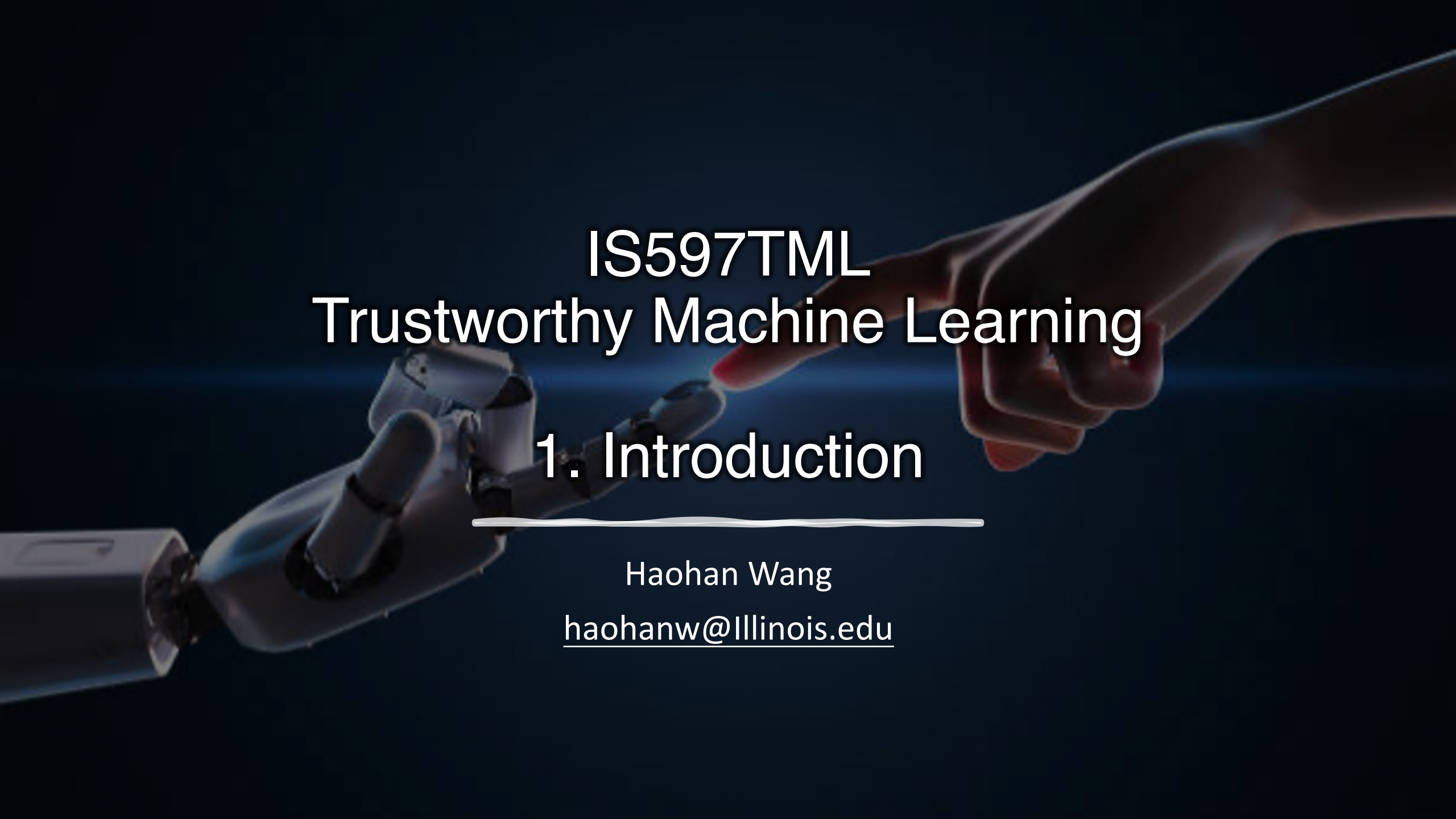


IS597TML
Trustworthy Machine Learning

1. Introduction

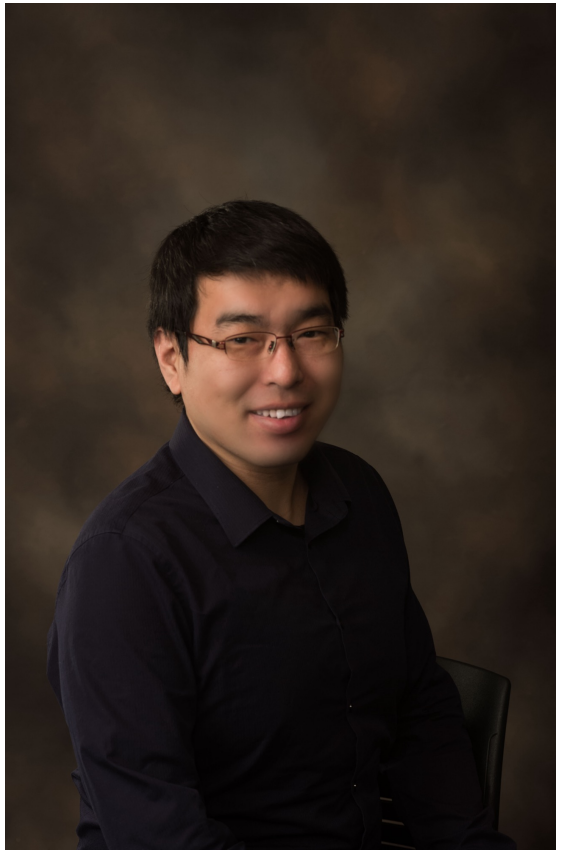
Haohan Wang

haohanw@Illinois.edu

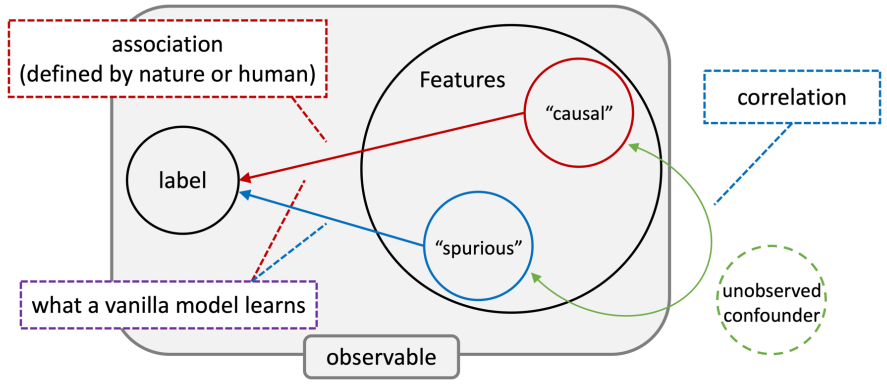
Agenda

- Logistics
- Introduction of Trustworthy Machine Learning
- Universal Solutions of Trustworthy Machine Learning
- Project Ideas

Instructor



- Haohan Wang
  - https://haohanwang.github.io/
  - Room 4123, 614 East Daniel St.
  - Office hours: 2-3pm Thursday
  - Research Interests
    - Trustworthy machine learning
      - Learning “true” signals out of data, instead of dataset artifacts
    - Computational Biology
      - Techniques to learn “true” signals from data to advance human knowledge of biology



Teaching Assistant – Haoyang Liu

- Informatics, advisor: Haohan Wang
- Research interest: dataset distillation, robustness in vision, AI for biomedical research
- Office: Room 4120, 614 East Daniel St.
- Email: hl57@illinois.edu
- Office hour: TBD

Expectations

- Knowledge in trustworthy machine learning
  - Module 1: Statistical Foundation
    - Weeks 1-4
    - Linear regression and regularizations (Lasso and Ridge)
    - Variable selection consistency and robustness to noises
  - Module 2: Robustness
    - Weeks 5-9
    - Domain Adaptation, Domain Generalization, Spurious Features, Adversarial Robustness
  - Module 3: Fairness
    - Weeks 10-12
    - Outcome Discrimination, Quality Disparity, Applications
  - Module 4: Interpretability and Privacy
    - Weeks 13-14

Expectations

- Generalization of the knowledge
  - This class will talk about the core techniques of trustworthy machine learning,
  - in the context of
    - Linear models
    - Deep learning models
    - Large language models
  - We will try to cover most of the core techniques in the first module in linear models
    - With works from actual papers
  - We will introduce all these techniques in the context of deep learning again
  - We will also include these techniques in the context of large language models
  - In the future, for newer techniques (even newer than large language models)
    - The students should be able to develop trustworthy methods themselves.

Expectations

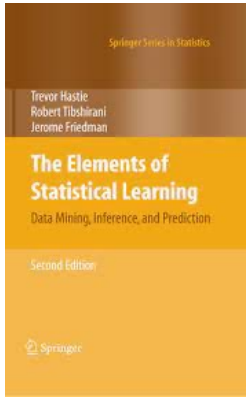
- Hands-on experience in projects
  - Team projects with 2-3 students
  - Delivery:
    - Proposal
    - Final report
    - Presentation

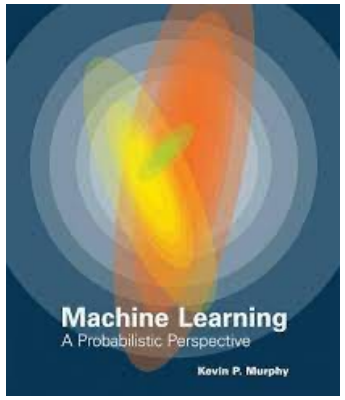
- Most importantly, to know that

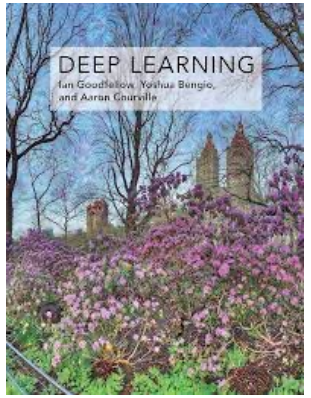
There is no magic in data science

Main Reference

- The “trustworthy” part will be based on high-impact papers of each field
- The statistical machine learning part will be based on several classical textbooks







- If you only want some high-level summary:
  - Check out our monograph:

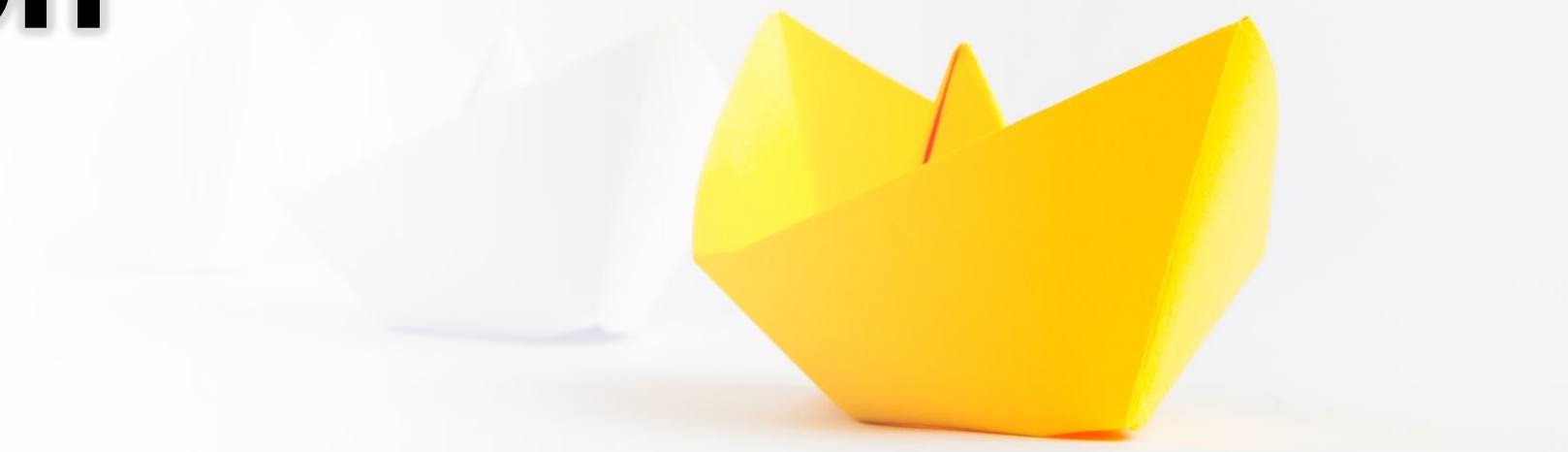
Towards Trustworthy and Aligned Machine Learning:
A Data-centric Survey with Causality Perspectives

