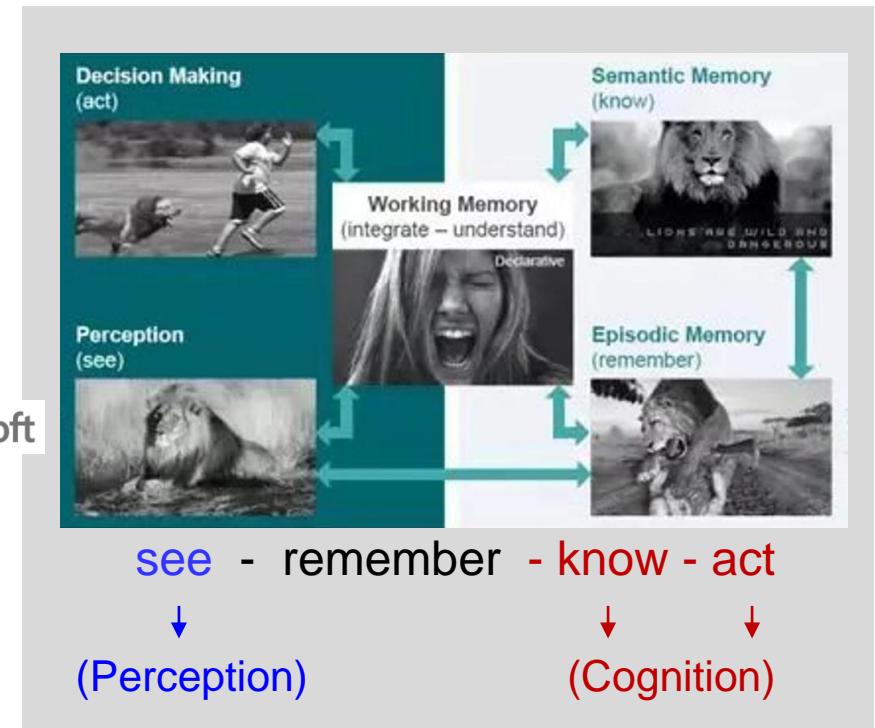
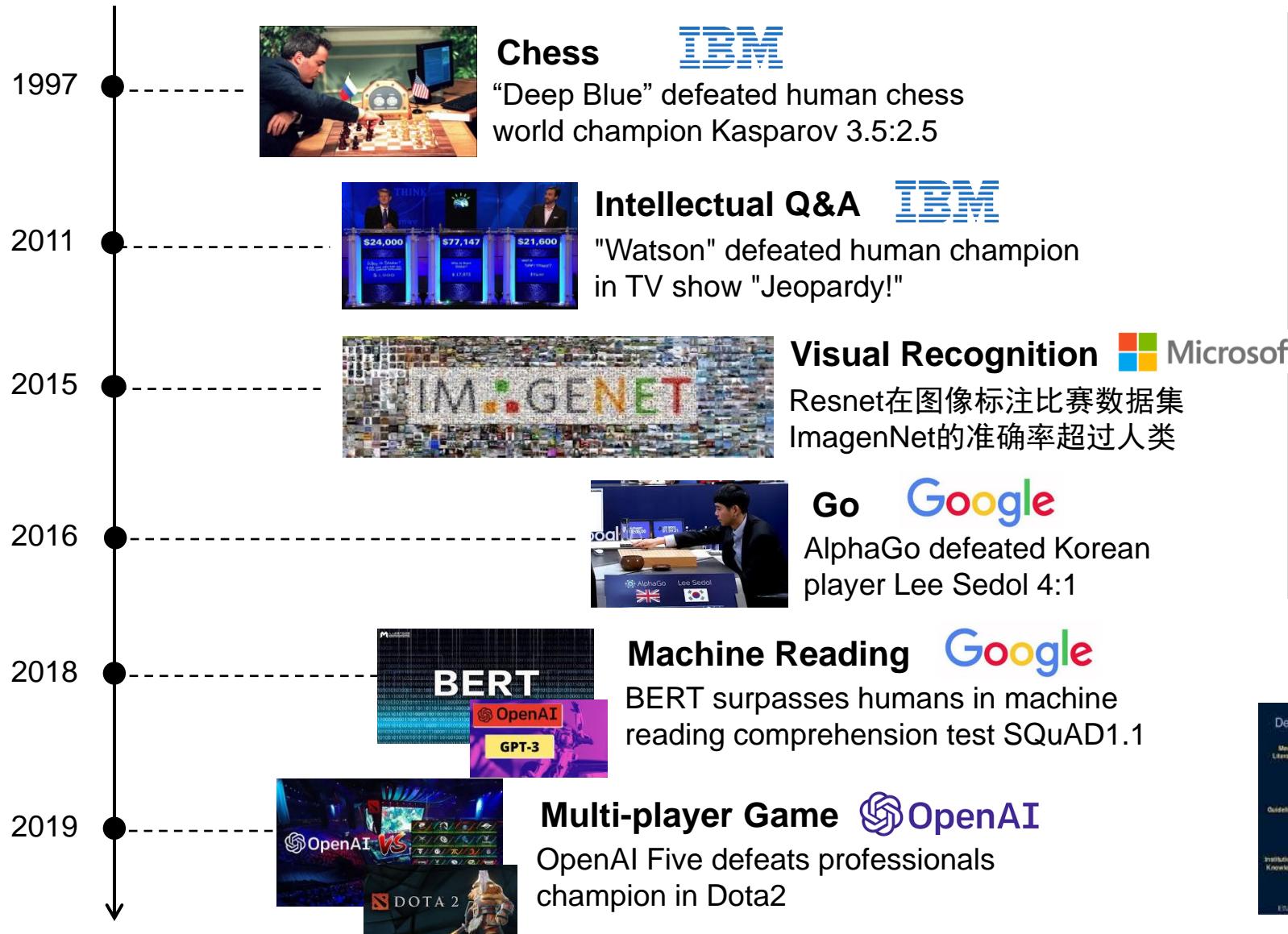


ACM Multimedia 2021 tutorial

Trustworthy Multimedia Analysis

Chengdu, Oct.20, 2021

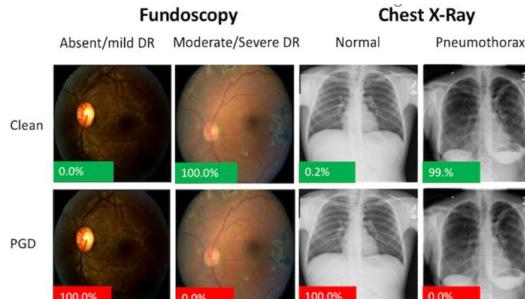
Machine vs Human: “CANNOT Use” → “OK to Use”



"OK to Use" ≠ "Good to Use"

Medical Diagnosis

poor Robustness



perturbation alters diagnostic results positive → negative

Unmanned Driving



trivial perturbation changes sign recognition result

violating Ethics & Commonsense

STAT Sections Topics Multimedia Newsletters More Q
EXCLUSIVE
IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show
量子位

Watson delivers bleeding drugs to patients with bleeding cancer



pedestrian detector is sensitive to gender, race, physical conditions

inefficiency to Test & Debug

babylon GP at hand
STAT Sections Topics Multimedia Newsletters More Q
EXCLUSIVE
IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show
量子位

Babylon cannot localize and rectify error with repeated complaints



accident error occurs without tracing and debugging

Barbarian at the Gate



Intelligent Medicine (Medical Experts)



Intelligent Medicine Chair
Prof. Zhaodong Wang
Director of National center for clinical research in geriatric diseases



Intelligent Medicine Chair
Prof. Xu Zhang
Dean of Capital Medical University School of Biomedical Engineering



Prof. Jitao Sang
Deputy Director of Department of computer science, Beijing Jiaotong University



Prof. Feng He
Vice President of Department of Medical Science and Engineering, Tianjin University



Prof. Gang Zhao
Neurology Director of Xijing Hospital of the Fourth Military Medical University



Zhenzhou Wu
CTO of Artificial Intelligence Department, National Center for Clinical Research in Neurological Diseases

Intelligent Medicine (Computer Experts)



Intelligent Medicine Chair
Shaoliang Peng
Professor of Hunan University



Intelligent Medicine Chair
Di Zhao
Associate Professor of Institute of Computing Technology, CAS



Yanchun Zhang
Professor of Victoria University, Thousand Talents Program



Wenli Cai
Director of Massachusetts General Hospital



郑海荣
中国科学院深圳先进技术研究院副院长、研究员



Jian Wu
Professor, Director of Medical Artificial Intelligence Research Center, Zhejiang University

Barbarian at the Gate

Google health



David Feinberg
former UCLA hospital CEO

VS



Greg Corrado
Tech Senior Director



Jeff Dean
Google Chief Architect, leads
MapReduce/TensorFlow

Official: Google Health shuts down because it couldn't scale adoption

Forbes

By Brian Dolan | June 24

Google has officially announced plans to shutter Google Health, its personal health records platform, come January 1. Data stored in Google Health will continue to be available.

Aug 21, 2021, 01:08pm EDT | 92,844 views

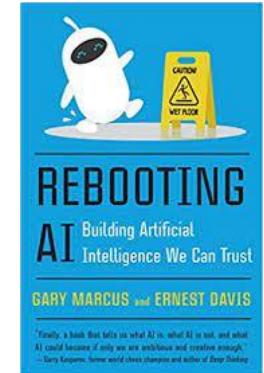
Google Dismantling Health Division

Global Edition Operations

Google dismantles health division in strategy overhaul

Expectation vs Reality

“Our biggest fear is not that machines will seek to obliterate us or turn us into paper clips; it’s that **our aspirations for AI will exceed our grasp [1]**.”



expectation



reality

[1] Gary Marcus, Ernest Davis. «Rebooting AI: Building Artificial Intelligence We Can Trust»

from Accuracy to Interpretation

Deep learning thus far

- 3.1. is data hungry
- 3.2. is shallow & has limited capacity for transfer
- 3.3. has no natural way to deal with hierarchical structure
- 3.4. has struggled with open-ended inference
- 3.5. is not sufficiently transparent
- 3.6. has not been well integrated with prior knowledge
- 3.7. cannot inherently distinguish causation from correlation
- 3.8. presumes a largely stable world
- 3.9. its answer often cannot be fully trusted
- 3.10. is difficult to engineer with

TODAY'S PAPER

arXiv.org > cs > arXiv:1801.00631

Computer Science > Artificial Intelligence

Deep Learning: A Critical Appraisal

Gary Marcus

(Submitted on 2 Jan 2018)



Gary Marcus, scientist, bestselling author, entrepreneur, and AI contrarian, was CEO and Founder of the machine learning startup Geometric Intelligence, recently acquired by Uber.

As a Professor of Psychology and Neural Science at NYU, he has published extensively in fields ranging from human and animal behavior to neuroscience, genetics, and artificial intelligence, often in leading journals such as Science and Nature.

» Terry Taewoong Um (terryt.um@gmail.com)



Gary Marcus. 2018. Deep learning: A critical appraisal. (2018).

