Course Orientation

Principles of Deep Learning (AI502/IE408/IE511)

Sungbin Lim (UNIST AIGS & IE)

Instructors & TA

Prof. Lim, Sungbin

Instructor

Yoon, Eunbi

TA

Kim, Seungwoo

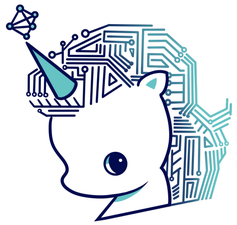
TA

Yoon, Taehyun

TA

Contact: ai502deeplearning@gmail.com

Research Area of LIM Lab



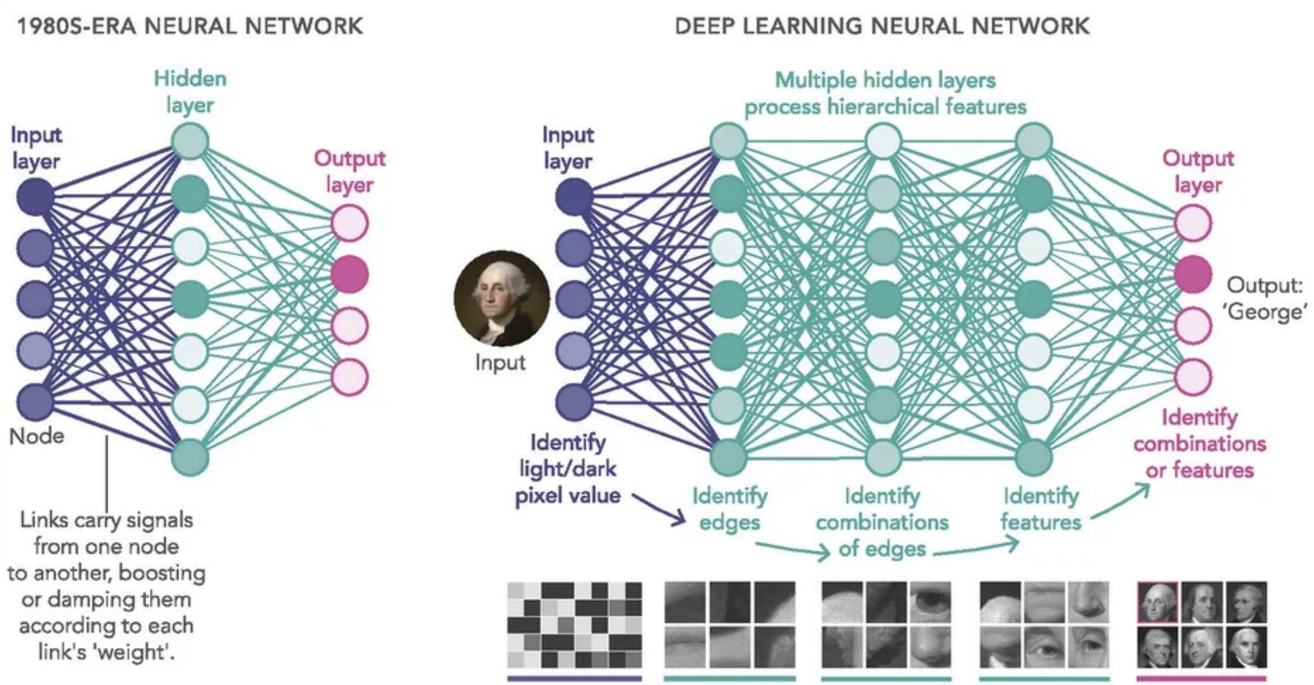