Haoyang Liu[†], Maheep Chaudhary[†*] and Haohan Wang
School of Information Sciences,
University of Illinois Urbana-Champaign
{hl57, haohanw}@illinois.edu, maheep001@e.ntu.edu.sg
[†] equal contribution

Tentative Schedule

Module	Week	Topics	Timeline	Module	Week	Topics	Timeline
Statistical Foundation	Week 1	Introduction		Spring Break			
	Week 2	Linear Models		Fairness	Week 10	Outcome Discrimination	
	Week 3	Ridge			Week 11	Quality Disparity	
	Week 4	Lasso	HW1 out		Week 12	Application in Healthcare	HW3 out
Robustness	Week 5	Domain Adaptation	Proposal due	Privacy and Interpretability	Week 13	Federated Learning	
	Week 6	Domain Generalization			Week 14	Interpretability	
	Week 7	Spurious Features		Presentations	Week 15	Final Presentations	
	Week 8	Adversarial Robustness	HW2 out		Week 16	Final Presentations	

Introduction



Machine Learning

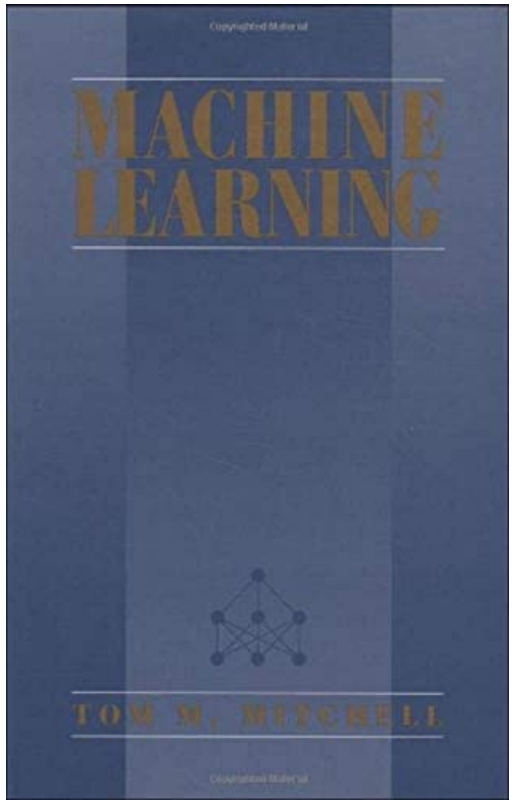
- Textbook definition

We talk about this all the time, what is machine learning after all?

A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**

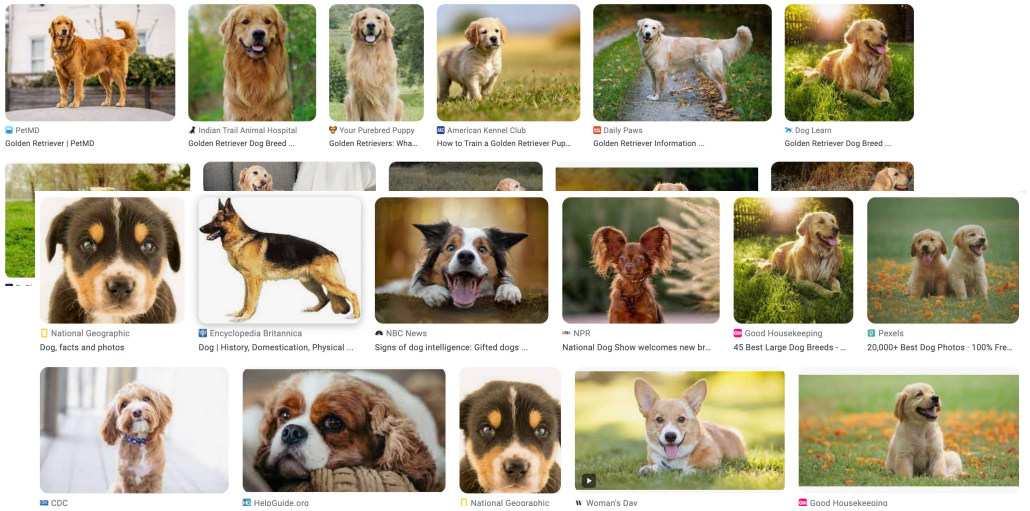
-- Tom Mitchell

Nah, too pedantic

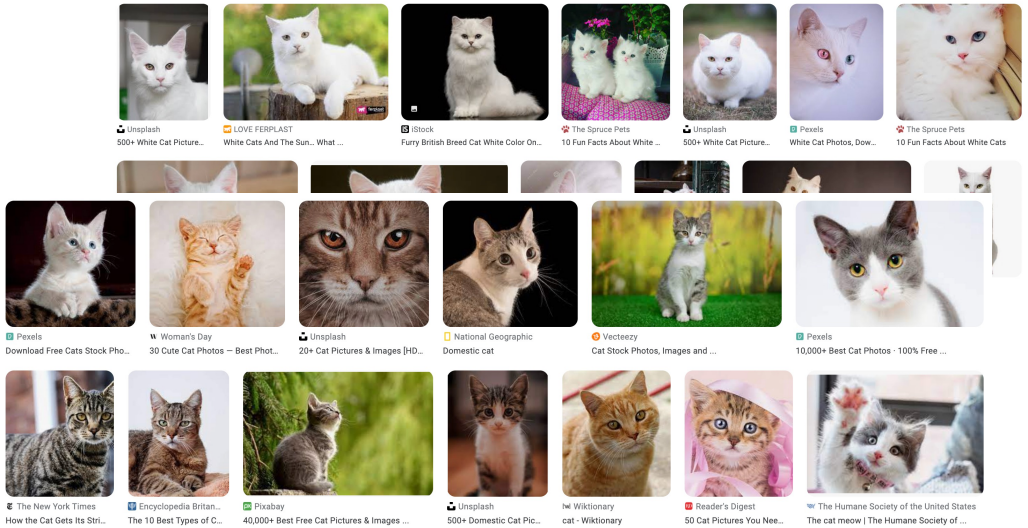


Machine Learning – the cat vs. dog example

- Let’s say we want to build a dog vs. cat machine learning model
  - Given an image, the model will tell us whether the image is depicting a dog or a cat

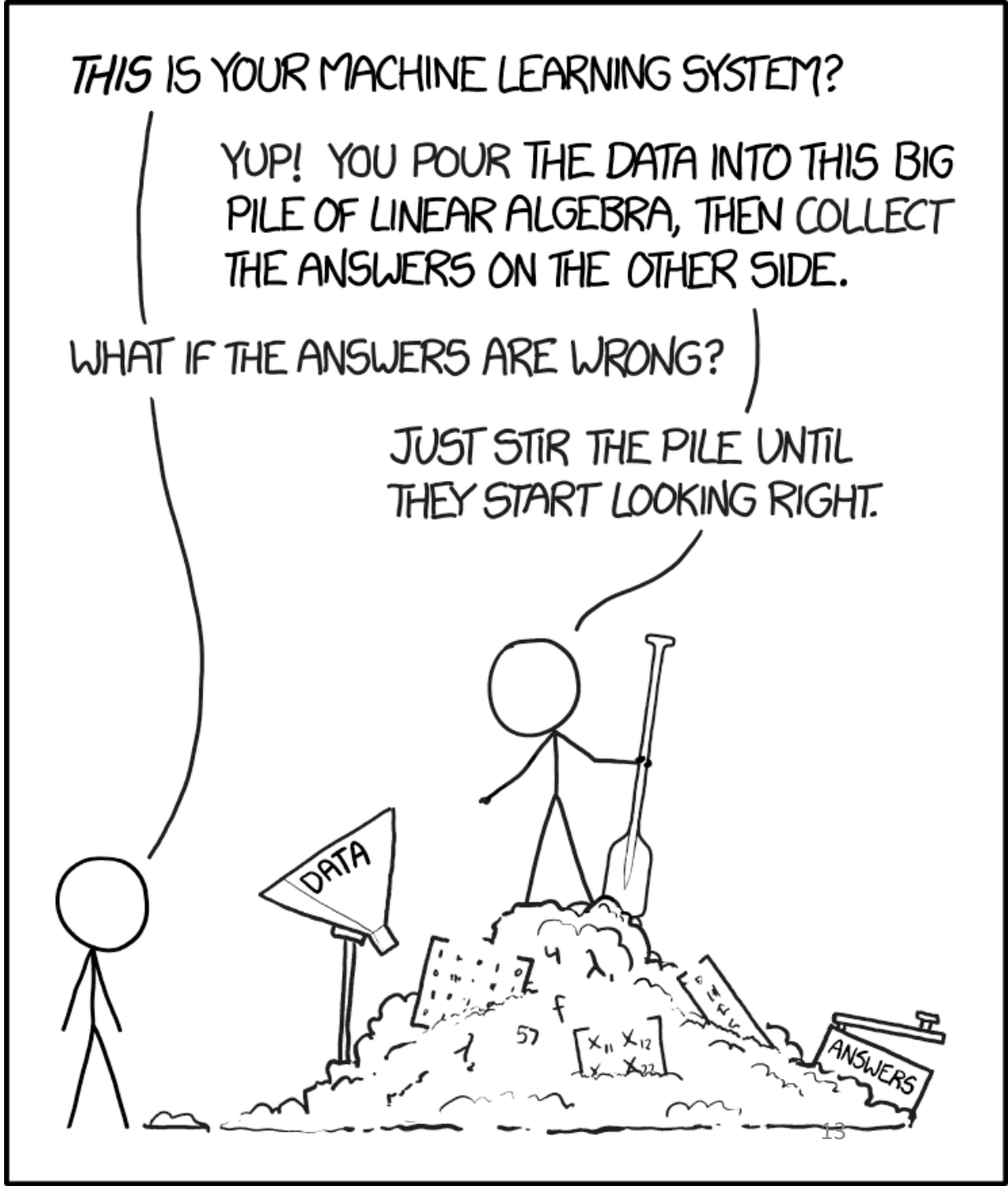


machine learning model



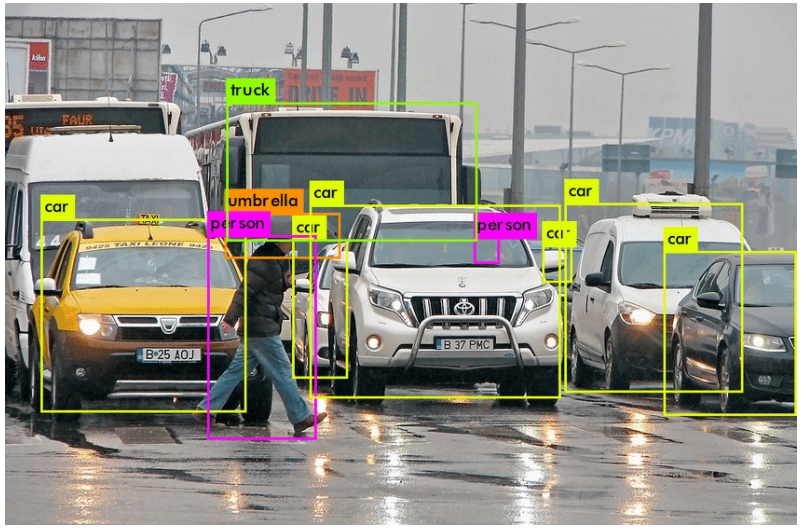
Machine Learning

- It's a fairly accurate to this field, but
  - What does "Right" mean?
    - Design of research question
    - Design of loss functions
    - Design of regularizations
  - What data to pour?
    - Data collection
    - Data preprocessing/augmentation
  - How to stir?
    - Model (building blocks) and inductive biases
    - Optimizations
    - Mathematical understandings