Interpretable ML Symposium

NIPS 2017

7 December, Long Beach, California

Debate

Time

8:30pm - 9:30pm on Thursday, December 7th (Hall C). [Video of debate.](#)

Proposition

Interpretability is necessary in machine learning

Participants

Team A [for the proposition]: Rich Caruana (A1), Patrice Simard (A2)

Team B [against the proposition]: Kilian Weinberger (B1), Yann LeCun (B2)



from Accuracy to Interpretation

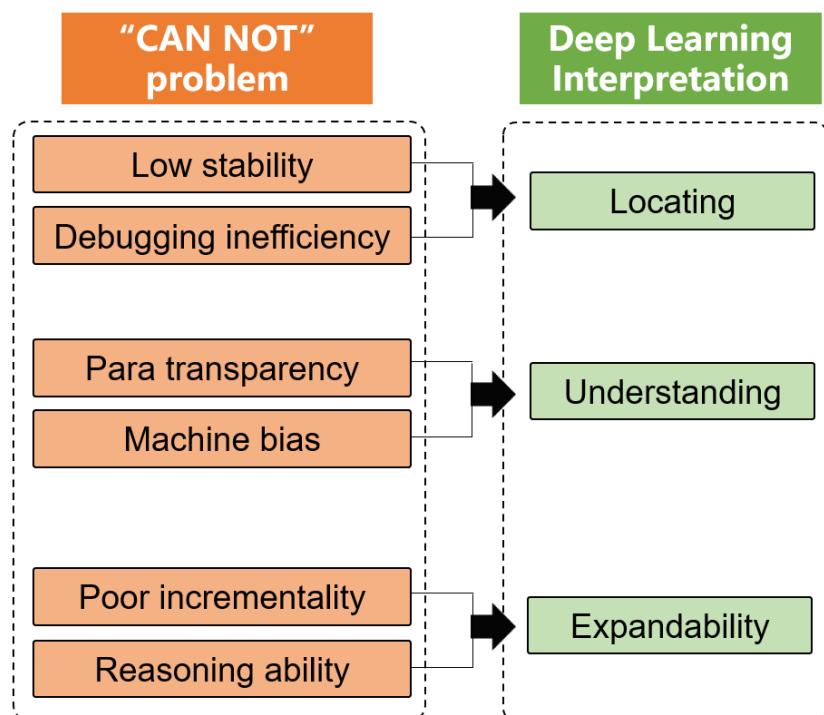
TUTORIAL

Deep Learning Interpretation

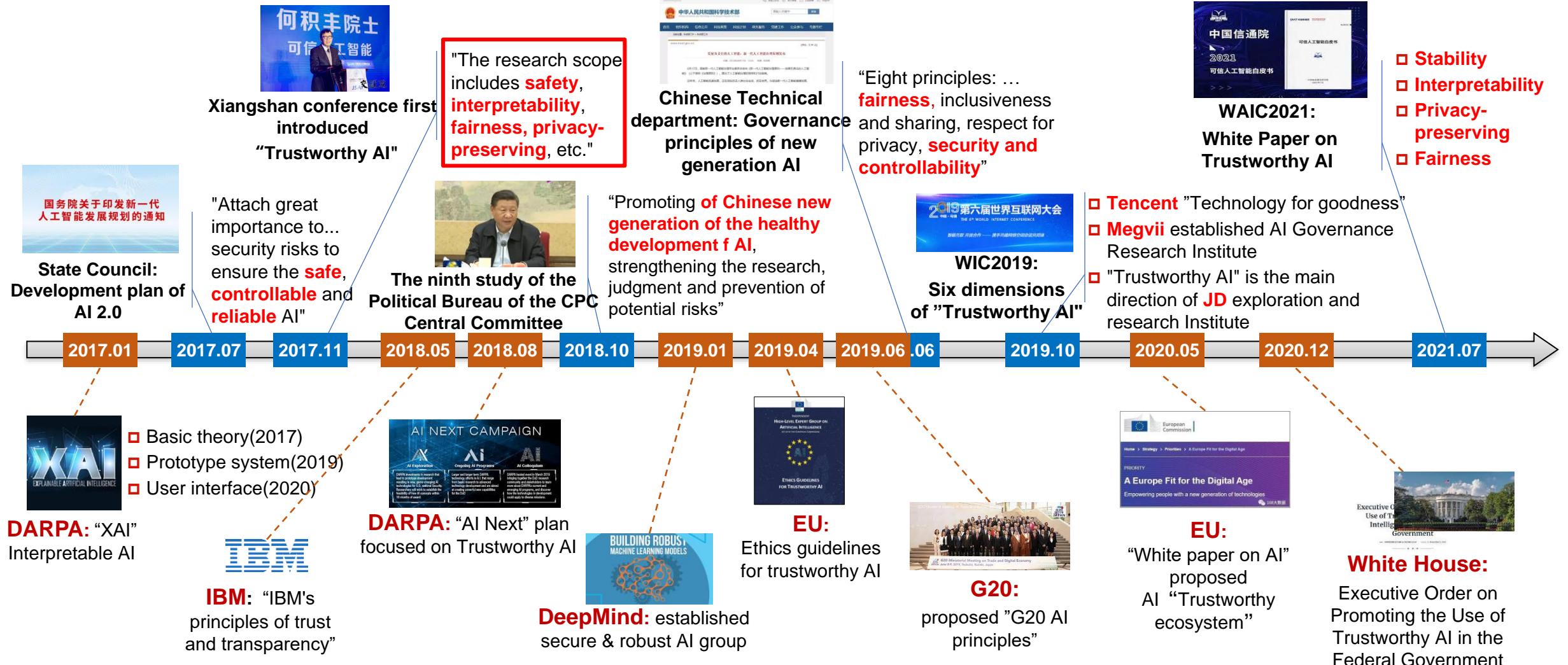
Jitao Sang

October 2018, pp 2098–2100 • <https://doi.org/10.1145/3240508.3241472>

Deep learning has been successfully exploited in addressing different multimedia problems in recent years. The academic researchers are now transferring their attention from identifying what problem deep learning CAN address to exploring what problem ...



Increasing Attention on Trustworthy

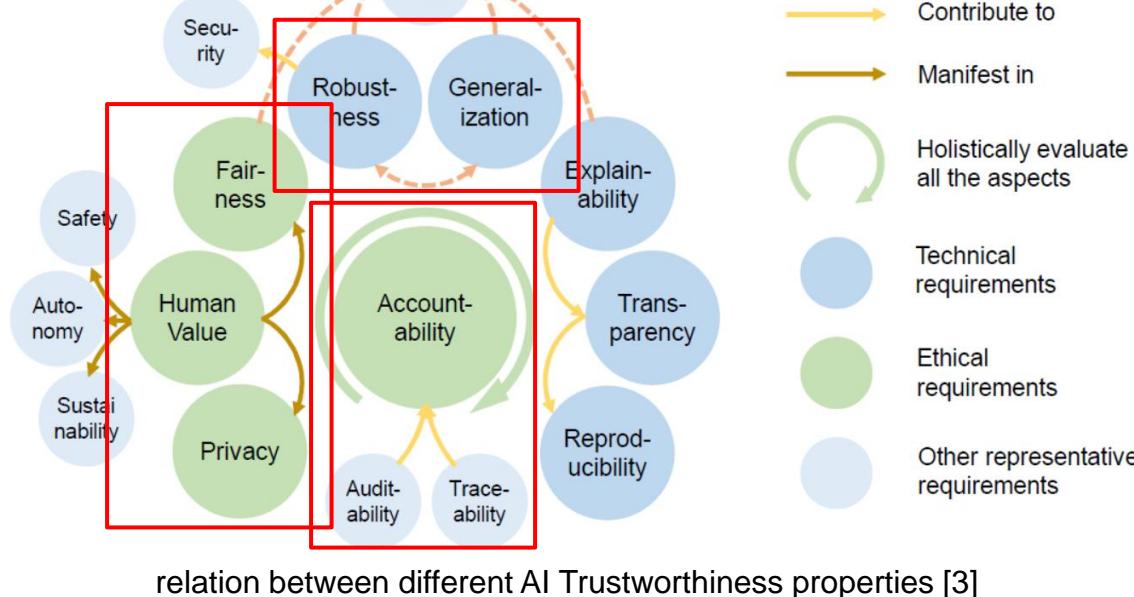


beyond Interpretation: Trustworthy AI

① A trustworthy machine learning system is one that has sufficient: 1. **basic performance**, 2. **reliability**, 3. **human interaction**, and 4. **aligned purpose**. [1]

② ✓ **Technical**: accuracy, **robustness**, and **explainability**.
✓ **User**: availability, usability, safety, **privacy**, and autonomy.
✓ **Social**: law-abiding, ethical, **fair**, **accountable**, and **environmentally friendly**. [2]

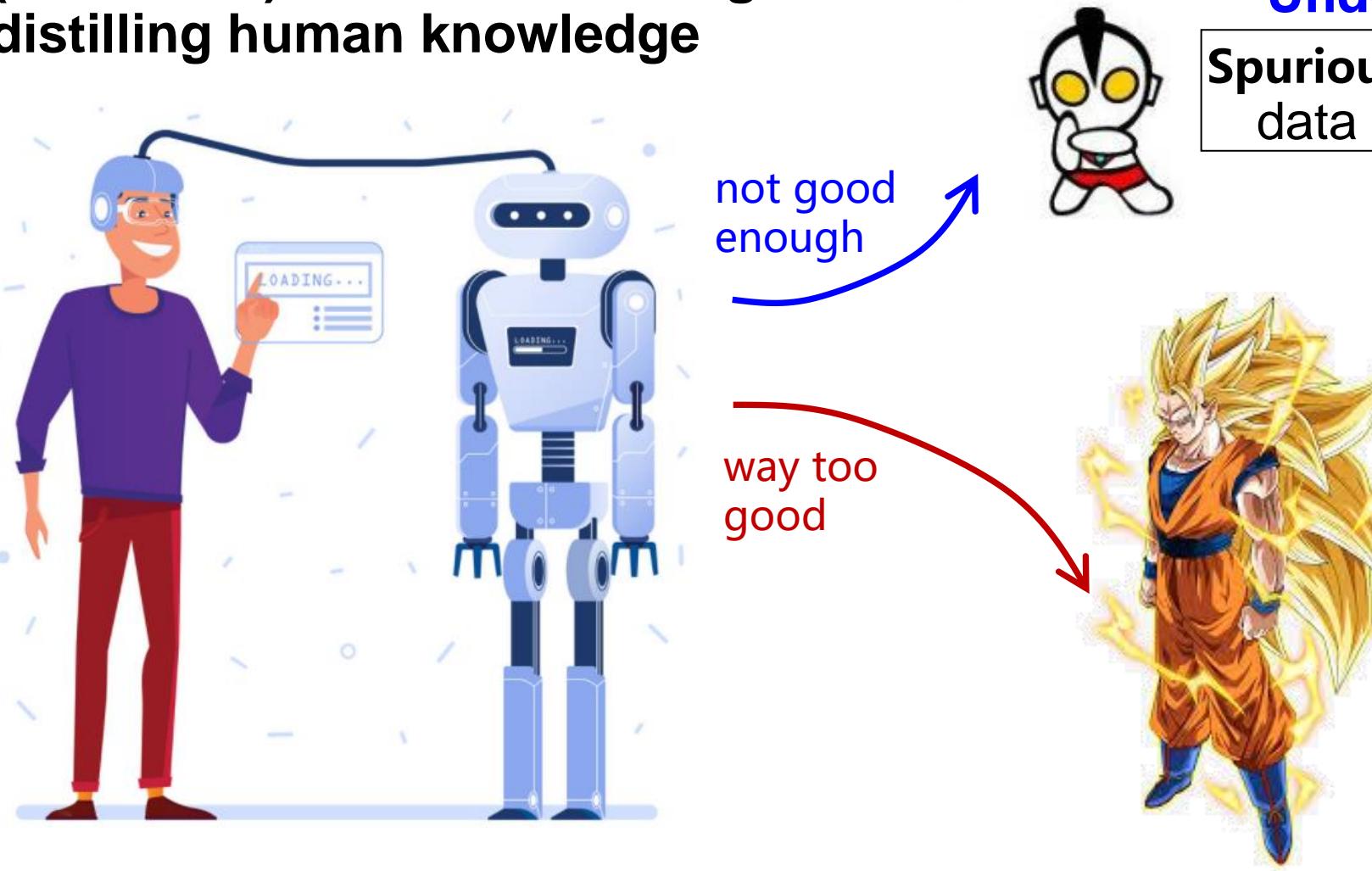
③



- [1] 《Trustworthy Machine Learning》 2021.09.13
[2] Trustworthy AI: a Computational Perspective. 2021.08.19
[3] Trustworthy AI: From Principles to Practices. 2021.10.04

Trustworthy AI: on Spurious Correlations

(Statistical) Machine Learning:
distilling human knowledge



Under-distillation

Spurious Correlations I:
data incompleteness

Over-distillation

Spurious Correlations II:
human-machine disparity

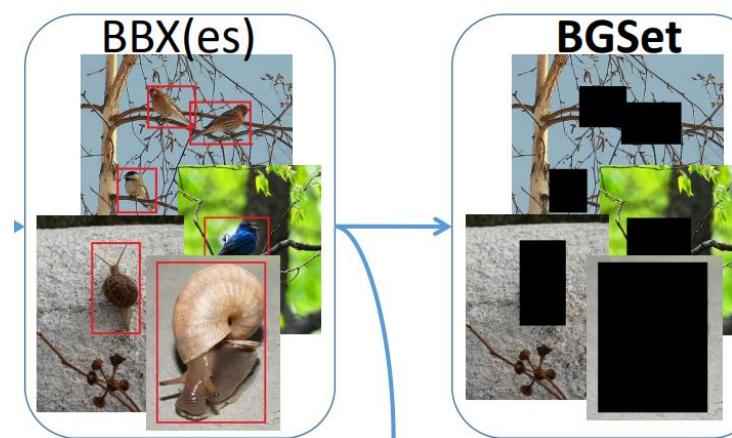
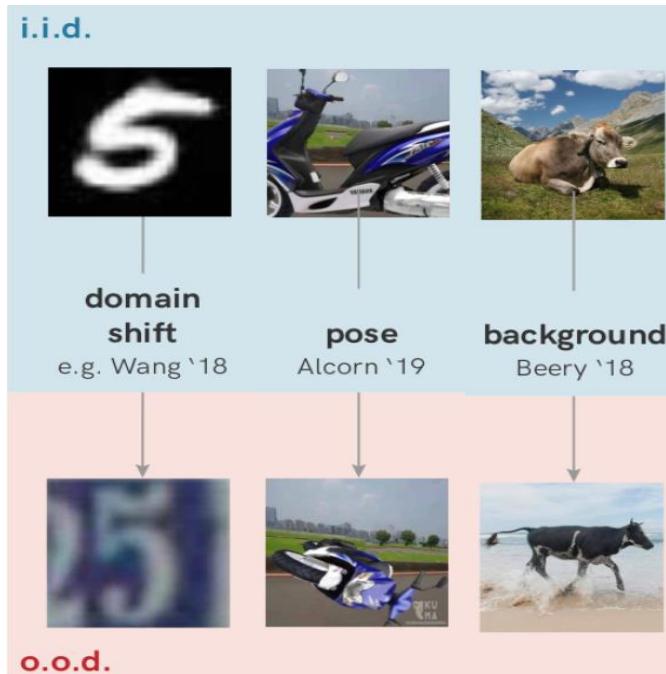
Spurious Correlations I: Under Distillation

Spurious Correlation: Statistical machine learning extracts feature from the correlation in training data, where some features lead to mistakes during system deployment or human interaction.