Learning Intelligent Machine Lab

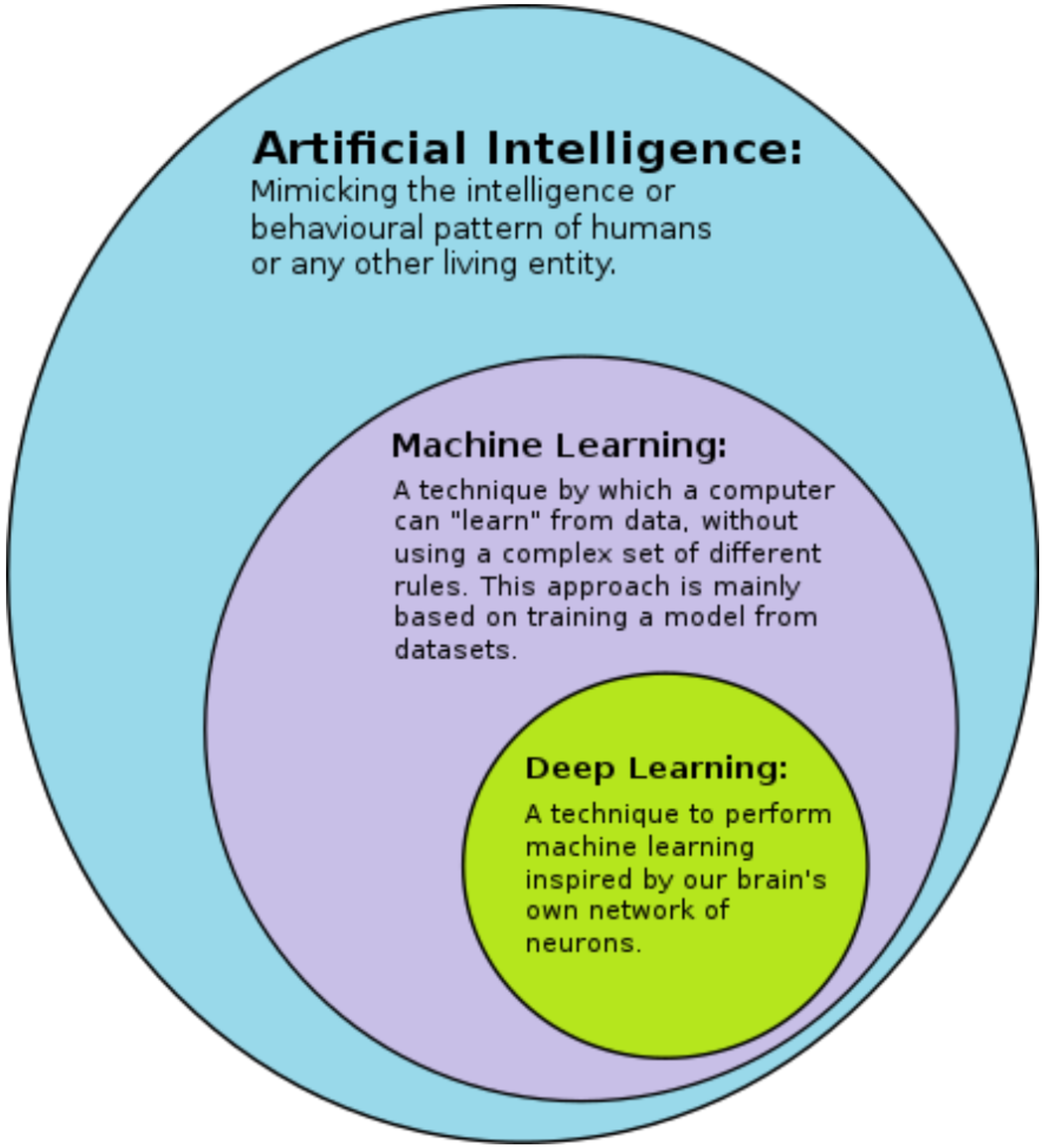
- Deep Learning
- Statistical Learning
- Stochastic Optimization
- Reinforcement Learning / Planning
- Causal Learning / Machine Reasoning
- AI Applications
  - robotics
  - operation research

Motivation

The Era of Deep Learning

- What is Deep Learning?

