internet 13, no. 4 (2021): 94.

Secure Multiparty Computation

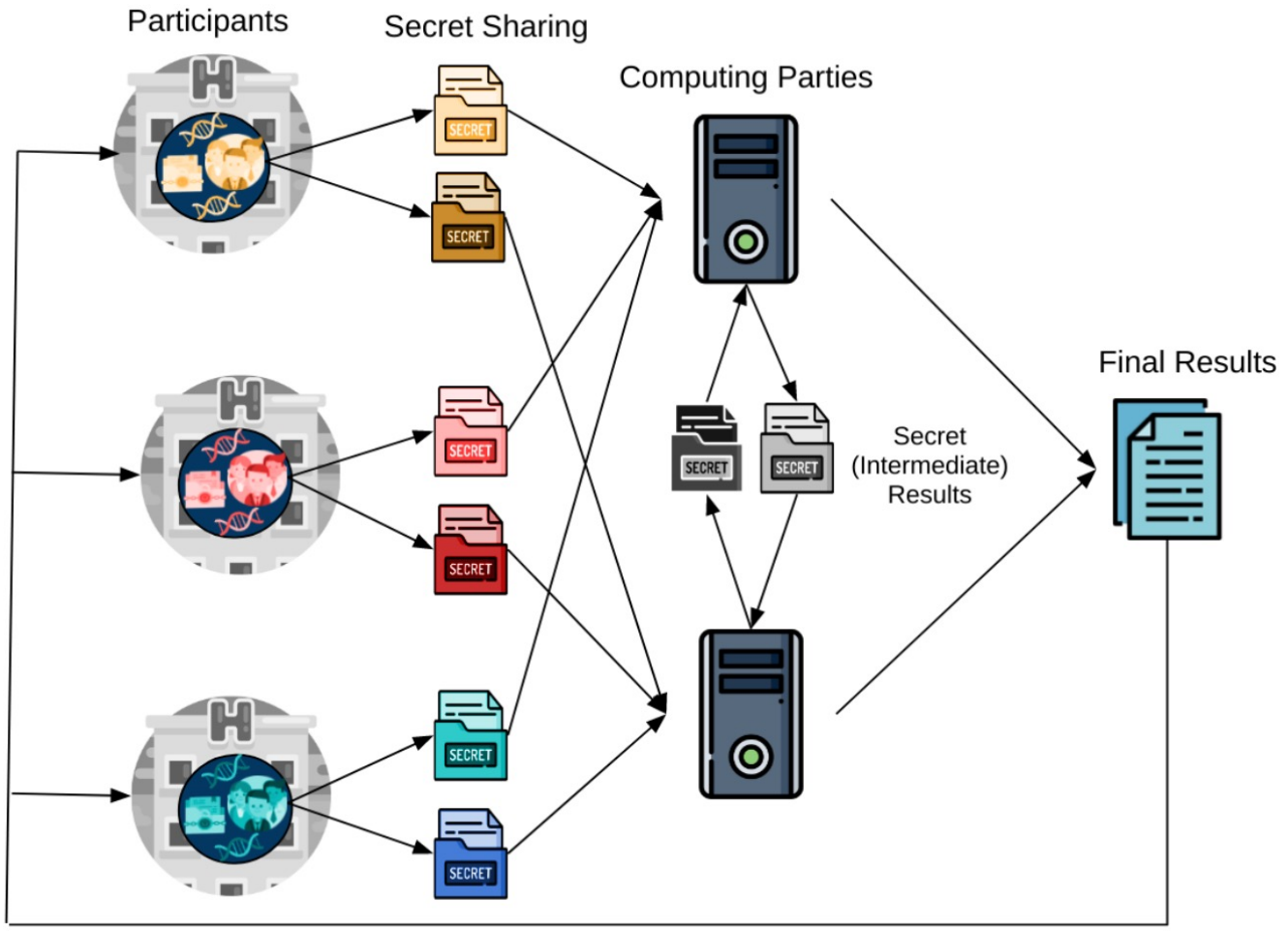


Image source: Torkzadehmahani, Reihaneh, Reza Nasirigerdeh, David B. Blumenthal, Tim Kacprowski, Markus List, Julian Matschinske, Julian Späth et al. "Privacy-preserving Artificial Intelligence Techniques in Biomedicine." arXiv preprint arXiv:2007.11621 (2020).

Conclusion

After this lecture, you should know:

- What is adversarial training and how to do it.
- What is the key principle of knowledge distillation.
- What is model interpretation and why do we need it.
- What is fairness and how to measure it?
- Why does privacy concern in machine learning?

Reference

- Recent Advances in Adversarial Training for Adversarial Robustness
- Knowledge Distillation: A Survey
- Interpretable Machine Learning
- A survey on bias and fairness in machine learning
- Privacy in deep learning: A survey

Thank you!

- Any question?
- Don’t hesitate to send email to me for asking questions and discussion. ☺