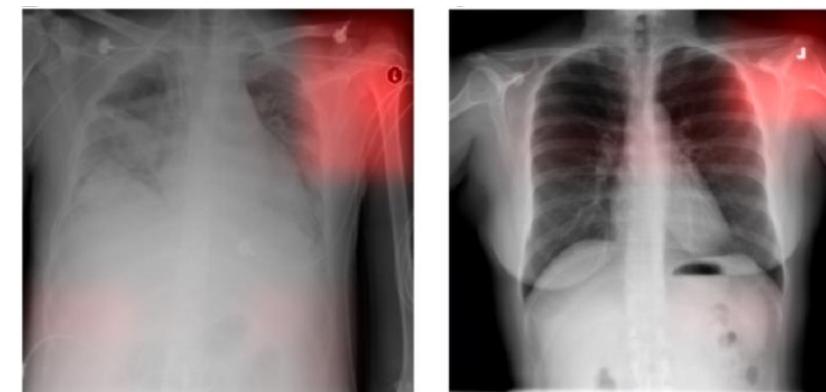
■ Spurious Correlation I: under distillation

Data Incompleteness → machine learns **local correlations** and fails to capture the intrinsic characteristics of task to be solved.

➤ Out-of-Distribution degradation



Identifying visual targets using only background information [2]



Relying on hospital marker for discrimination [3]

OoD degradation regarding domain/pose/background [1]

[1] Shortcut Learning in Deep Neural Networks

[2] Object Recognition with and without Objects

[3] Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

Spurious Correlations I: Under Distillation

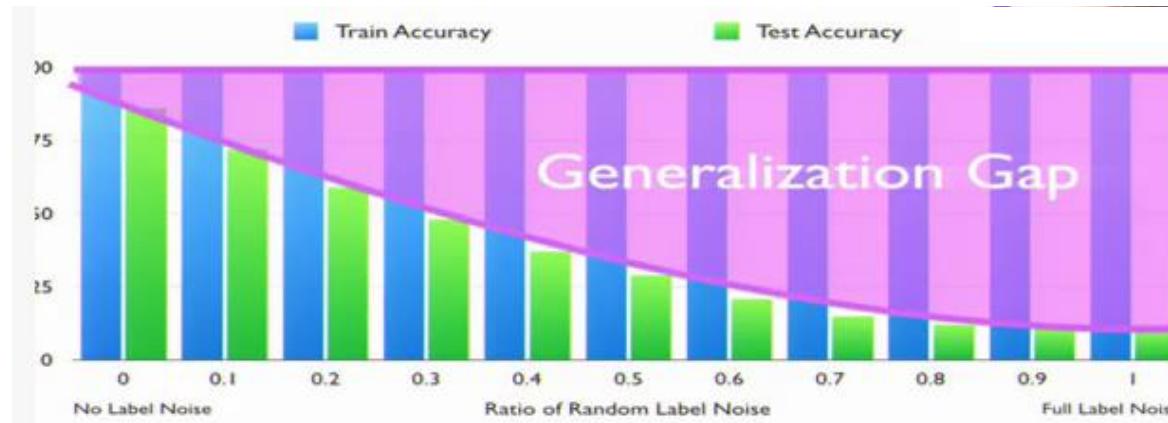
Spurious Correlation: Statistical machine learning extracts feature from the correlation in training data, where some features lead to mistakes during system deployment or human interaction.

■ Spurious Correlation I: under distillation

Data Incompleteness → machine learns local correlations and fails to capture the intrinsic characteristics of task to be solved

➤ Random noise

shuffle the training label → fits in training but degrades in test



Generalized gap increases as label noise increases^[1]

➤ Reward hacking in decision-making



Agent learns solutions to specific tasks, and even converges to local maximum reward^[2]

[1] Understanding Deep Learning Requires Rethinking Generalization.

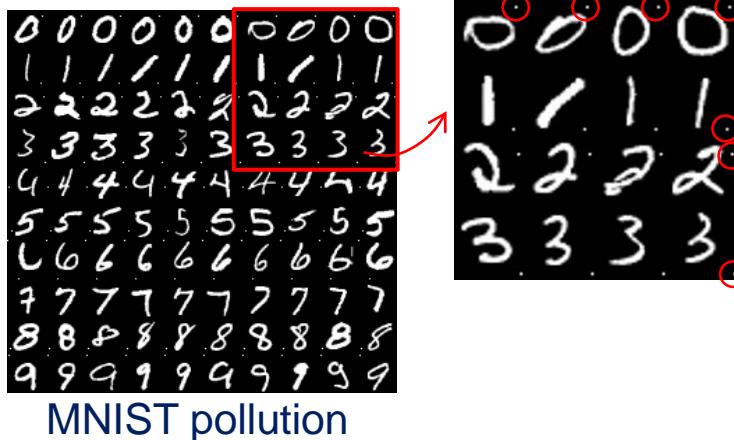
[2] The Animal-AI Environment : Training and Testing Animal-Like Artificial Cognition.

Spurious Correlations I: Under Distillation

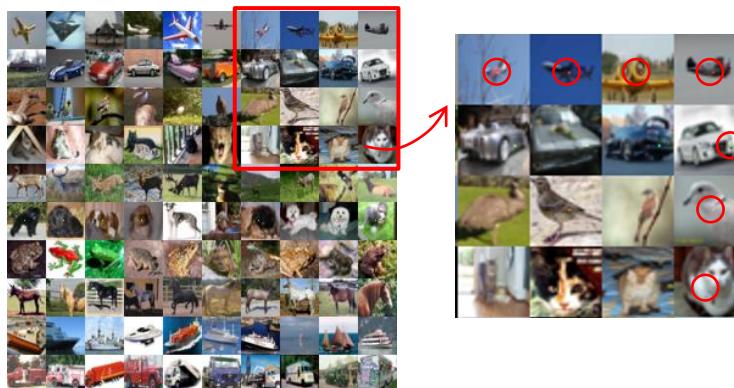
Spurious Correlation: Statistical machine learning extracts feature from the correlation in training data, where some features lead to mistakes during system deployment or human interaction.

■ Spurious Correlation I: under distillation

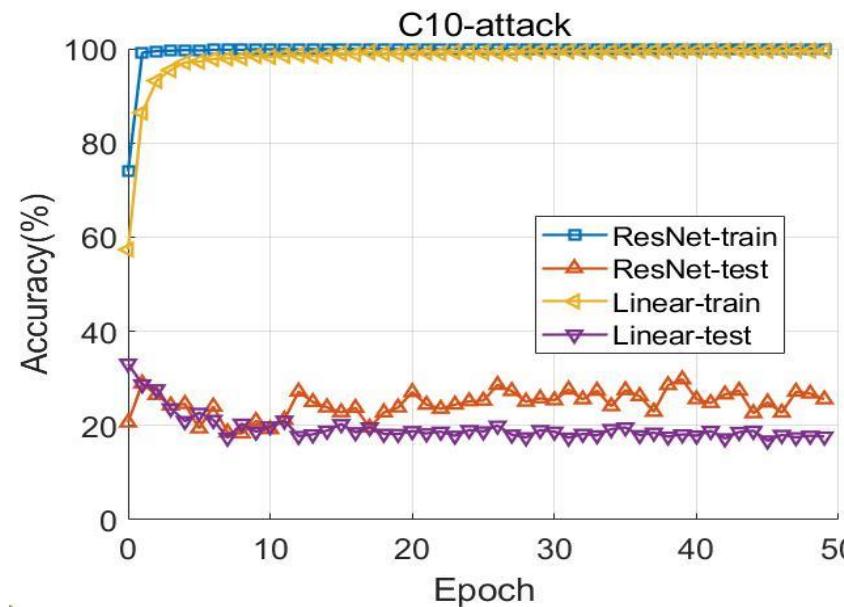
Data Incompleteness → machine learns local correlations and fails to capture the intrinsic characteristics of task to be solved



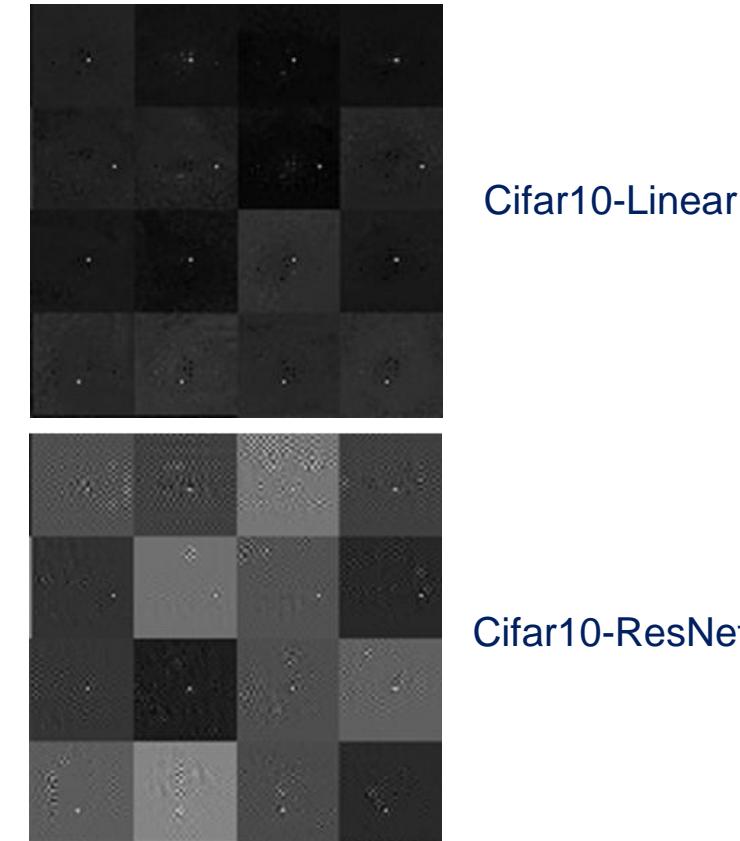
MNIST pollution



Cifar10 pollution



Training/test error curve (ResNet&Linear)



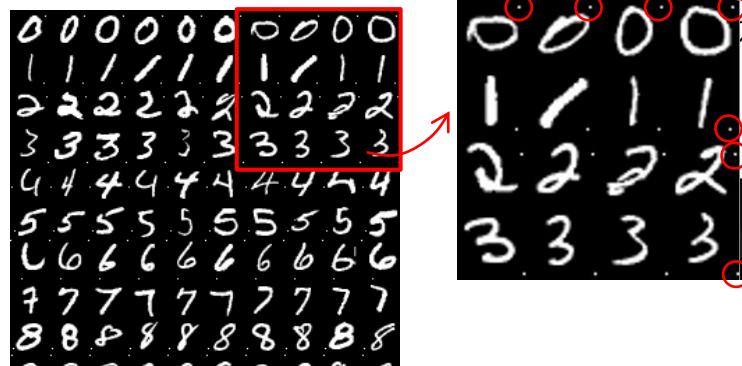
Integrated Gradient attribution map

Spurious Correlations I: Under Distillation

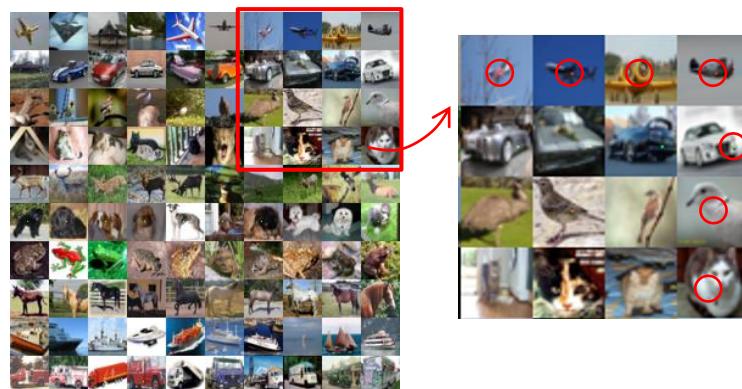
Spurious Correlation: Statistical machine learning extracts feature from the correlation in training data, where some features lead to mistakes during system deployment or human interaction.