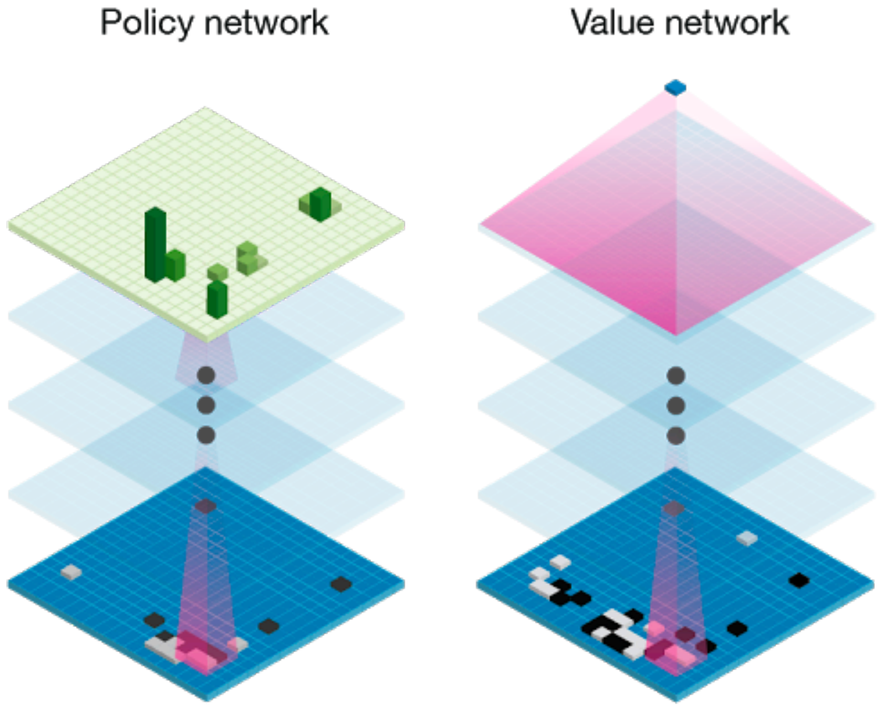




The Era of Deep Learning

- What is Deep Learning?
- Why do people pay attention to deep learning today?

AlphaGo



The Era of Deep Learning

- What is Deep Learning?
- Why do people pay attention to deep learning today?

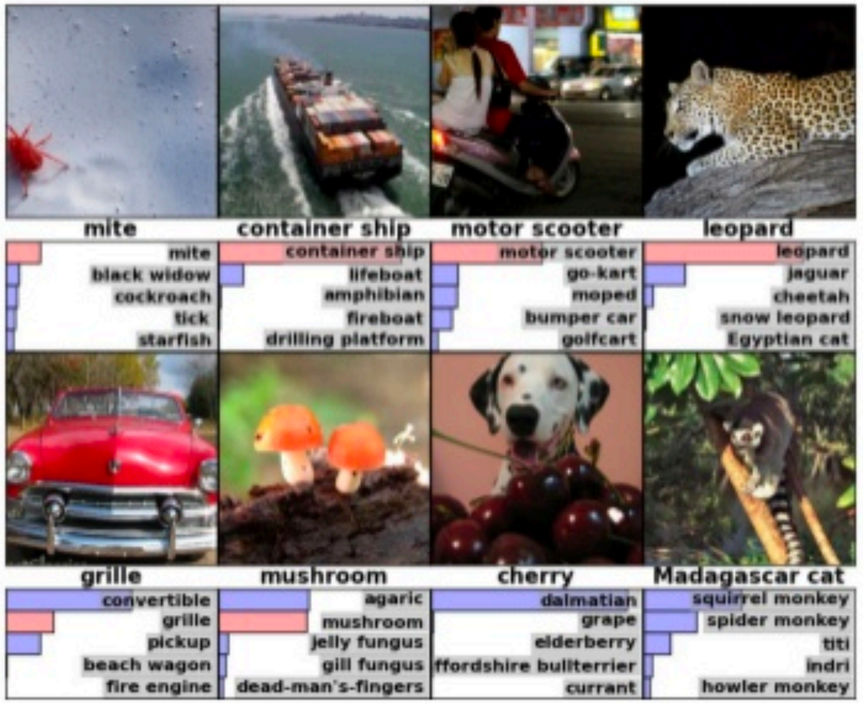
The Era of Deep Learning

- What is De
- Why do pe

ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



4

The Era of Deep Learning

- What is Deep Learning?
- Why do people pay attention to deep learning today?