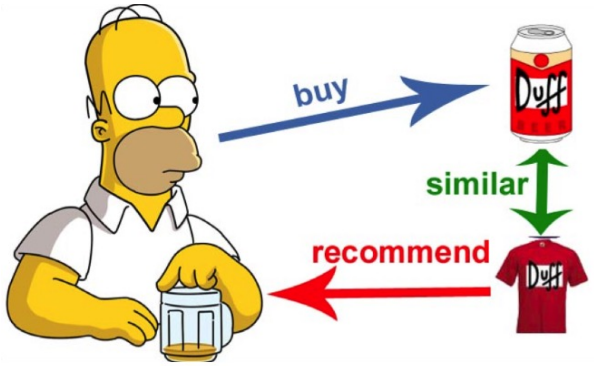
Application and Impact of Machine Learning

- Applications



Object detection

Machine translation



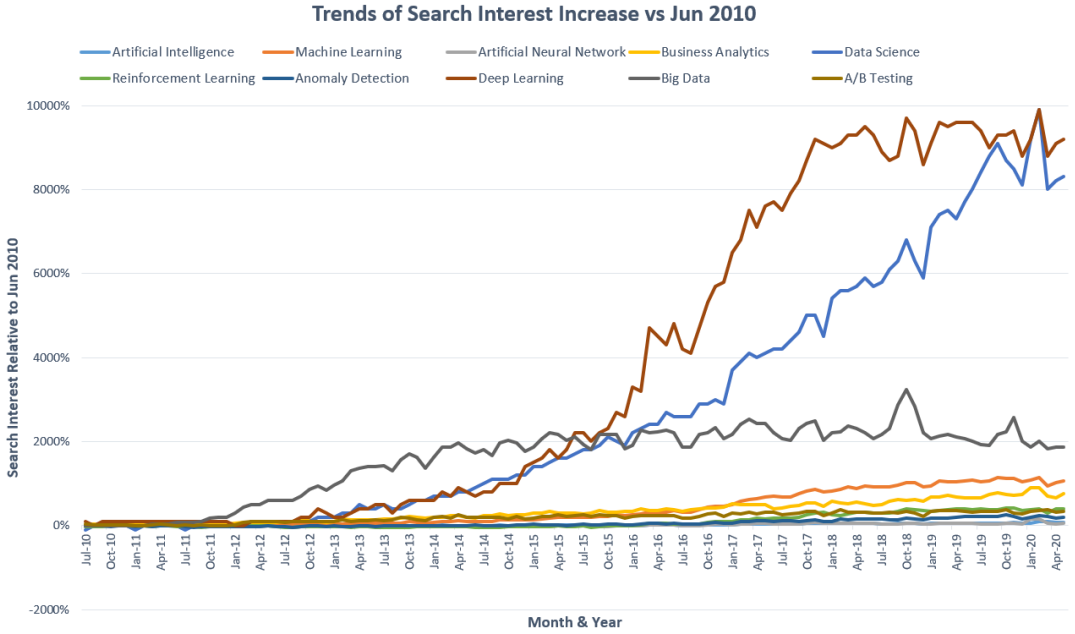
Recommendation System

Sentiment analysis

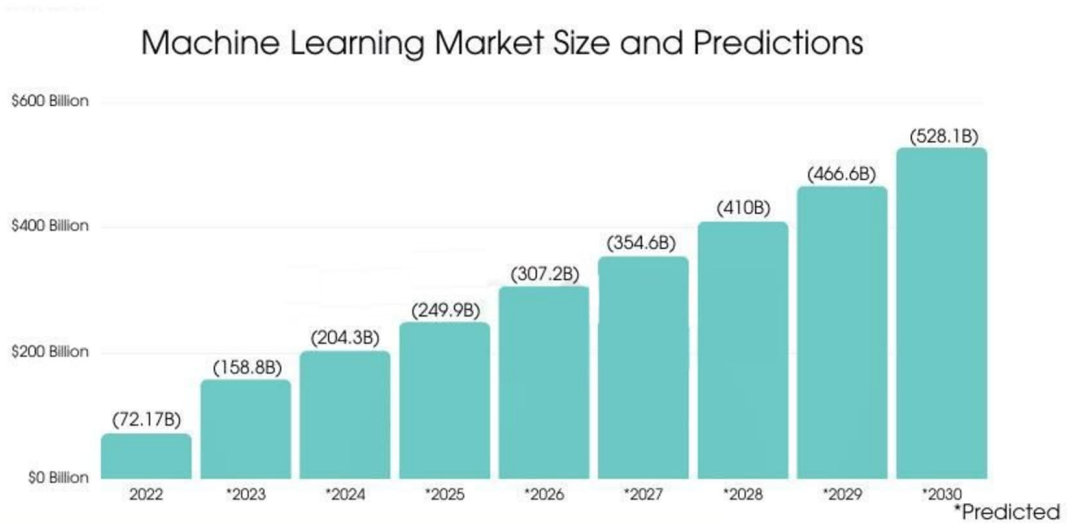
Large language models

The Unprecedented Popularity of Machine Learning

- Machine Learning has become one of the most popular and rapidly evolving fields in technology and science



Source: https://towardsdatascience.com/has-interest-in-data-science-peaked-already-437648d7f408



Source: https://whatsthebigdata.com/top-machine-learning-statistics/

Discussion

- What is so special of this machine learning topic that it can advance so fast?
  - There might be many answers
  - One possibility:
    - It heavily values the empirical results
      - No matter what the method is, as long as it outperforms previous methods, it's a good method

the “Soaring” Status of Machine Learning

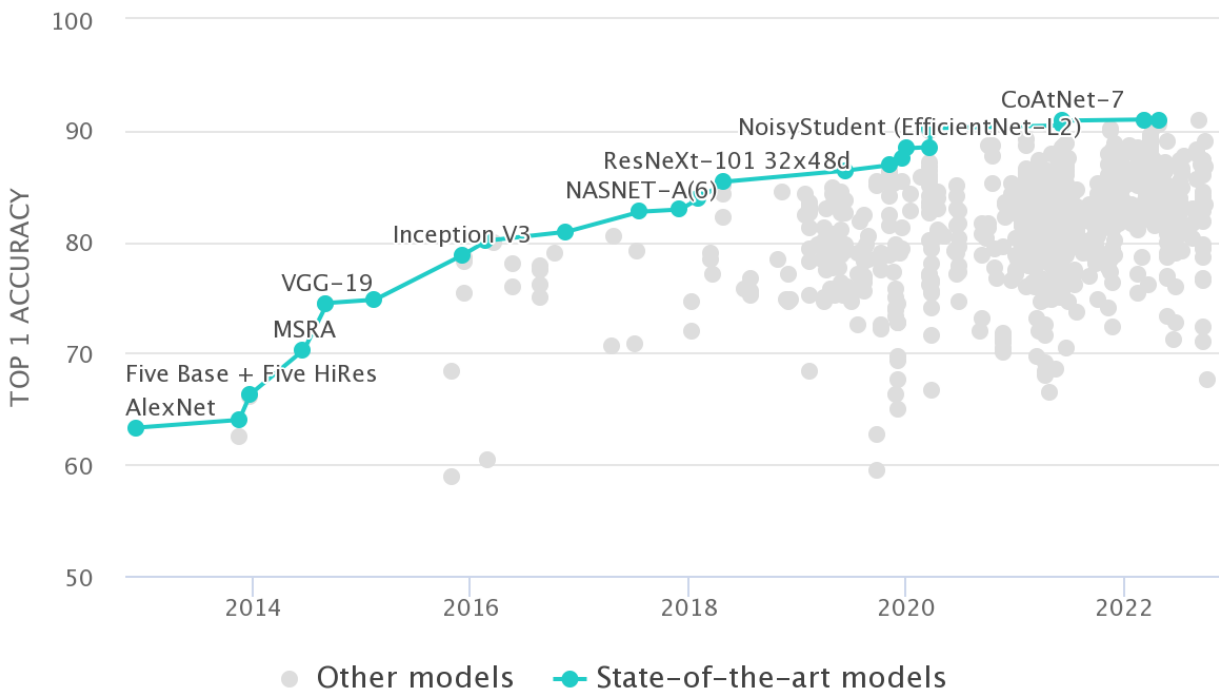
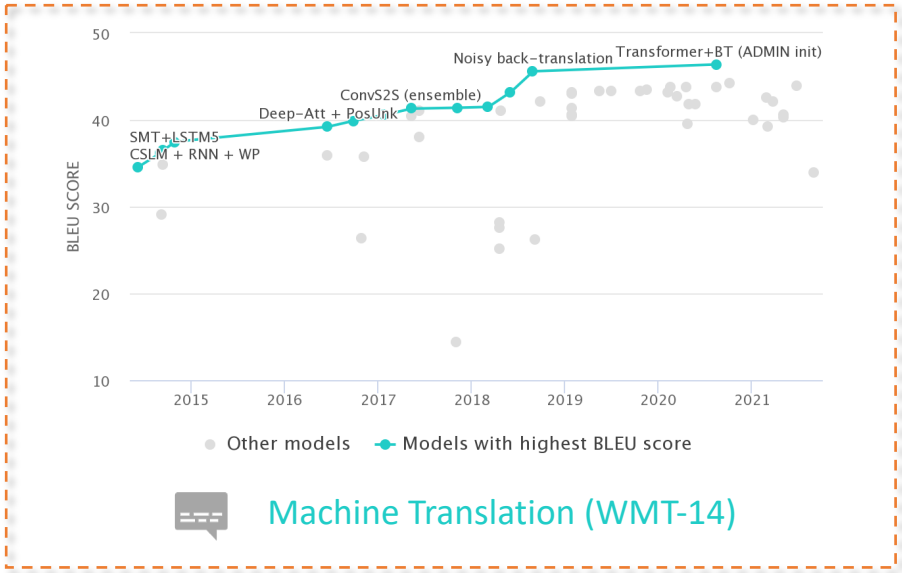


Image Classification (ImageNet)



Machine Translation (WMT-14)

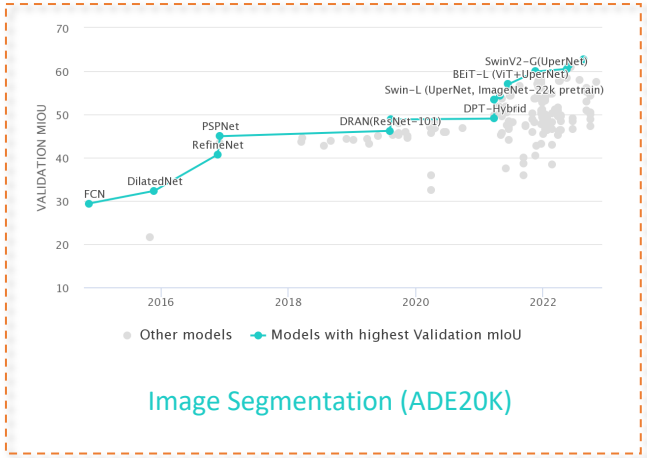
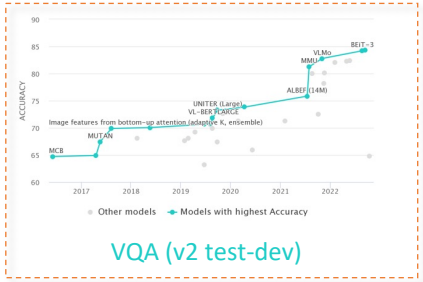
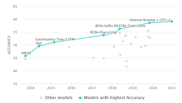


Image Segmentation (ADE20K)



VQA (v2 test-dev)



source: papers with code

However

- Highly accurate models can make mistakes
  - We often see news title like this

Tesla behind eight-vehicle crash was in 'full self-driving' mode, says driver

San Francisco crash is the latest in a series of accidents blamed on Tesla technology, which is facing regulatory scrutiny

Isn't that the models are supposed to have very good performances before they are deployed onto the road? Why are there still often crashes?