■ Spurious Correlation I: under distillation

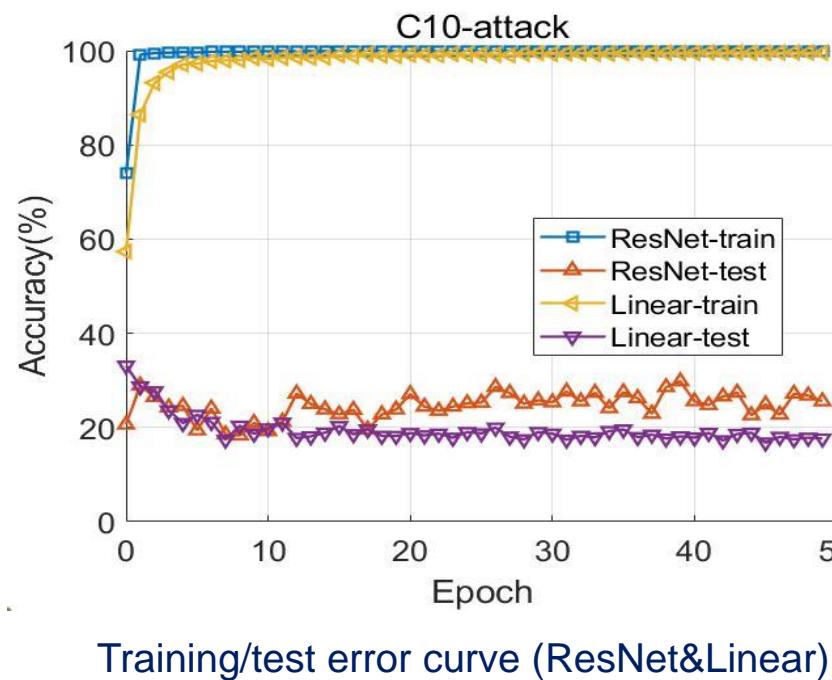
Data Incompleteness → machine learns local correlations and fails to capture the intrinsic characteristics of task to be solved



MNIST pollution



Cifar10 pollution



Adversarial Backdoor Attack: pollute the training set to increase test error while minimizing training error

$$\max_{r_i} L_{D_{X,Y}} \left(\operatorname{argmin}_h \frac{1}{n} \sum_{i=1}^n loss(h(x_i + r_i), y_i) \right) \\ \text{s.t. } \|r_i\| < \varepsilon$$

- Internal: minimizing training error
- External: maximizing test error
- vs Adversarial Attack:
 - modify training vs testing data
 - optimize model vs fixed model
 - Maximize error for testing set vs single sample

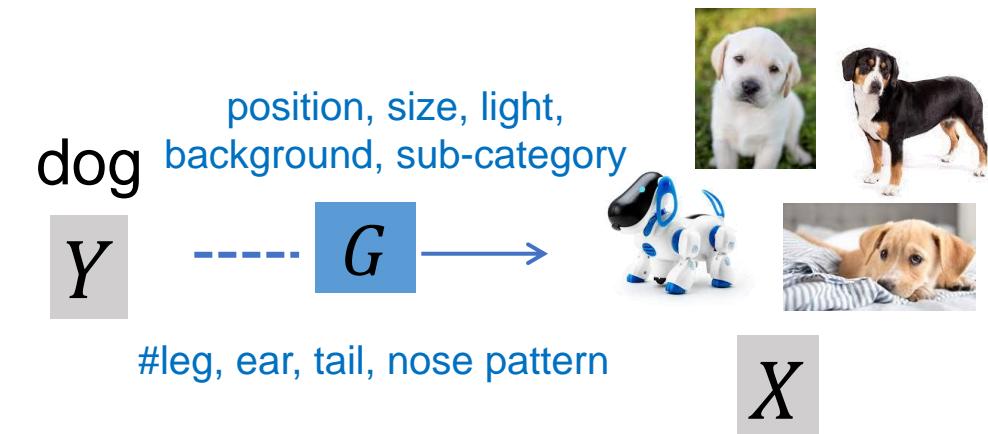
Spurious Correlations I: Under Distillation

■ Spurious Correlation I: under distillation

Data Incompleteness → machine learns local correlations and fails to capture the intrinsic characteristics of task to be solved

□ Task-irrelevant feature:

- Strongly correlated in training set, but fails to generalize to test set
- Generalization from data generation
 - Generative variable G



Spurious Correlations I: Under Distillation

■ Spurious Correlation I: under distillation

Data Incompleteness → machine learns local correlations and fails to capture the intrinsic characteristics of task to be solved

□ Task-irrelevant feature:

➤ Strongly correlated in training set, but fails to generalize to test set

➤ Generalization from data generation

➤ Generative variable G

➤ Task-relevant G_Y vs task-irrelevant G_U :

➤ Independent and Task-Identically Distributed assumption (ITID): considering task-relevant data distribution $\forall D_i, D_j: P_{D_i}(G_Y, Y) = P_{D_j}(G_Y, Y)$

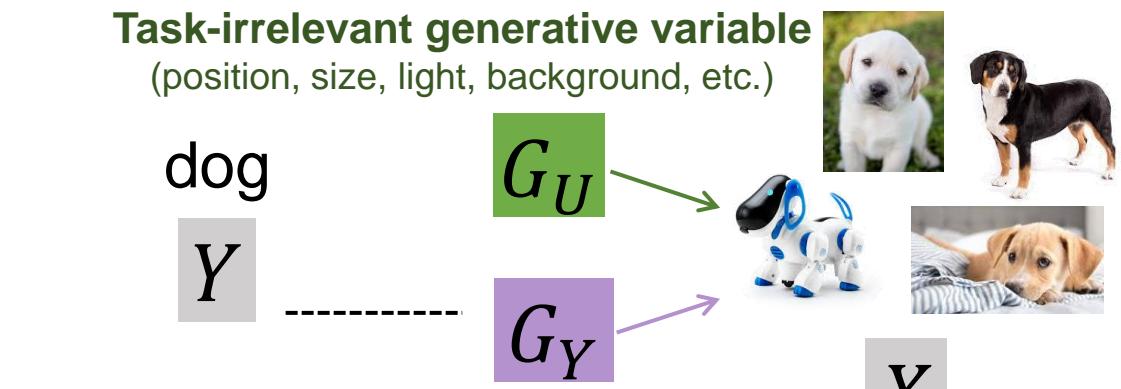
➤ ITID-based generalization gap: generalization performance \propto dependence on task-irrelevant variable

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \sqrt{\frac{2TK \ln 2 + \ln(1/\delta)}{n}} + \frac{\gamma}{\log 2}$$

generalization error
for optimal model

Task complexity

- ✓ T : #state of task-relevant variable
- ✓ K : #state of output variable
- ✓ n : #training data



Dependence on task-irrelevant variable
 h has γ -dependence on G_U : $H(\hat{Y}|G_U) \leq \gamma$

[A Generalization Theory based on Independent and Task-Identically Distributed Assumption.](#)

Spurious Correlations I: Under Distillation

■ Spurious Correlation I: under distillation

Data Incompleteness → machine learns local correlations and fails to capture the intrinsic characteristics of task to be solved

□ Task-irrelevant feature:

- Strongly correlated in training set, but fails to generalize to test set

- Generalization from data generation

- Independent and Task-Identically Distributed

Assumption (ITID): considering task-relevant data distribution $\forall D_i, D_j: P_{D_i}(G_Y, Y) = P_{D_j}(G_Y, Y)$

- ITID-based generalization gap: generalization performance \propto dependence on task-irrelevant variab

□ Employing task-irrelevant feature:

- Generalization

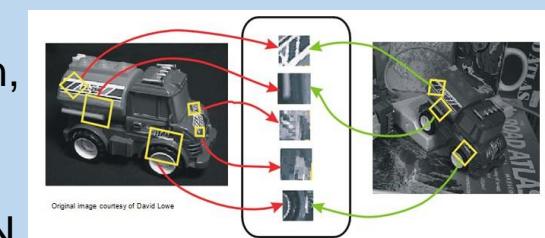
- **Data:** data augmentation reduces the correlation of task-irrelevant features by balancing data distribution.



- **Model:** regularize the learning of task-irrelevant features

- Target: invariant features to translation, rotation, and zoom

- Solution: hand-designed invariant features + convolution&pooling in CNN



[A Generalization Theory based on Independent and Task-Identically Distributed Assumption.](#)

Spurious Correlations I: Under Distillation

■ Spurious Correlation I: under distillation

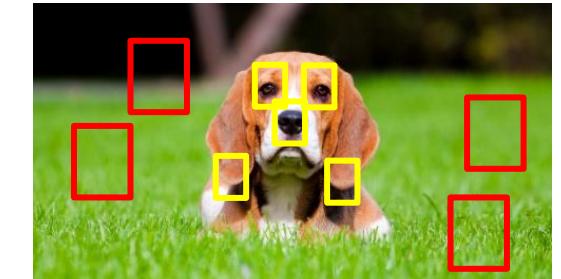
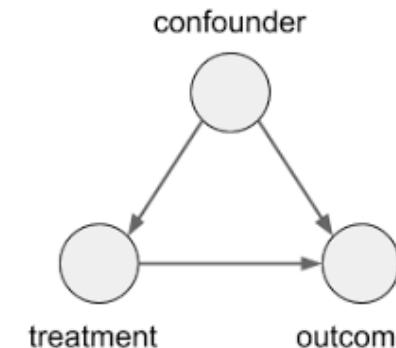
Data Incompleteness → machine learns **local correlations** and fails to capture the intrinsic characteristics of task to be solved

□ Task-irrelevant feature:

- Strongly correlated in training set, but fails to generalize to test set
- Generalization from data generation
- **Independent and Task-Identically Distributed Assumption (ITID)**: considering task-relevant data distribution $\forall D_i, D_j: P_{D_i}(G_Y, Y) = P_{D_j}(G_Y, Y)$
- **ITID-based generalization gap**: generalization performance \propto dependence on task-irrelevant variable

□ Problems of employing task-irrelevant feature:

- **Generalization**
- **Causality**: task-irrelevant feature is semantic-oriented
→ confounder variable
- **Fairness**: confounder variable is socially sensitive
→ bias variable



Task-irrelevant feature: confounder variable
(e.g., background)



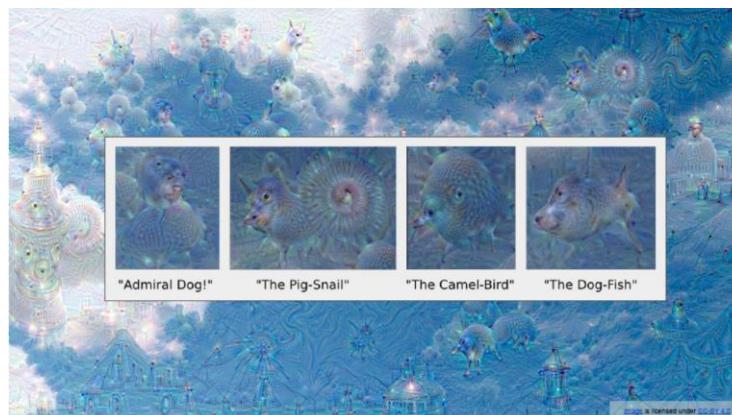
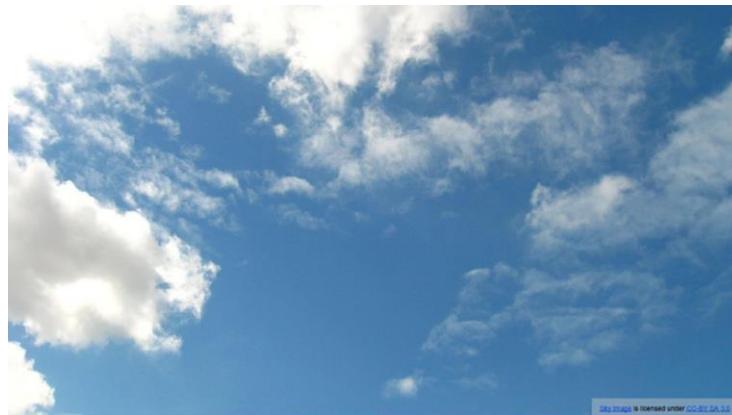
Task-irrelevant feature: bias variable (e.g., race, gender)

Spurious Correlations II: Over Distillation

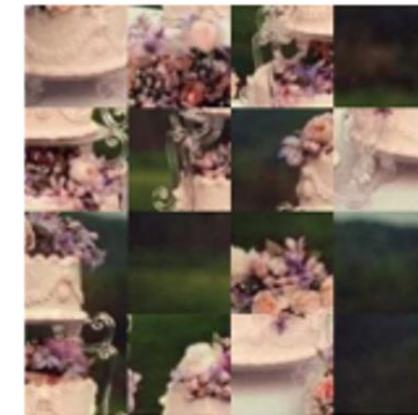
■ Spurious Correlation II: over distillation

Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ What can you “see”?



□ What is this?



CNN recognizes objects with trivial shape (relying more on texture) [2]

“DeepDream” effect using VGG16 trained on ImageNet [1]

[1] These Google “Deep Dream” Images Are Weirdly Mesmerising. *Wired*, 2015.

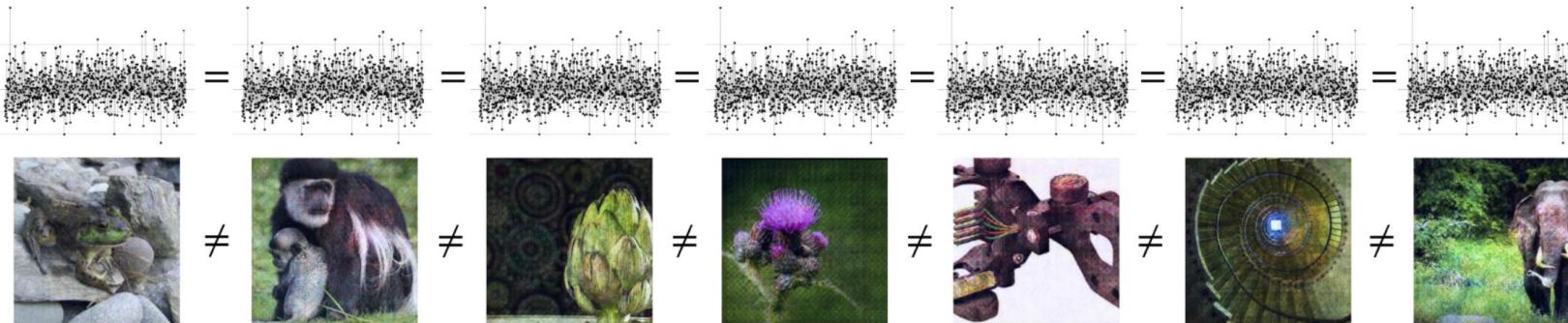
[2] ImageNet-trained CNNs are biased towards texture. *ICLR* 2019.

Spurious Correlations II: Over Distillation

■ Spurious Correlation II: over distillation

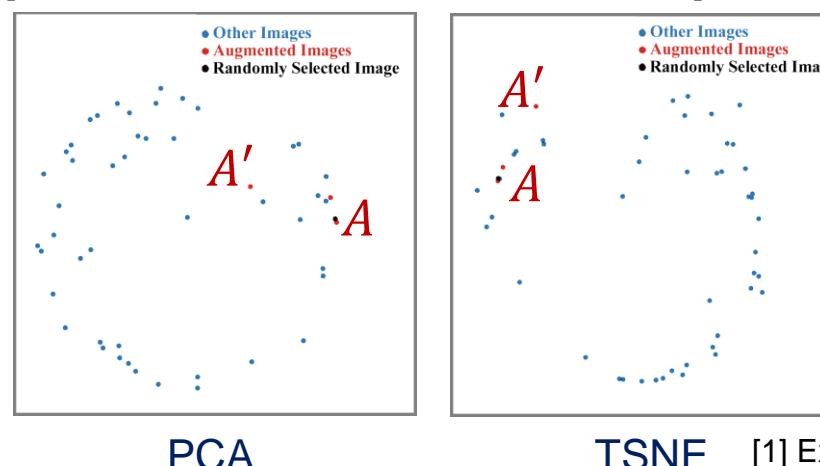
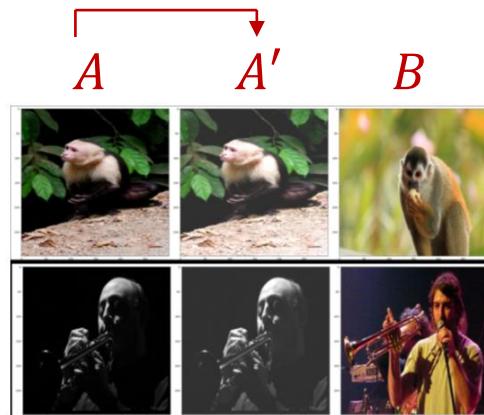
Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ Different human perception && same model representation [1]



□ Similar human perception && different model representation [2]

modify light and contrast



$$d(x_A, x_{A'}) > d(x_A, x_B)$$

[1] Excessive Invariance Causes Adversarial Vulnerability.

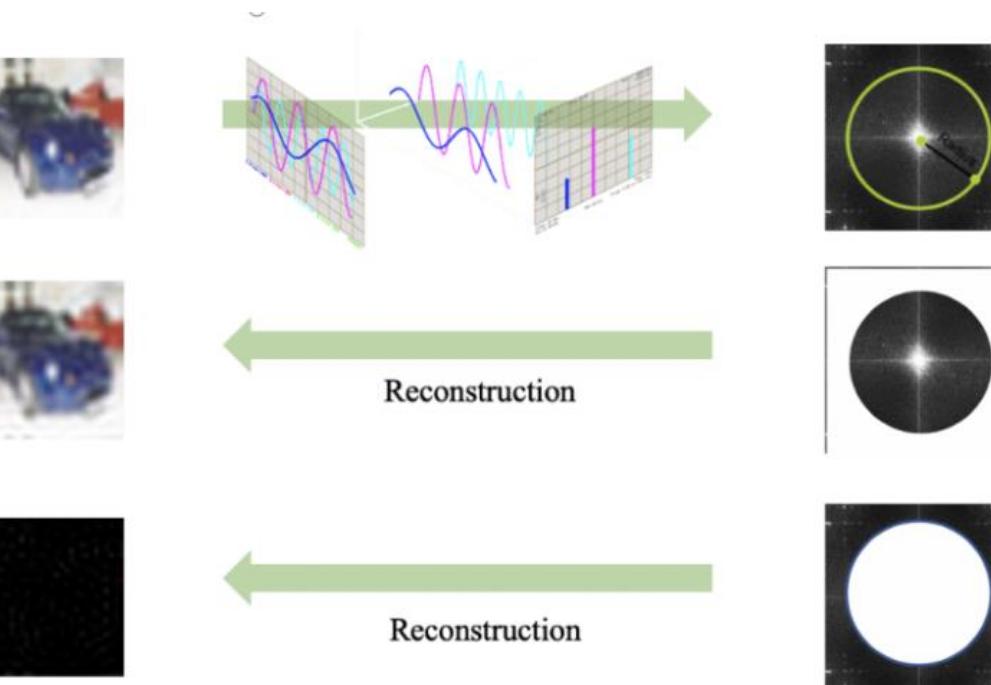
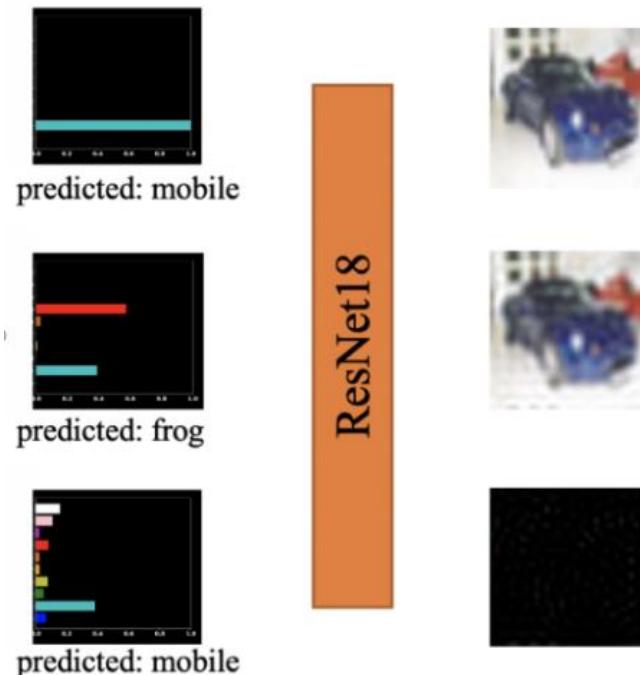
[2] An Experimental Study of Semantic Continuity for Deep Learning Models.

Spurious Correlations II: Over Distillation

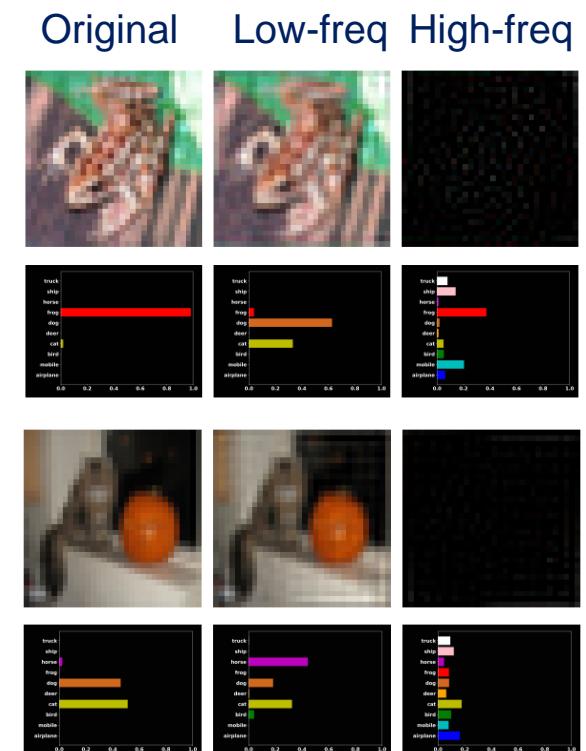
■ Spurious Correlation II: over distillation

Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ Machine: high frequency vs Human: low frequency



Images reconstructed from high-frequency is unrecognizable to human,
but well classified by DNN model [1]



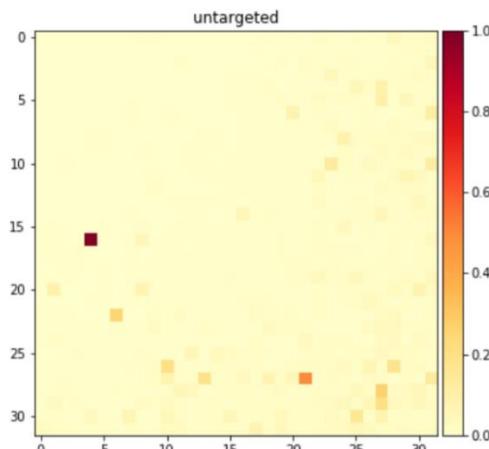
[1] High Frequency Component Helps Explain the Generalization of Convolutional Neural Networks.

Spurious Correlations II: Over Distillation

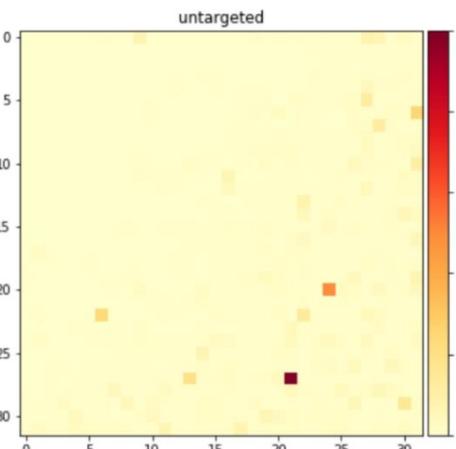
■ Spurious Correlation II: over distillation

Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ Adversarial attack occurs more in high-frequency components

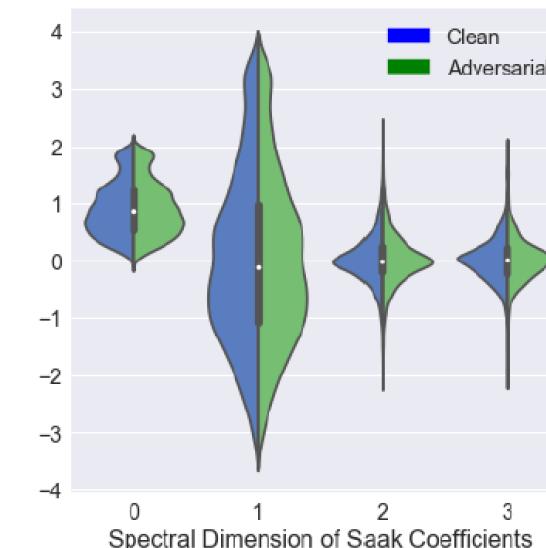


(a) PGD Attack

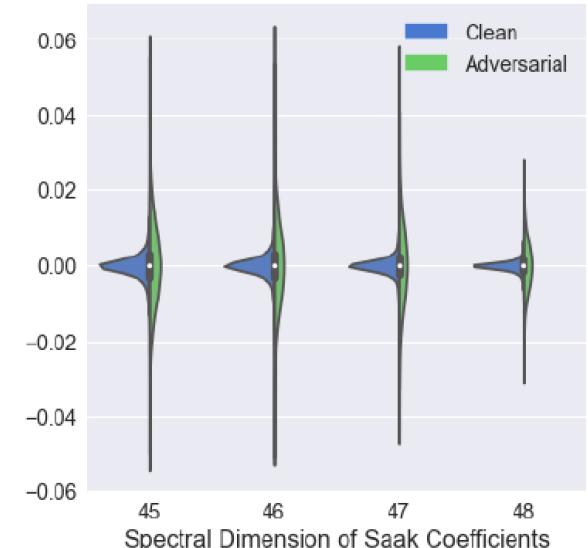


(b) CW attack

RCT maps for original-adversarial change of CIFAR-10 images^[1]



The distributions of Saak coefficients of clean images and adversarial images^[2]



[1] Towards Frequency-Based Explanation for Robust CNN.