The long-lasting mindset of machine learning development

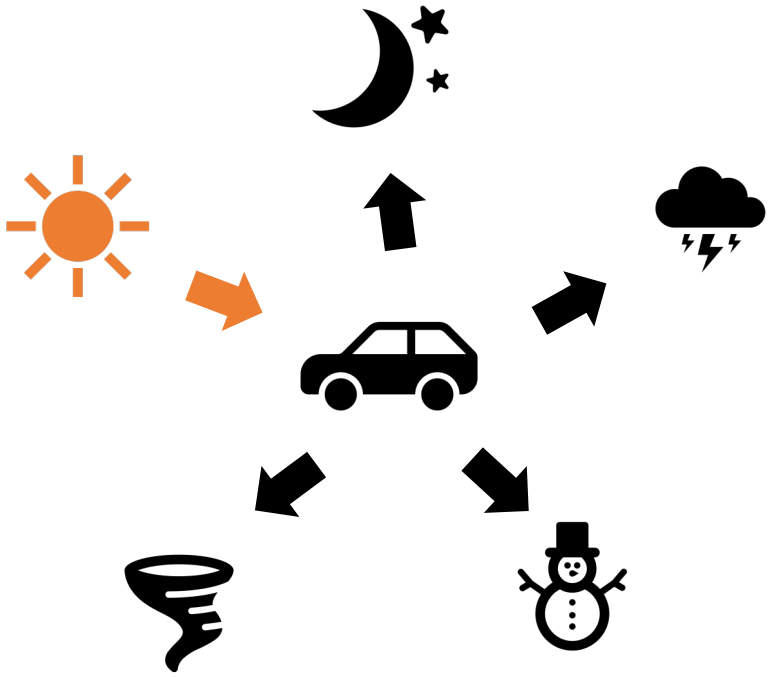
What the Real-world Needs

More than Prediction Accuracy

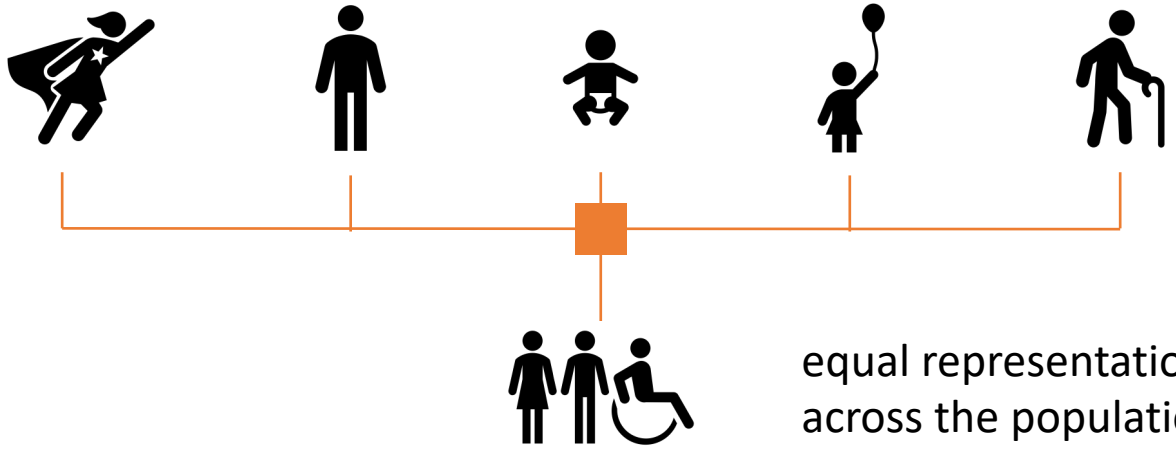
Entertaining | Social media | Automatic production | Virtual assistant | Scientific discovery | Medical care | Self-driving car

For many of these applications, prediction accuracy is not the only thing that matters

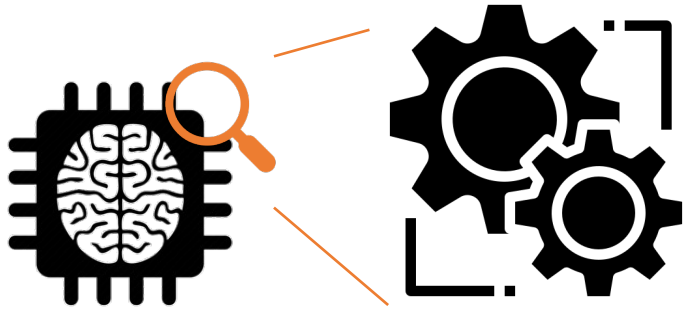
Desired Properties of ML Application



stable behavior across different situations



equal representation
across the population



working mechanism
understood by users

Trustworthy Machine Learning

- So, what is trustworthy machine learning anyway?

- There might be different textbook level definitions
- Here, we use “trustworthy” as an umbrella term that encompasses several properties of machine learning, e.g.,
  - Robustness
  - Fairness
  - Privacy-preserving
  - Interpretable
  - Transparent

With these definition, can we train a machine learning model?

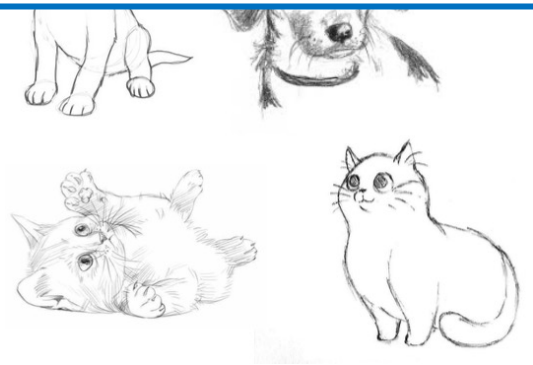
What do these terms mean?

Robustness

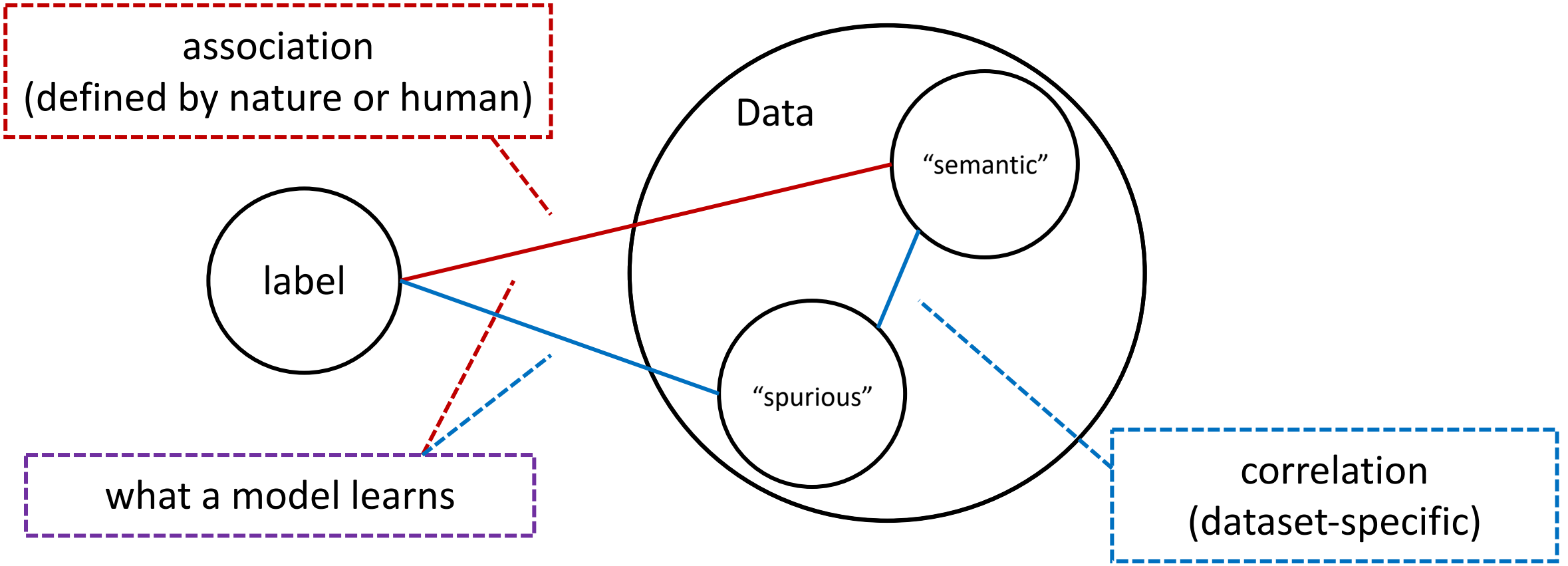
What are the challenges here?

- Even robustness can mean many different things
  - Let's start with a simple thing
  - Our aim is to train this machine learning model on color images and hope it can generalize to sketch:

If we can get high-accuracy on the training set, does it mean we learn the cat vs. dog concepts? What could be the issues behind this?

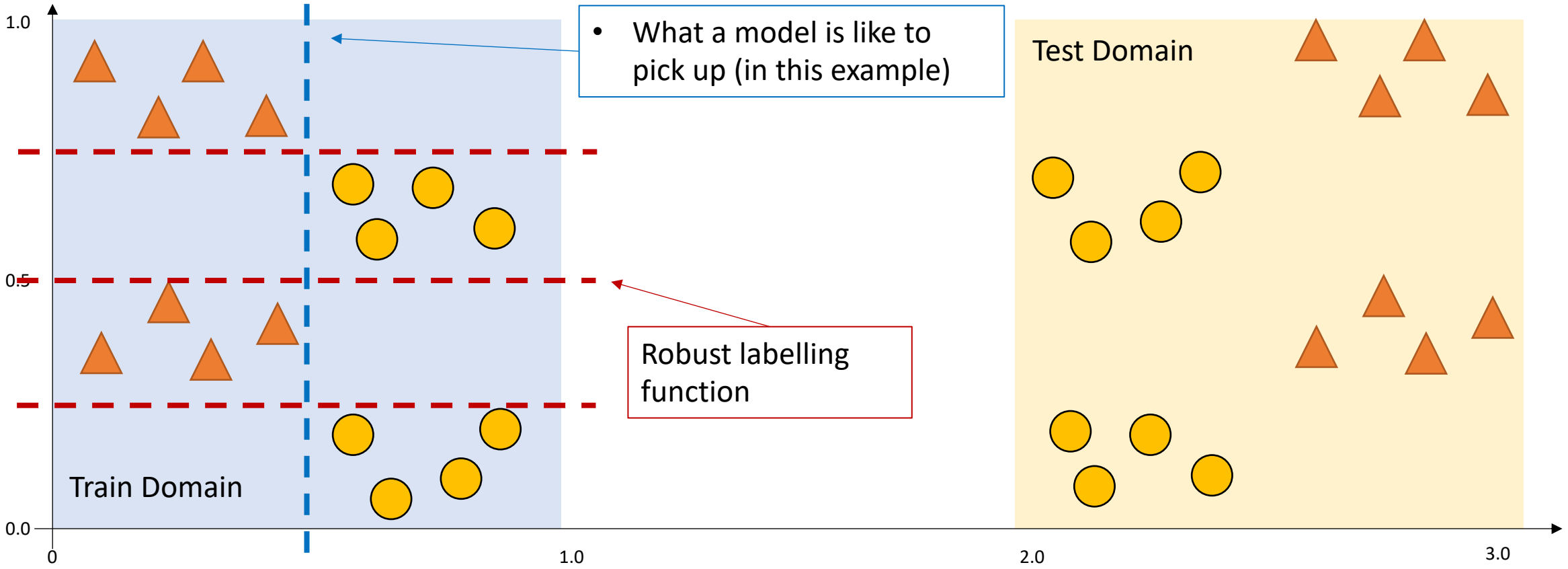


A Conjecture: high-accuracy but do not generalize

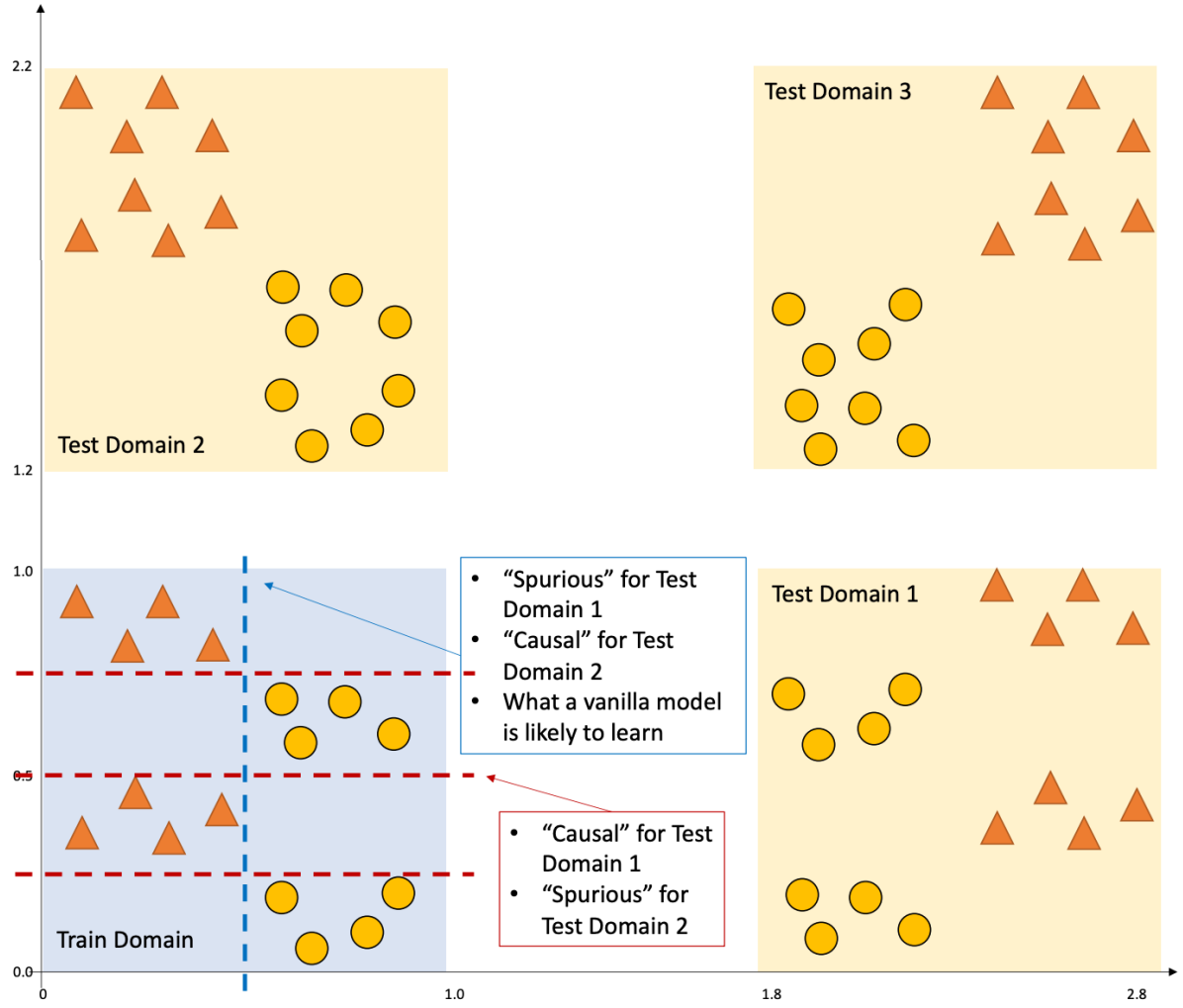


The Challenge Illustrated

Does this mean we should always drop the easiest learning function and picker the harder one?



The Challenge Illustrated with More Dimension



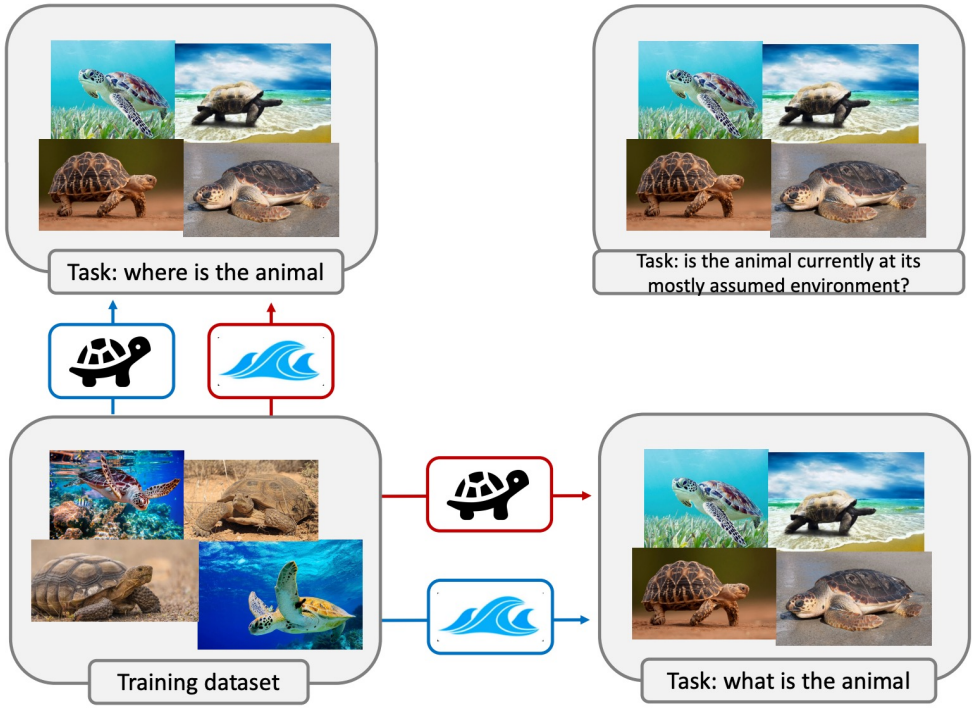
What is a good function depends on what do we want

In other words, there won't exist a method, that help learn a robust model without context

The Challenge Illustrated with A More Concrete Example

- In a task of predicting sea turtles and tortoise
  - The background can highly correlate with the animal

VS

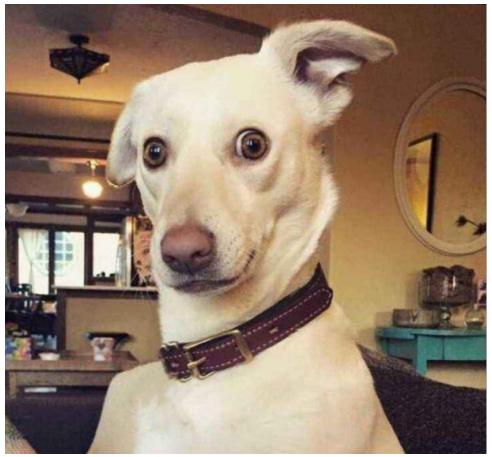


Machine Learning Robustness

- Machine Learning Robustness
  - Must specify what the robustness is against (the context of the problem)
    - If you find a paper that talks about robustness in genera, claim it will learn a robust model across all datasets and applications
      - Feel free to throw it away, it's likely lying.

- Can we say the same thing for other trustworthy properties?
  - Interpretability
  - Fairness
  - Privacy

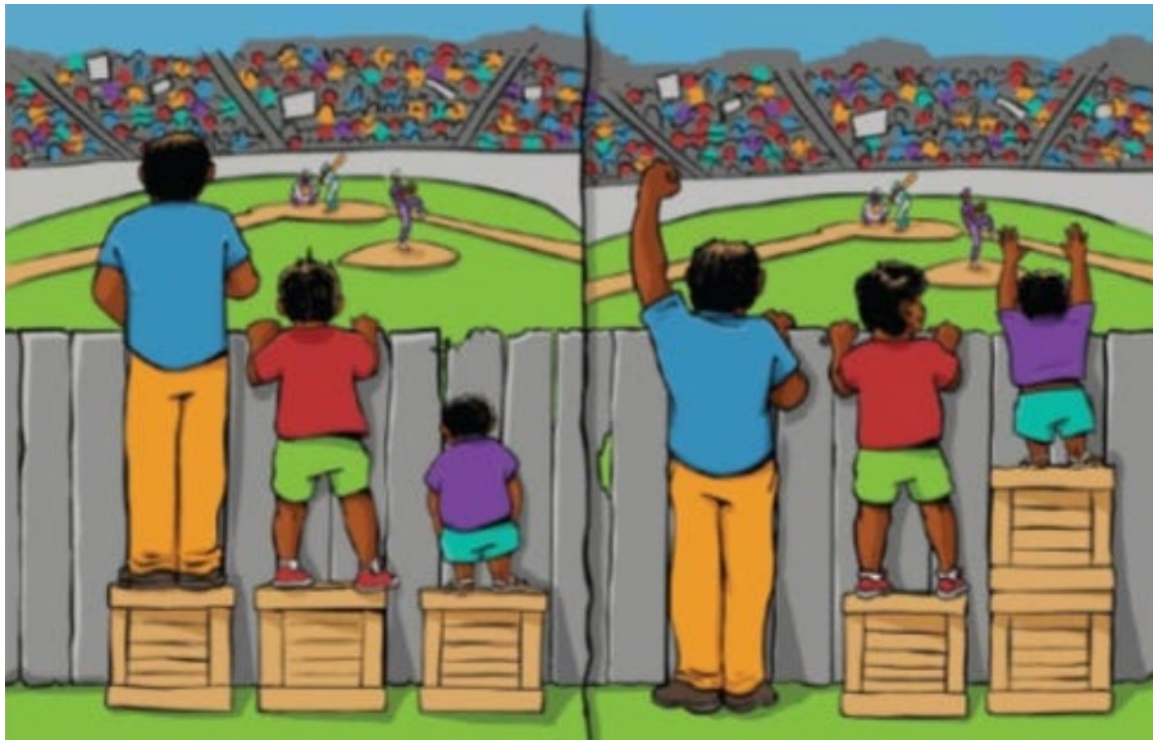
Interpretability



- Probably also subjective
  - First of all:
    - It must be interpretable to the users
      - Whether one can understand it can be subjective
  - If a machine learning mode is interpretable to a dog, but not to human, can we call it interpretable?
- Later we will see two examples:
  - One paper argues that the fewer features a model use, the more interpretable the model is, so they propose a method that introduces sparsity regularization, and their proposed method outperform every others.
  - One paper argues that the more connected the features a model use, the more interpretable the model is, so they propose a method that introduces smoothness regularization, and their proposed method outperform every others