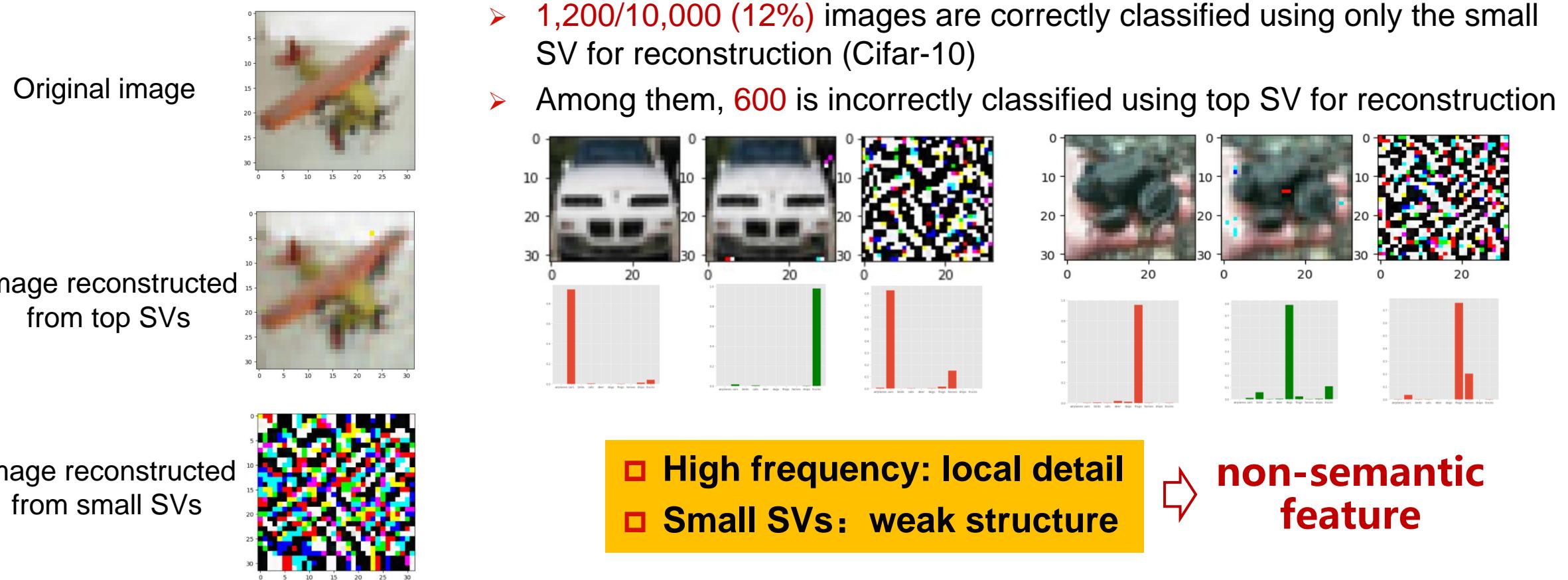
[2] Defense Against Adversarial Attacks with Saak Transformation.

Spurious Correlations II: Over Distillation

■ Spurious Correlation II: over distillation

Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ Machine: fuzzy structure vs Human: vivid structure



Spurious Correlations II: Over Distillation

■ Spurious Correlation II: over distillation

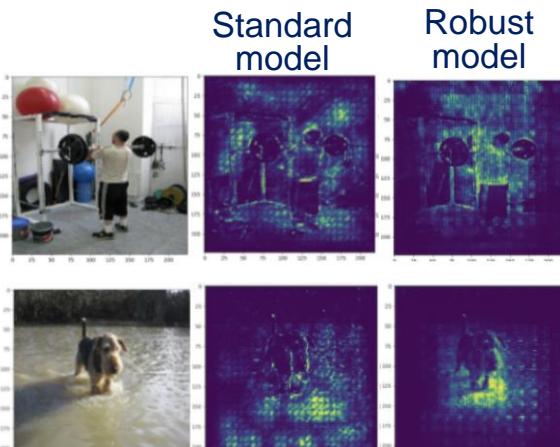
Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ Non-semantic feature:

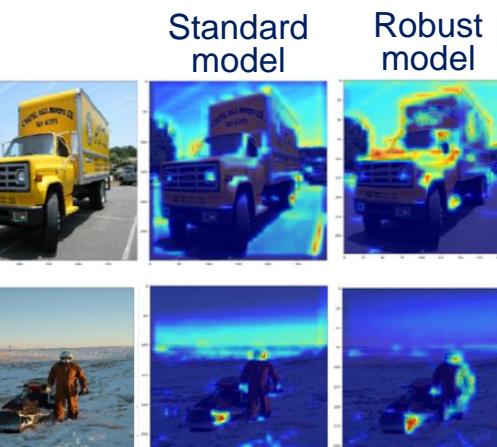
- Imperceptible or incomprehensible to human
- Disturbance doesn't affect human perception

□ Problems of using non-semantic feature:

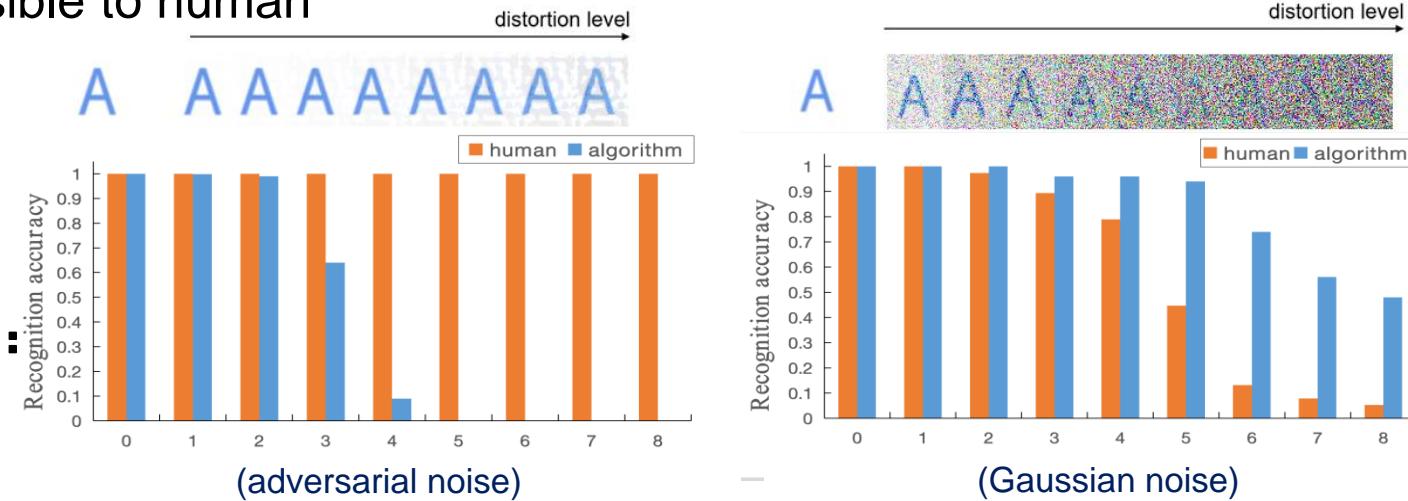
- Adversarial robustness
- Interpretation [2]



Integrated Gradients



Grad-CAM



Human-machine disparity on different disturbances [1]

[1] [Robust CAPTCHAs towards Malicious OCR](#).

[2] [An Experimental Study of Semantic Continuity for Deep Learning Models](#).

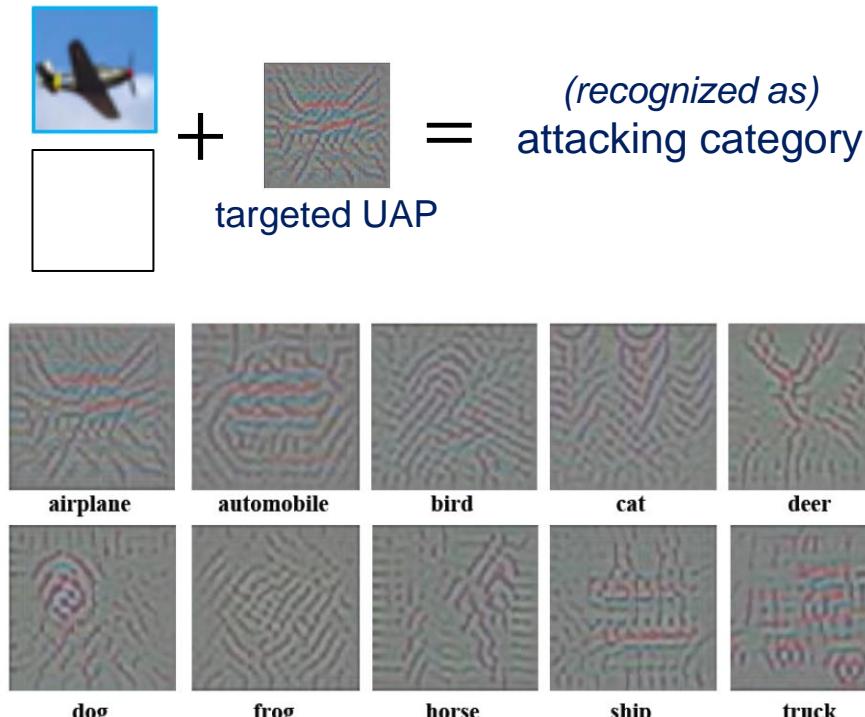
Spurious Correlations II: Over Distillation

■ Spurious Correlation II: over distillation

Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

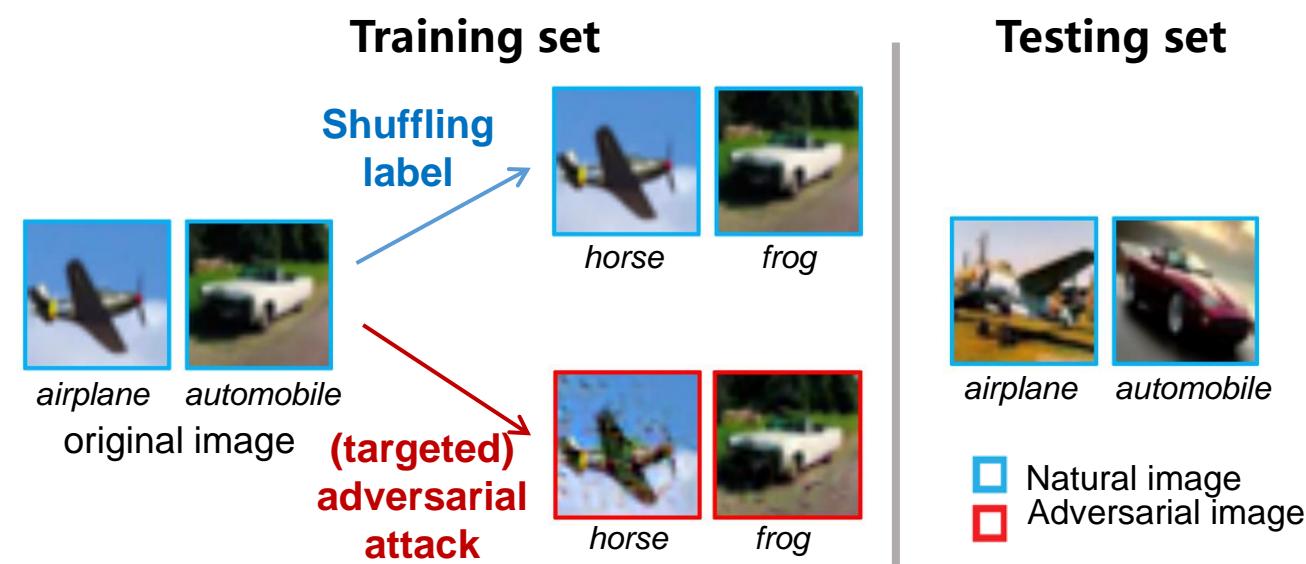
□ Non-semantic features can be task-relevant

① Adversarial attack provides **discriminative** features



targeted UAP(Universal Adversarial Perturbation) of CIFAR10

② Adversarial attack provides **generalizable** features



accuracy	Training set	Testing set
Random noise	99.89%	10.41%
Adv attack	99.77%	66.30%

[Benign Adversarial Attack: Tricking Algorithm for Goodness.](#)

Spurious Correlations II: Over Distillation

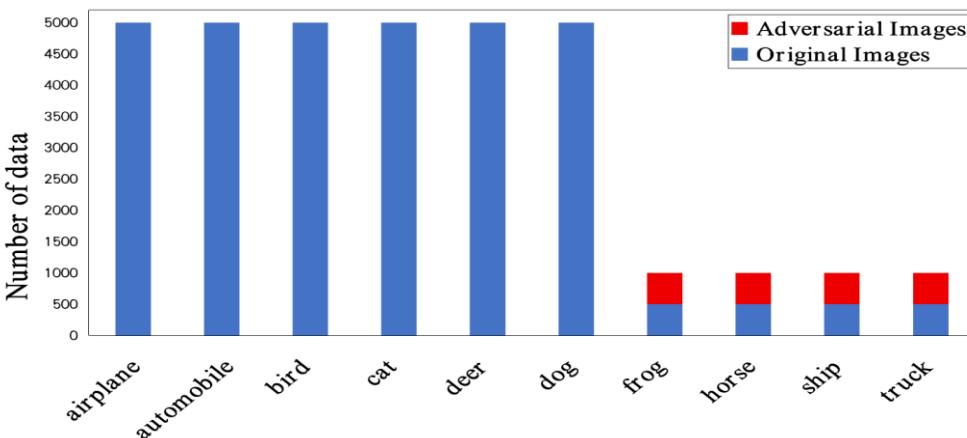
■ Spurious Correlation II: over distillation

Human-machine disparity → machine learns information patterns imperceptible or incomprehensible to human

□ Non-semantic features can be task-relevant

③ Adversarial attack provides complementary features

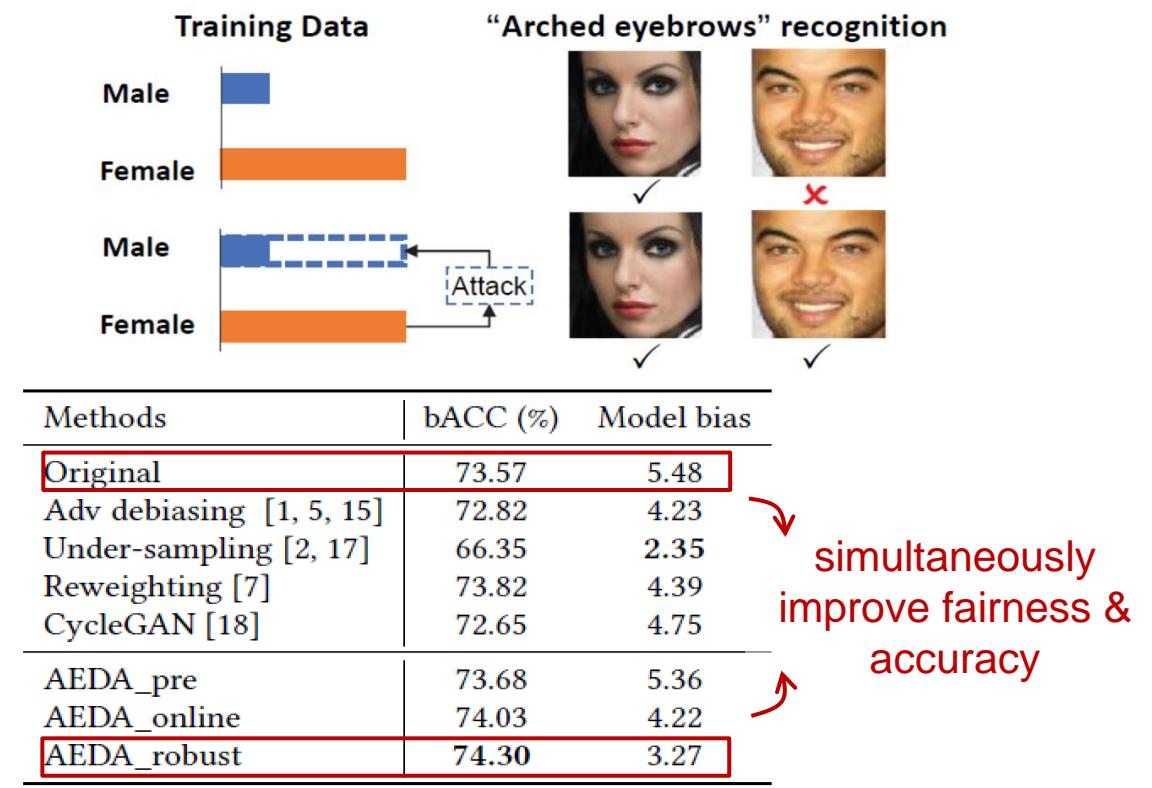
- **Few-shot:** non-semantic features in adv samples can make up for the lack of natural samples [1]



Models	Classification accuracy				
	frog	horse	ship	truck	average
fimb	44.7%	42.5%	56.6%	54.1%	49.5%
fauq	53.7%	45.4%	58.0%	61.7%	54.7%

↑ 5%+

- **Fairness:** generate adversarial pseudo sample to balance bias variable, improve fairness & accuracy [2]



[1] Benign Adversarial Attack: Tricking Algorithm for Goodness.

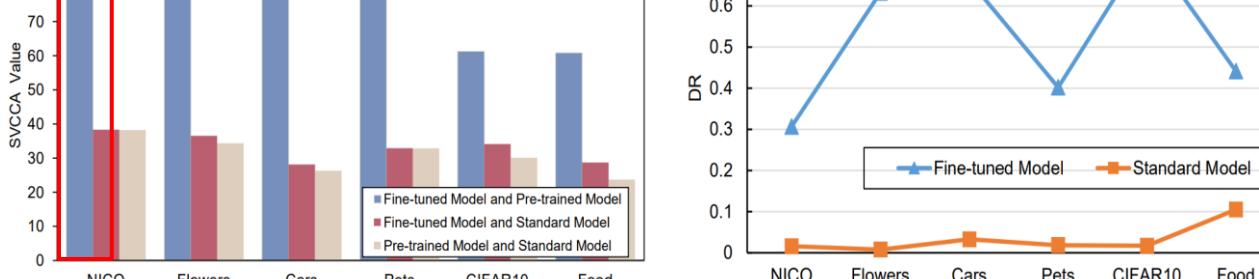
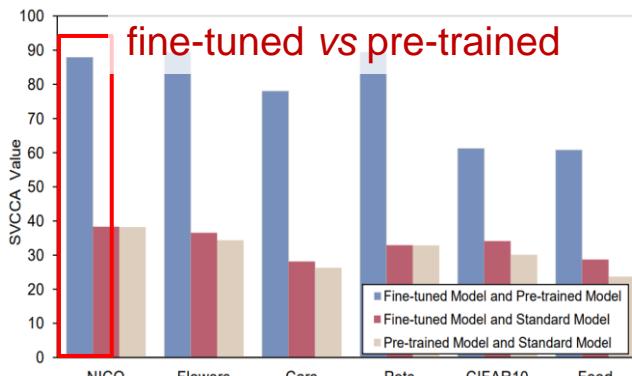
[2] Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing.

Spurious Correlations II: Over Distillation

Transfer Learning inherits Non-semantic features → robustness & interpretation ↓

- Pre-trained model: trained on large-scale source set
- Fine-tuned model: initialized with pre-trained model and fine-tuned on target dataset
- Standard model: directly trained on limited target set

① Fine-tuned model is more close to pre-trained model

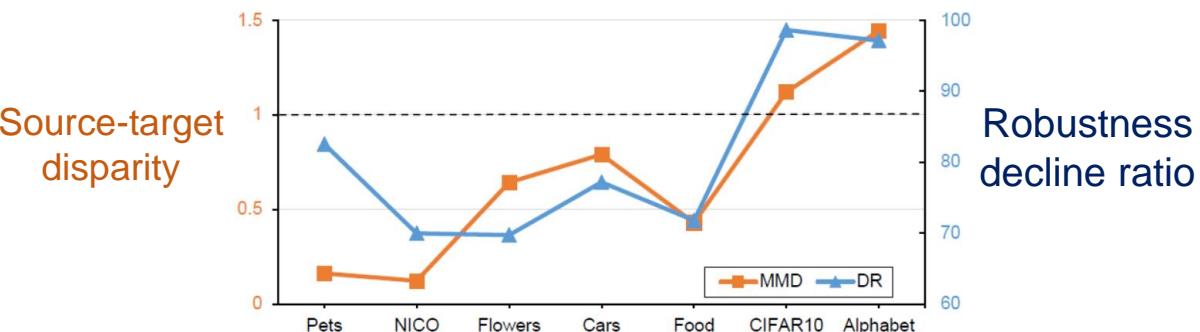


On the learned knowledge (SVCCA)



Target dataset (source dataset: ImageNet)

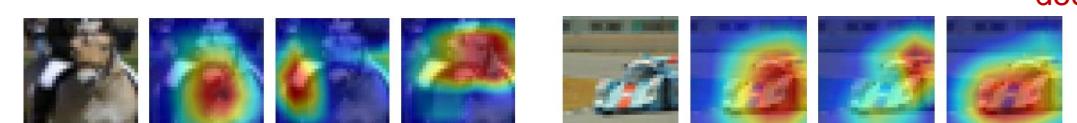
② Degradation of robustness & interpretation ~ Source-Target disparity



③ Robust & Interpretable Pre-training

- Steepness regularization: reduce LLF of pre-trained feature space on target samples

Method	Pets	NICO	Flowers	Cars	Food	CIFAR10	Alphabet	accuracy
Full Fine-tuned	AOI	88.74	92.71	89.17	60.63	71.80	92.38	100.00
	AAI	10.19	20.15	18.38	0.31	1.50	0.02	2.90
	DR	88.51	78.26	79.39	99.48	97.90	99.97	97.10
robustPT	AOI	86.48	91.71	87.17	64.83	70.04	95.62	99.96
	AAI	77.73	85.50	81.41	53.46	47.93	88.63	99.90
	DR	10.11	6.77	6.60	17.53	31.56	7.31	0.05



standard fine-tuned
robust fine-tuned

Pre-training also Transfers Non-Robustness.

More to Explore

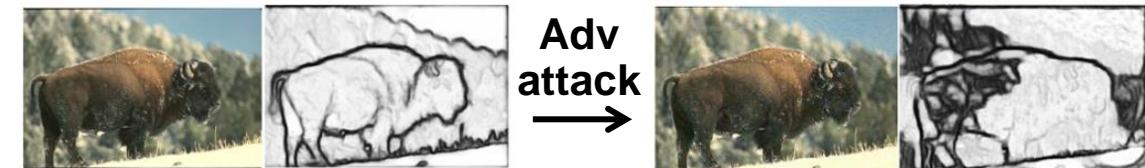
□ Spurious correlation in transfer learning:

- Transfer learning (e.g., pre-training) becomes popular for large-scale industrial applications.
- It can be seen as a process of **self knowledge distillation**.
- **Echo-chamber** in self knowledge distillation accumulates spurious correlations in the transferred knowledge.

<i>transferred knowledge</i>	<i>solution</i>	Pre-training/ Self-supervised learning	Knowledge distillation/ Auto annotation	Few-shot/ Semi-supervised learning
Feature extractor $T: x \rightarrow T(x)$	✓			✓
Predictor $f: x \rightarrow y$			✓	✓

□ Unifying the problems within the same type:

- Fairness-Generalization
Employing task-irrelevant features (social attribute)
- Interpretation-Adversarial Robustness
Robustness to input disturbance: **gradient (1 order vs output (0 order))**



Adversarial attack destroys the continuity of image space

□ Aligning the problems between two types:

- Tradeoff?:
 - Generalization results from non-semantic employment
 - In current training setting, penalizing non-semantic will sacrifice generalization.
- Alignment:
 - Stability: sampling + noise disturbances
 - **Sampling disturbance**: generalization;
 - **Noise disturbance**: adversarial robustness.