Fairness

- What is fair is likely subjective
  - E.g., equal opportunity vs. equal outcome
  - Later we will see many more different examples

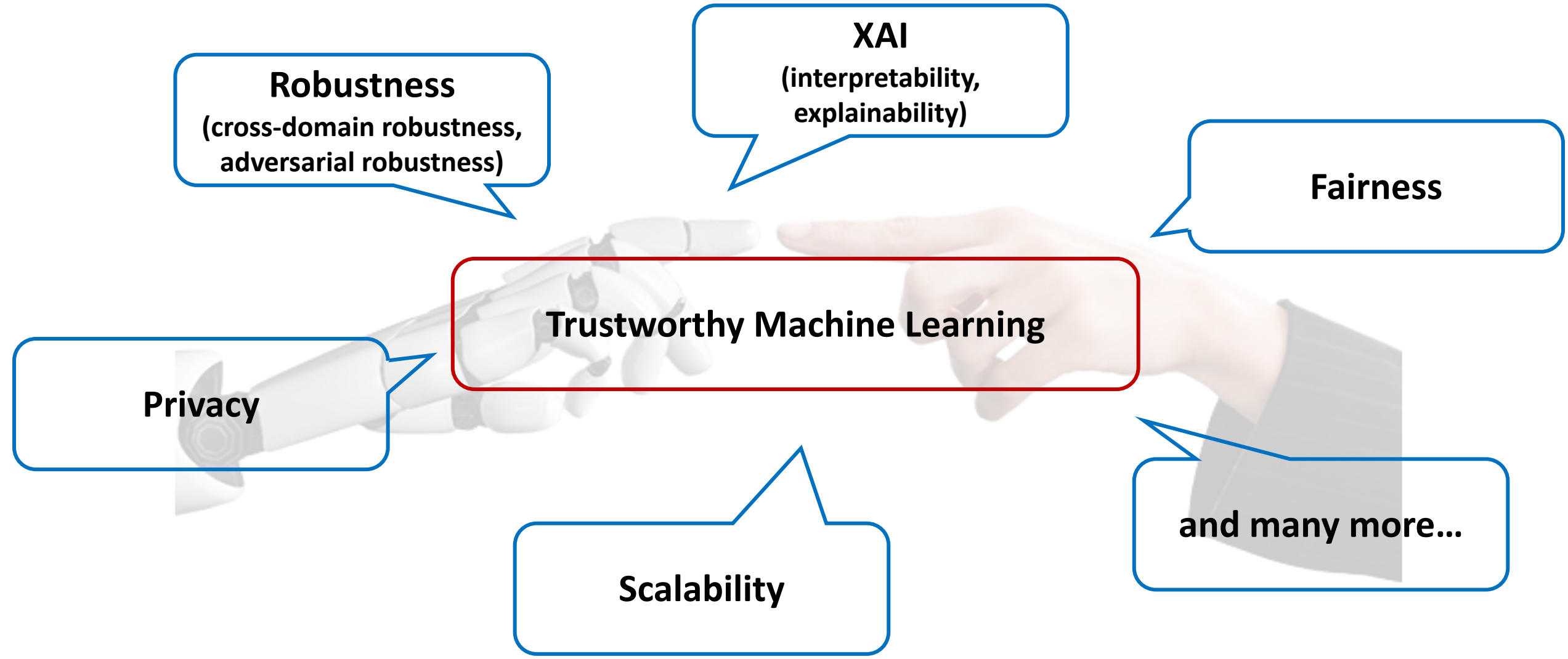


Trustworthy Machine Learning

- **Trustworthy machine learning cannot be studied without prior knowledge**
  - There must be a definition that help specify the subjective part
  - Then, the method is built upon this definition

There is no magic in data science


Solutions

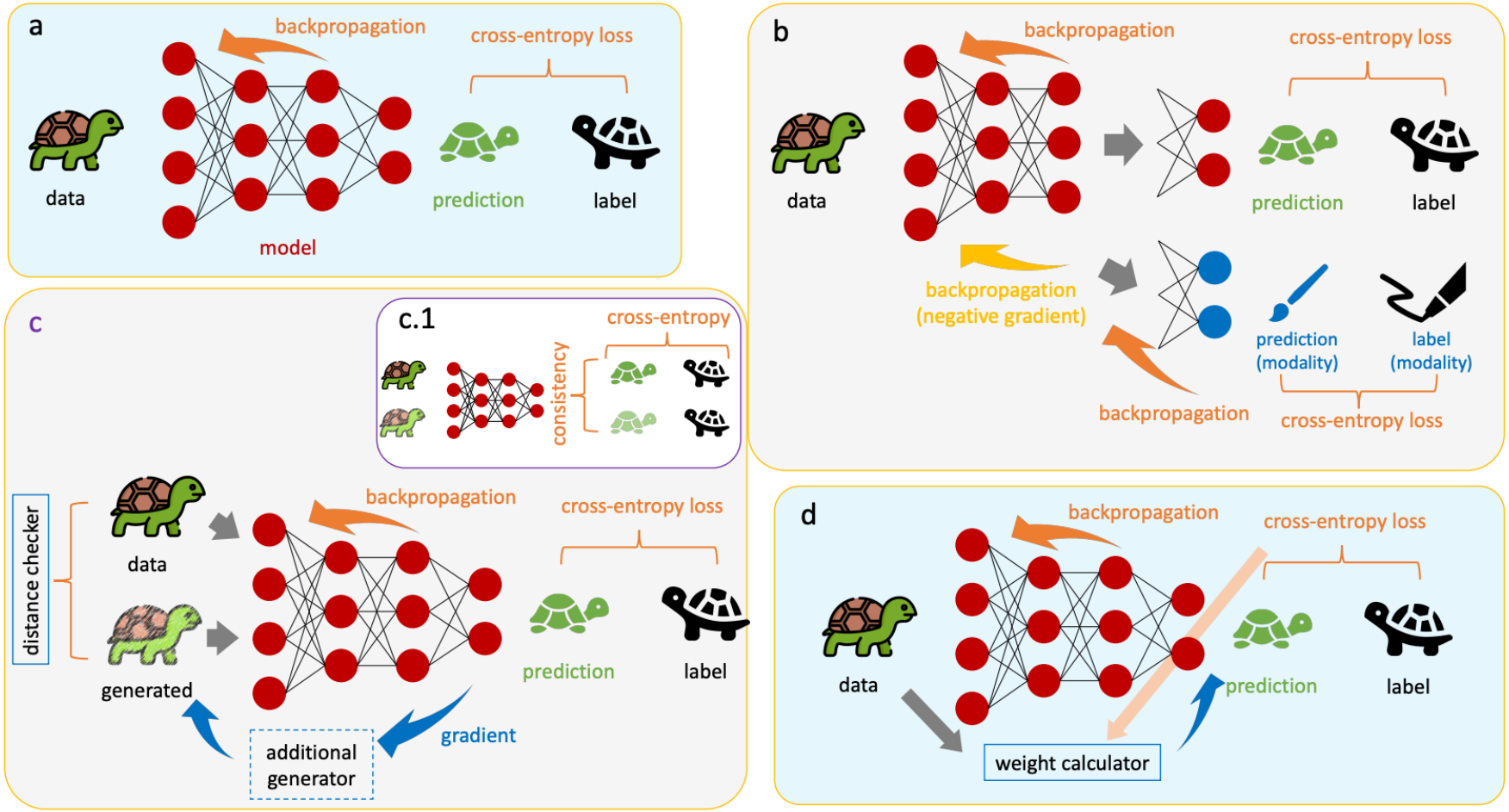

Robustness
(cross-domain robustness, adversarial robustness)
XAI
(interpretability, explainability)
Fairness
Trustworthy Machine Learning
Privacy
Scalability
and many more…

Trustworthy Machine Learning

- There are so many different topics under the umbrella of trustworthy
- There are also many techniques under each of these topics
- Will we cover all of them?
  - We will try to touch many of them
  - But not because we want to memorize everyone of them
  - We only want to touch many enough so that we can summarize the universal understanding behind it.

Trustworthy Machine Learning

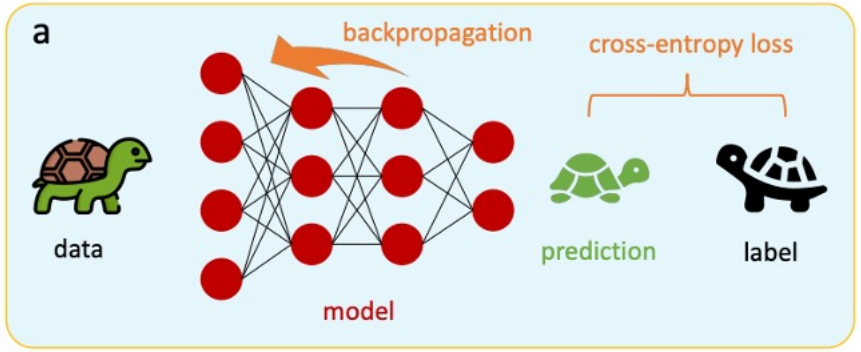
- Universal Solutions



Trustworthy Machine Learning

- Universal Solutions
  - A Baseline Network

$$\arg\min_{\theta} \frac{1}{n} \sum_{(x,y)\in(\mathbf{X},\mathbf{Y})} l(f(x;\theta),y)$$



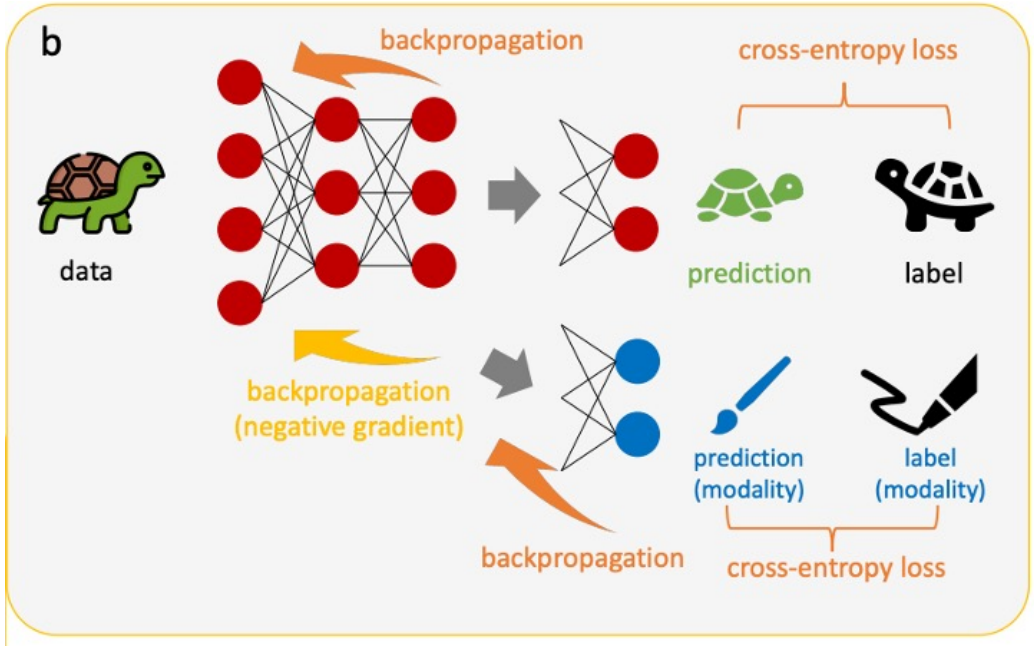
  - Make sure you are comfortable with the notations:
    - Data
    - Model and parameters
    - Loss functions
    - argmin

Trustworthy Machine Learning

- Universal Solutions
  - Through invariance regularizations

$$\arg\min_{\theta} \frac{1}{n} \sum_{(x,y)\in(\mathbf{X},\mathbf{Y})\cup(\mathbf{Z},\emptyset)} l(f(x;\theta),y) - \lambda l(h(f_k(x;\theta);\phi),d),$$

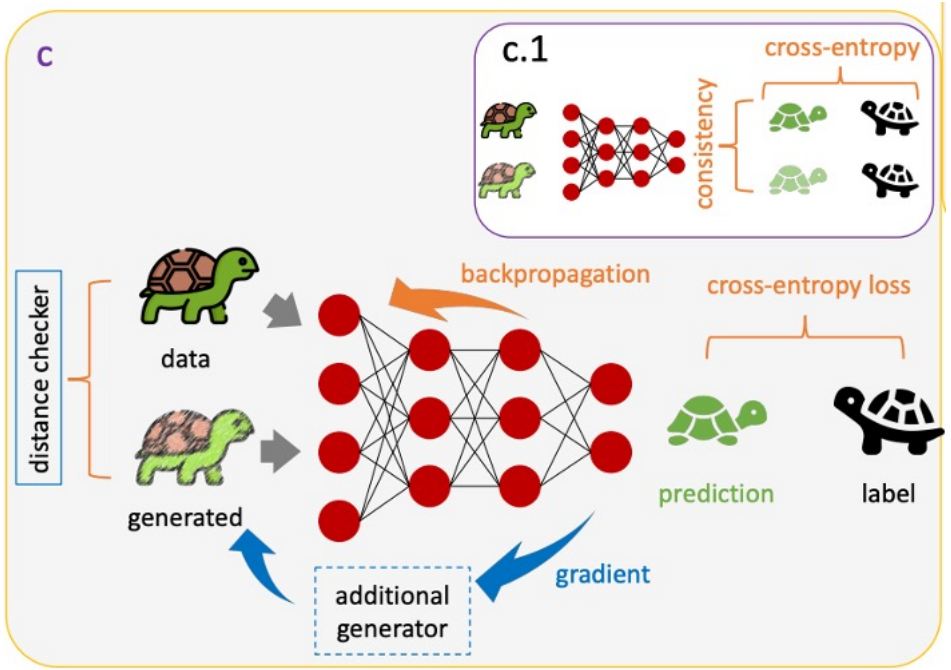
$$\arg\min_{\phi} \sum_{(x,y)\in(\mathbf{X},\mathbf{Y})\cup(\mathbf{Z},\emptyset)} l(h(f_k(x;\theta);\phi),d),$$



Trustworthy Machine Learning

- Universal Solutions
  - Through worst-case augmentation

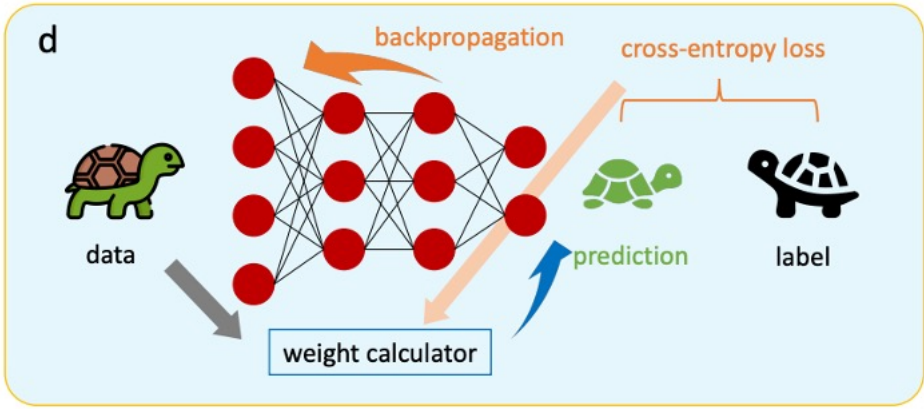
$$\arg\min_{\theta} \frac{1}{n} \sum_{(x,y)\in(\mathbf{X},\mathbf{Y})} \max_{x';d(x',x)\leq\epsilon} l(f(x';\theta),y).$$



Trustworthy Machine Learning

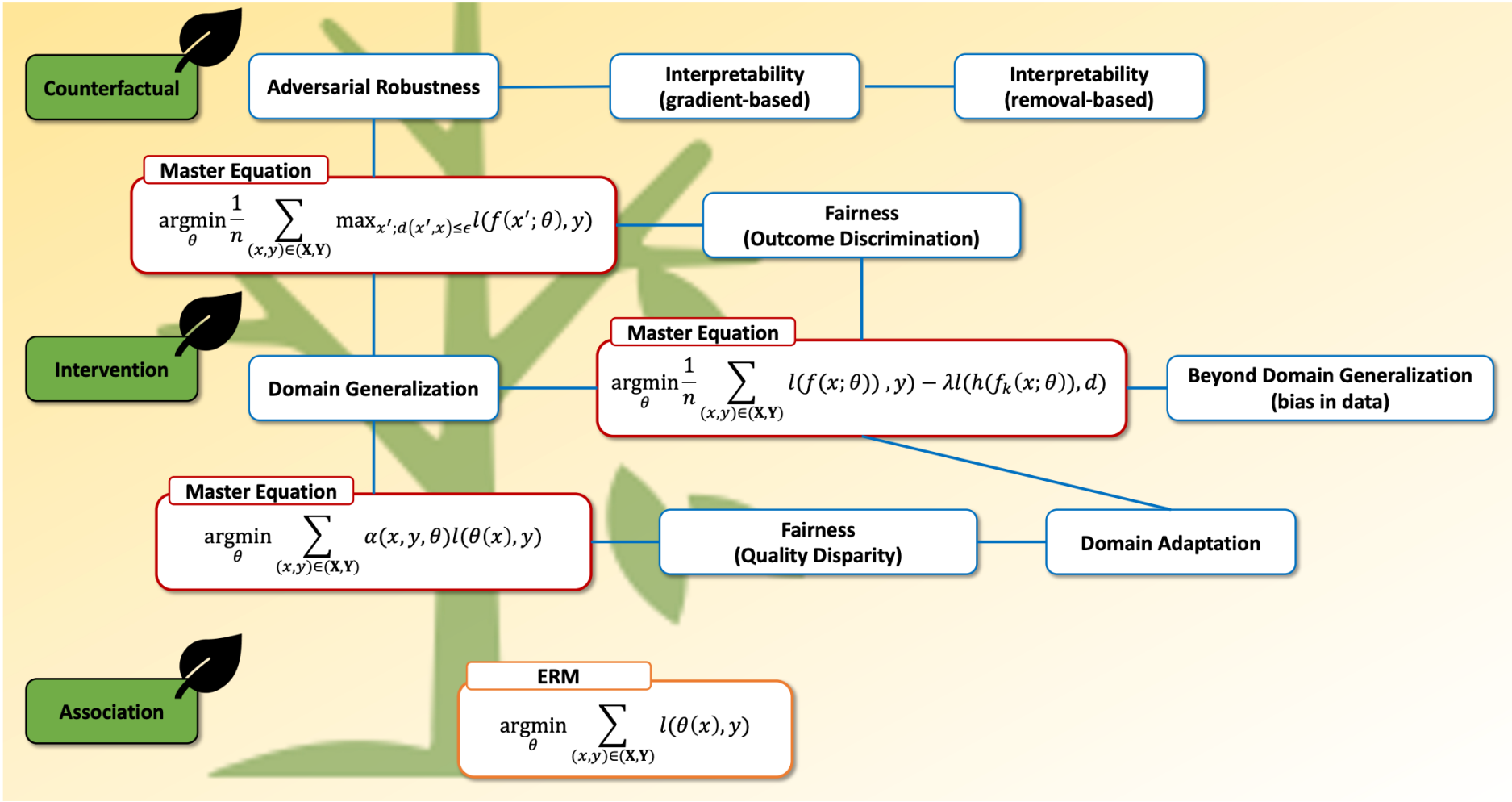
- Universal Solutions
  - Through sample re-weighting

$$\underset{\theta}{\operatorname{argmin}} \sum_{(x,y)\in(\mathbf{X},\mathbf{Y})} \alpha(x,y,\theta) l(\theta(x),y)$$



Trustworthy Machine Learning

- Universal Solutions



Methods for Pretrain Models

- Some methods that are relevant:
  - Fine-tuning
  - Parameter efficient fine-tuning
  - Prompting
    - Machine learning generated prompts

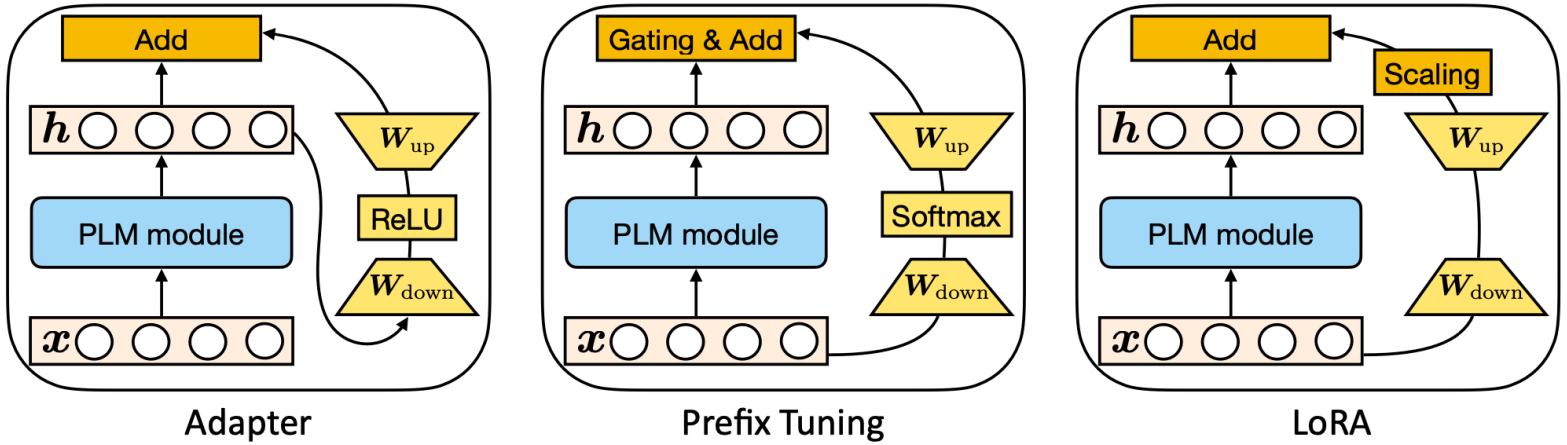
Fine-tuning

- A continued train of pretrained models over new data
  - So, basically ERM, with a new initialization