Modifying X	Loss design	Goal
Sampling disturbance (changing semantic)	Y: discriminative loss (class-sensitive)	Task-relevant feature
Noise disturbance (changing non-semantic)	Y: contrastive loss (sample-sensitive)	Semantic feature

Self-supervised Stability learning

Spurious Correlation → Trustworthy Multimedia Analysis

Trustworthy: as expected

Machine Learning = Software 2.0

Input: what to do → Output: how to do

□ **Understanding task: what to do**

➤ No adequate training data

□ **Executing task: how to do**

➤ No sufficient supervision

Spurious Correlation → Trustworthy Multimedia Analysis

Trustworthy: as expected

Machine Learning = Software 2.0

Input: what to do → Output: how to do

❑ Understanding task: what to do

- No adequate training data

❑ Executing task: how to do

- No sufficient supervision

Spurious Correlation

❑ Under Distillation: data incompleteness

- Task-irrelevant feature:

Generalization/causality/fairness



Task-relevance

visual feature taxonomy coordinate

Spurious Correlation → Trustworthy Multimedia Analysis

Trustworthy: as expected

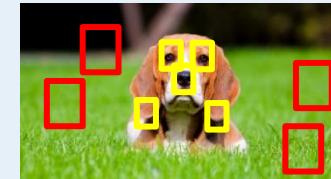
Machine Learning = Software 2.0

Input: what to do → Output: how to do

- **Understanding task: what to do**
 - No adequate training data
- **Executing task: how to do**
 - No sufficient supervision

Semantic-orientation

Task-irrelevant Semantic feature



Task-relevant Semantic feature



Spurious Correlation

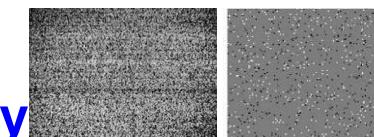
□ Under Distillation: data incompleteness

- Task-irrelevant feature:

Generalization/causality/fairness

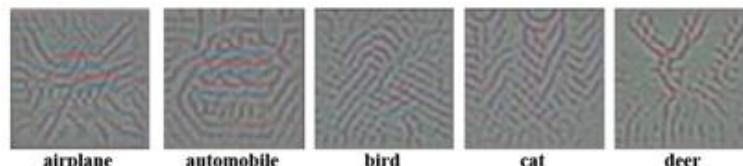


Task-irrelevant Non-semantic feature



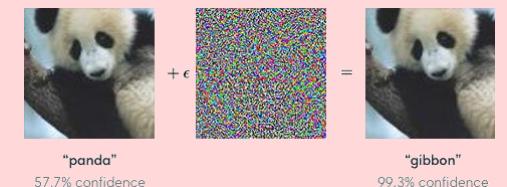
□ Over Distillation: human-machine disparity

- Non-semantic feature: adversarial robustness/interpretability



Task-relevance

Task-relevant Non-semantic feature

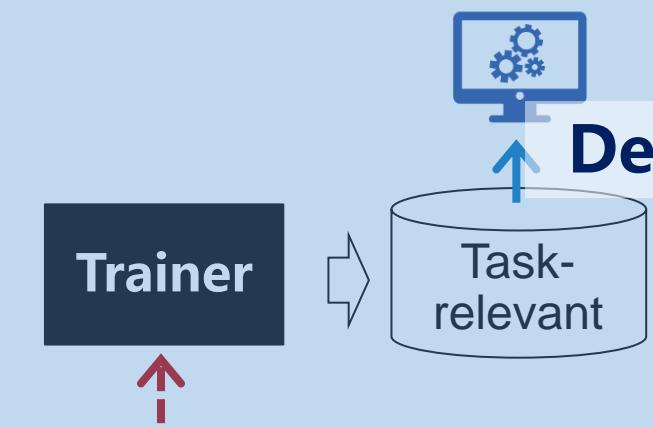


visual feature taxonomy coordinate

Trustworthy Multimedia Analysis

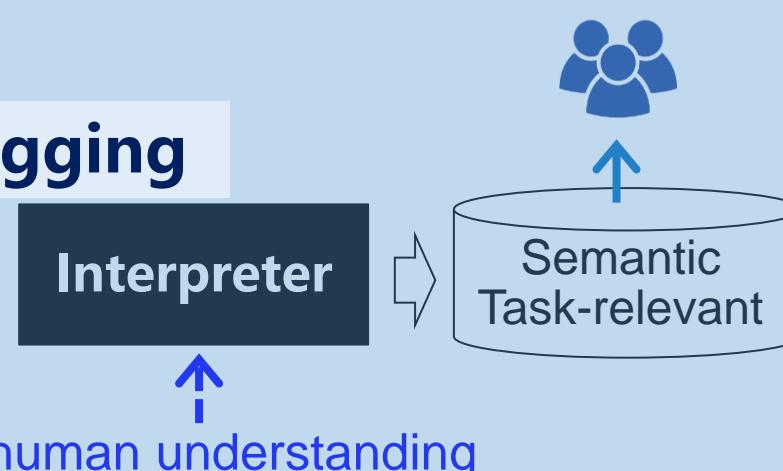
Trainer

- Goal: generalize to unseen data
- Output: Task-relevant feature → system deployment



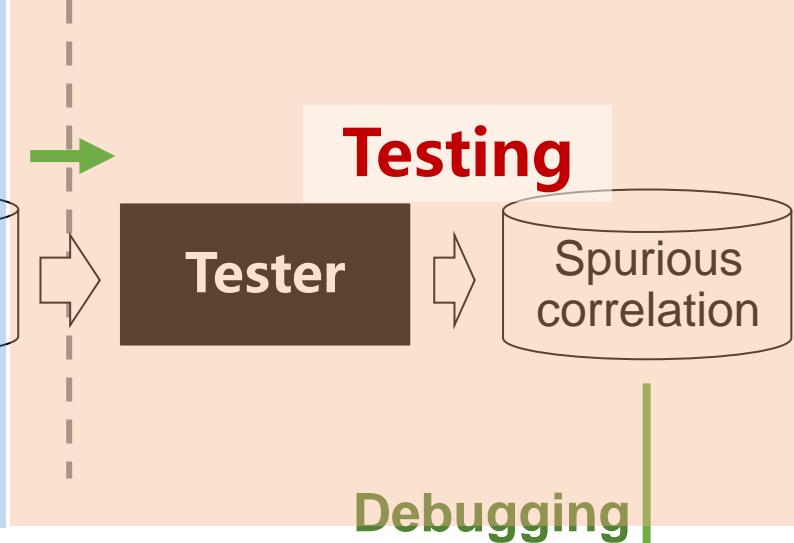
Interpreter

- Goal: human understanding
- Output: Task-relevant semantic feature → human interaction



Tester

- Goal: detecting bugs
- Output: spurious correlations → corrected by trainer and interpreter



Accuracy-Compatible
Fairness Computing

Human-like Adversarial
Robustness & Interpretation

Spurious Correlation-oriented
Algorithm Testing

Benign Adversarial Attack:
Adversarial Privacy Preserving

Trustworthy User Modeling

Outline

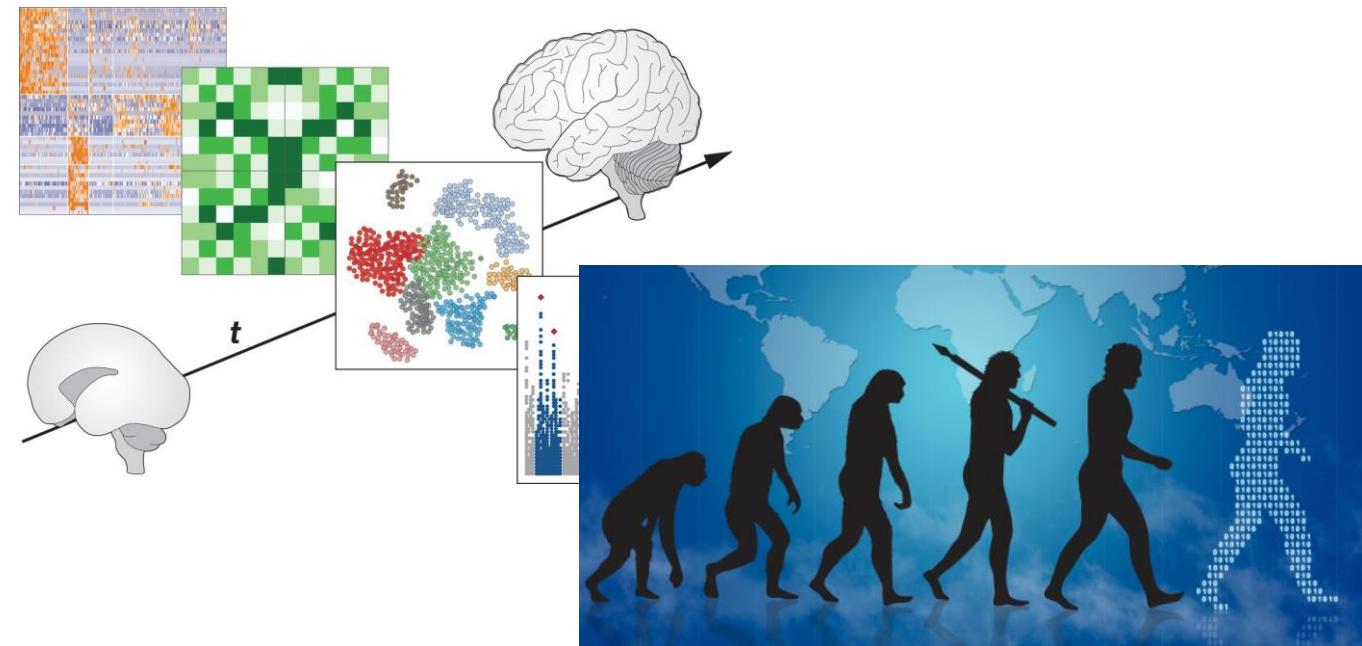
- ✓ Part I: On the Two Types of Spurious Correlations
- ❑ Part II: Accuracy-compatible Fairness Computing (30mins)
by Yi Zhang
- ❑ Part III, Benign Adversarial Attack: Adversarial Privacy (30mins)
by Jiaming Zhang
- ❑ Part IV, Trustworthy User Modeling (60mins)
by Dr. Xiaowen Huang



On Spurious Correlation I: Small Data vs Big Data

□ Biometric neural network: small data vs big data

- **Small data:** individual perspective, babies learn by analogy
- **Big data:** group perspective, hundreds of millions of years of data accumulation
→ structure prior in genetic/physiological evolution [1]



[1] Leslie Valiant: Evolution as Learning.

On Spurious Correlation II: Human-like *vs* beyond-Human

▢ Human-like: Trustworthy AI

→ Safety-critical and high-stake applications:
interaction is necessary with the guarantee of
understanding



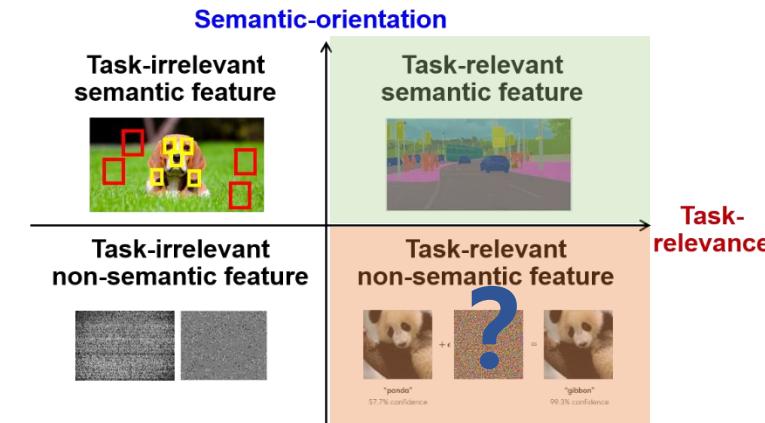
“no crossing the line”

□ beyond Human: explore what human CANNOT

→ Address tasks human cannot solve, and even discover new ones



chatbots create non-human languages
to communicate with each other



Current learning mechanism is for human-like target and cannot fully explore the potential of non-semantic information

- **Why** machine learns non-semantic features?
 - **What** characteristics the non-semantic feature haves?
 - **How** to extract and better employ non-semantic features?

Reference links

[A Generalization Theory based on Independent and Task-Identically Distributed Assumption. 2019.](#)

[Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. 2020](#)

[Robust CAPTCHAs towards Malicious OCR. 2020.](#)

[An Experimental Study of Semantic Continuity for Deep Learning Models. 2020](#)

[Benign Adversarial Attack: Tricking Algorithm for Goodness. 2021](#)

[Pre-training also Transfers Non-Robustness. 2021](#)

Thanks