$$\text{argmin}_\theta \mathbb{E}_{(x,y)} l(f(x;\theta), y)$$

Parameter-efficient Fine-tuning

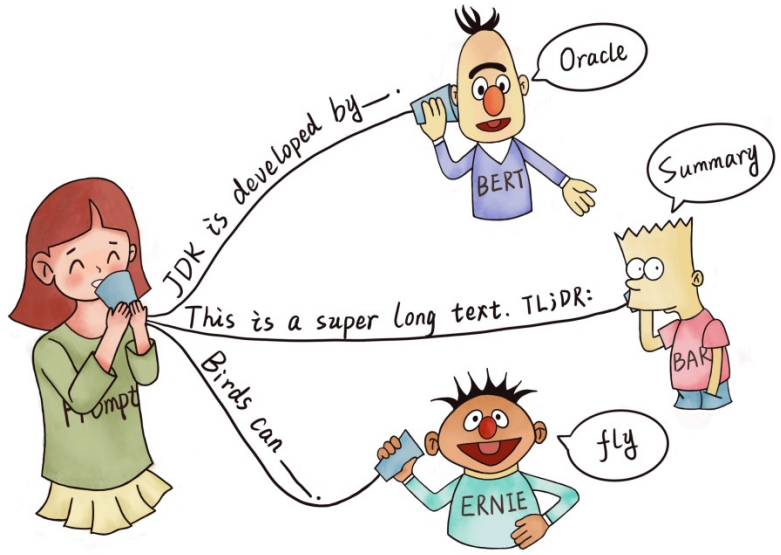
- Plug-in a component to the new model and fine-tunes the new component



$$\text{argmin}_{\theta}\, \mathbb{E}_{(x,y)}\, l(f(x; [\Theta; \theta]), y)$$

Prompting (auto-generated prompting)

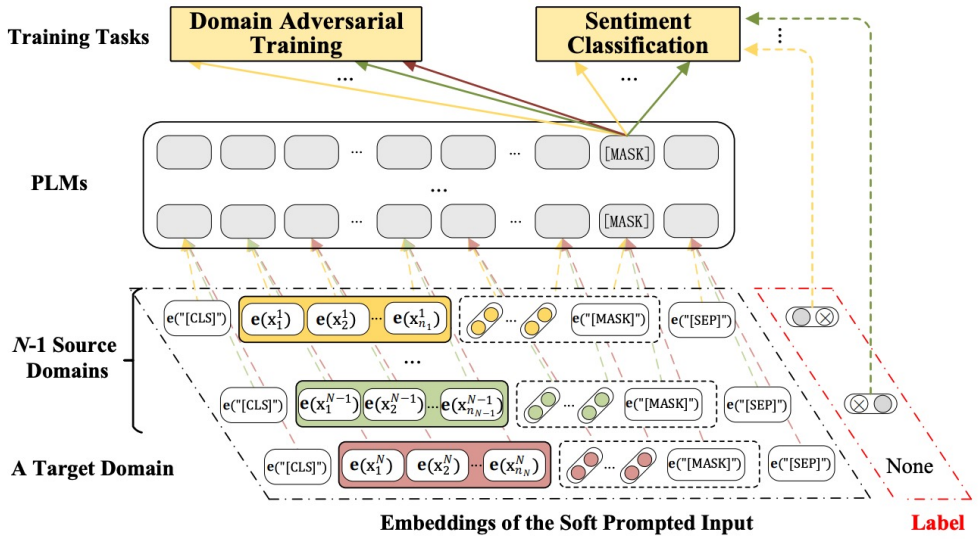
- How do we find an algorithm that can generate prompts to always induce trustworthy output

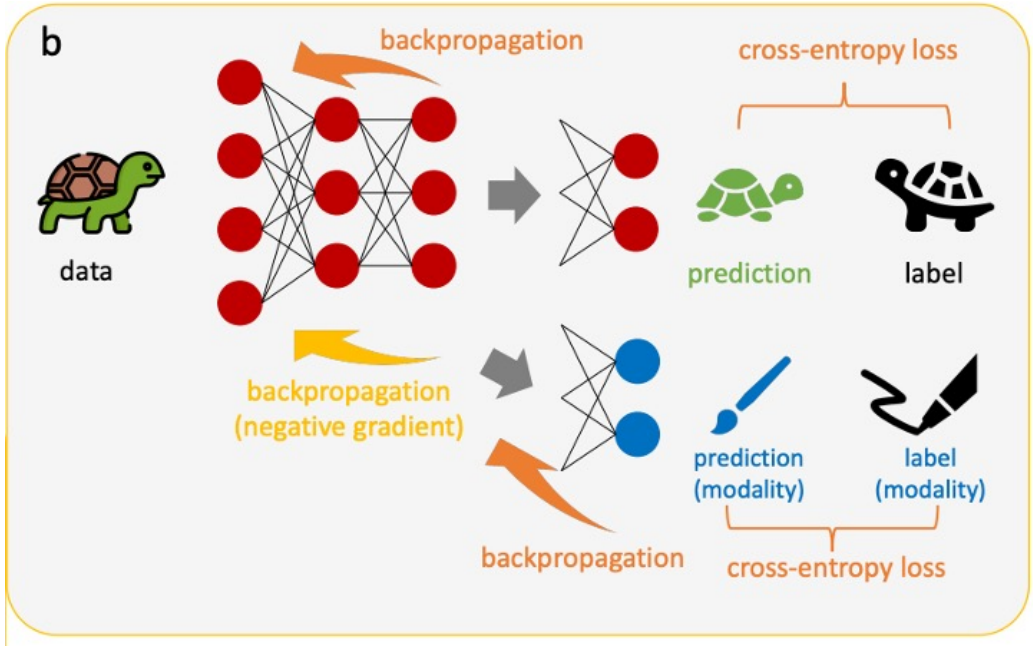


$$\text{argmin}_\theta \mathbb{E}_{(x,y)} l(g(f(x;\theta);\Theta), y)$$

Example Solutions for Large Models

- With invariance regularizations

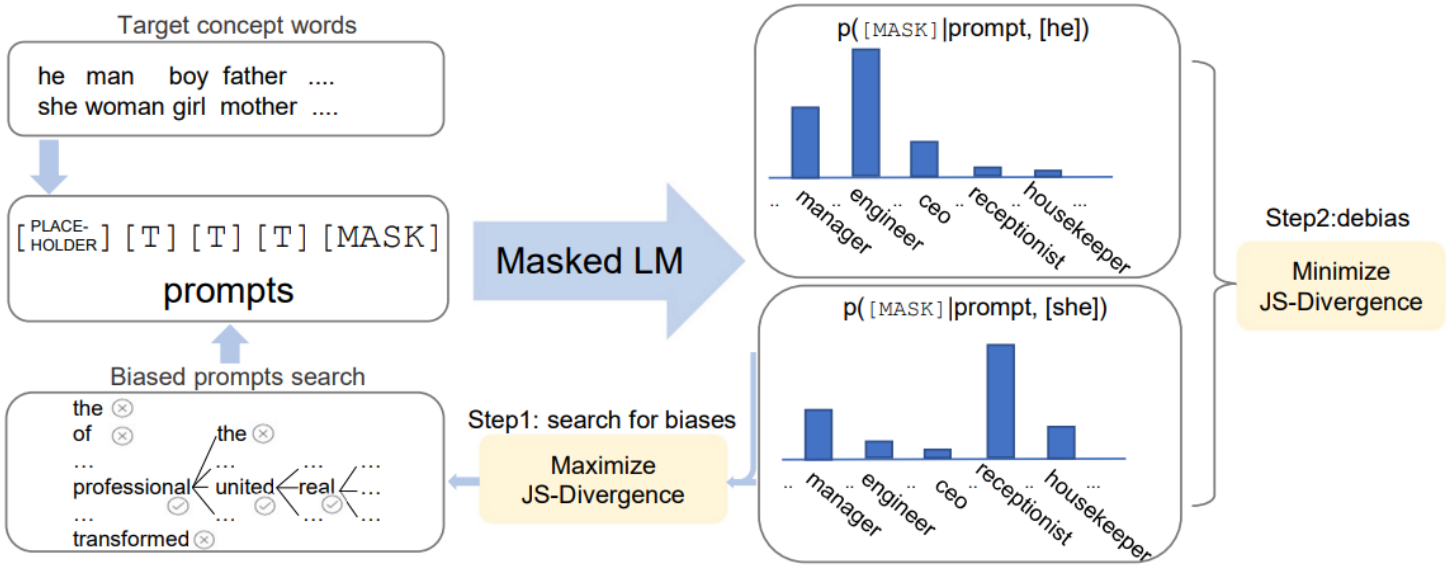


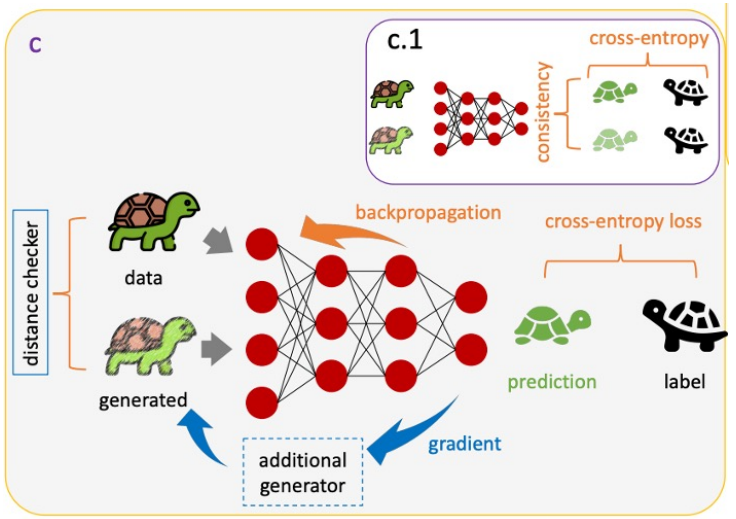


Wu, Hui, and Xiaodong Shi. "Adversarial soft prompt tuning for cross-domain sentiment analysis." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Example Solutions for Large Models

- With worse-case generation and consistency loss





Guo, Yue, Yi Yang, and Ahmed Abbasi. "Auto-debias: Debiasing masked language models with automated biased prompts." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Trustworthy Machine Learning

- Take-aways:
  - It must be studied with context
    - Since the definition of trustworthy is potentially subjective
  - It is an umbrella term that encompasses many subjects like robustness, fairness etc.
  - There are universal solutions across different topics
  - With such principled understanding of universal solutions,
    - We can even predict future machine learning solutions
      - For large language models and even beyond
        - Probably applicable to what has not been invented yet.
      - Check out our plan-to-release blog “a thousand ideas for your next paper on trustworthy machine learning”
  - This class is not just about more details of the above, but also about the deep thoughts that lead to the above.

Trustworthy Machine Learning

- If you want to have something to read for what has been discussed so far,
- Check out our monograph

Towards Trustworthy and Aligned Machine Learning:

A Data-centric Survey with Causality Perspectives

Haoyang Liu[†], Maheep Chaudhary[†*], and Haohan Wang

School of Information Sciences,
University of Illinois Urbana-Champaign
{hl57, haohanw}@illinois.edu, maheep001@e.ntu.edu.sg

[†] equal contribution

Agenda

- Logistics
- Introduction of Trustworthy Machine Learning
- Universal Solutions of Trustworthy Machine Learning
- Project Ideas