ImageNet Classification with Deep Convolutional Neural Networks

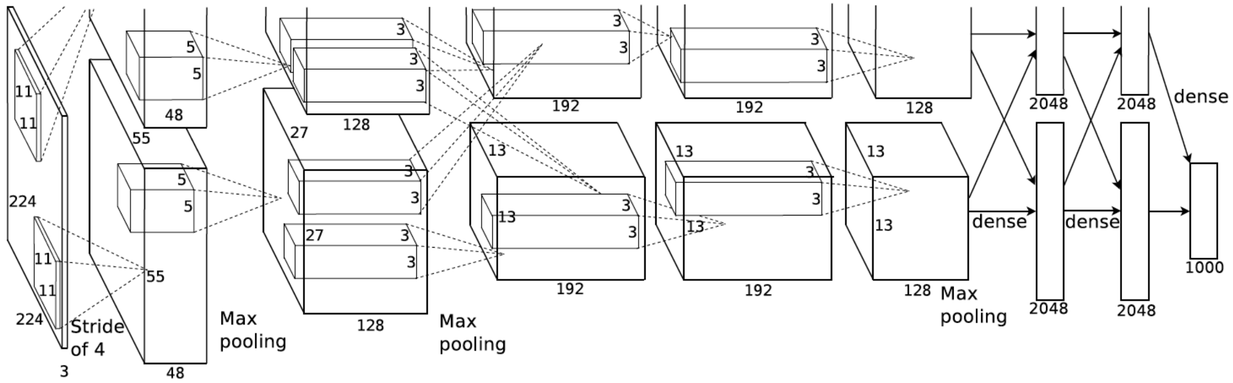
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

115,000 citations (2012, NIPS)

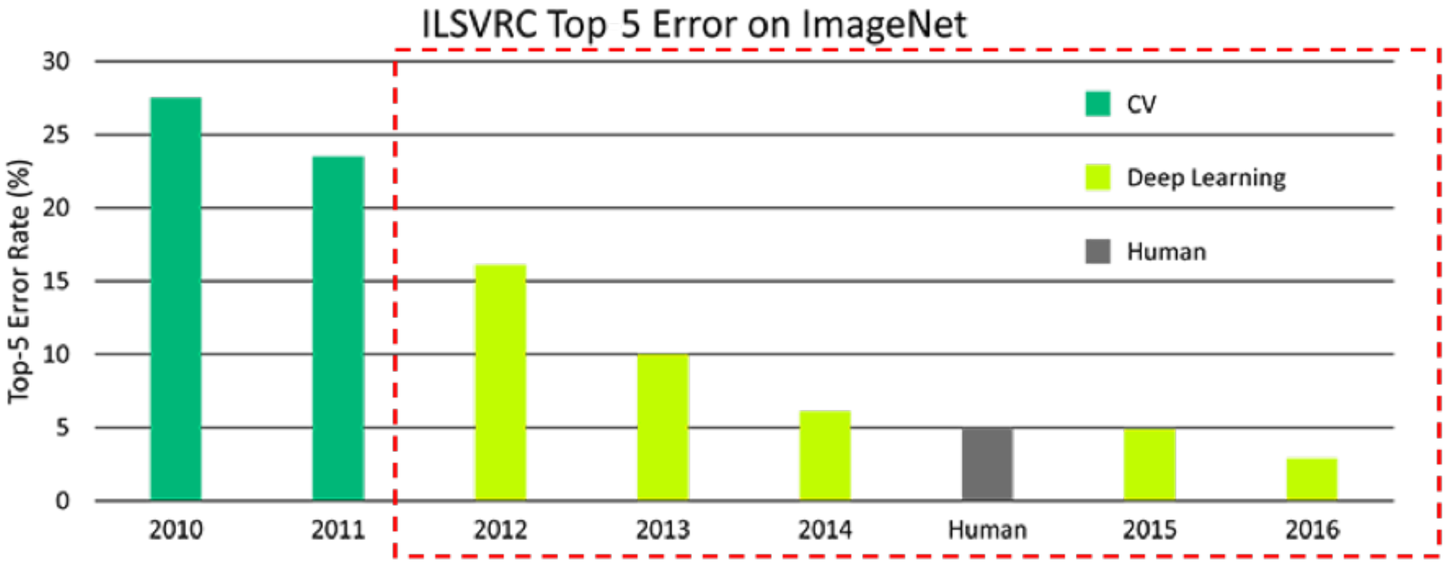
cited more than
DNA, special relativity theory..



AlexNet

The Era of Deep Learning

- What is Deep Learning?
- Why do people pay attention to deep learning today?

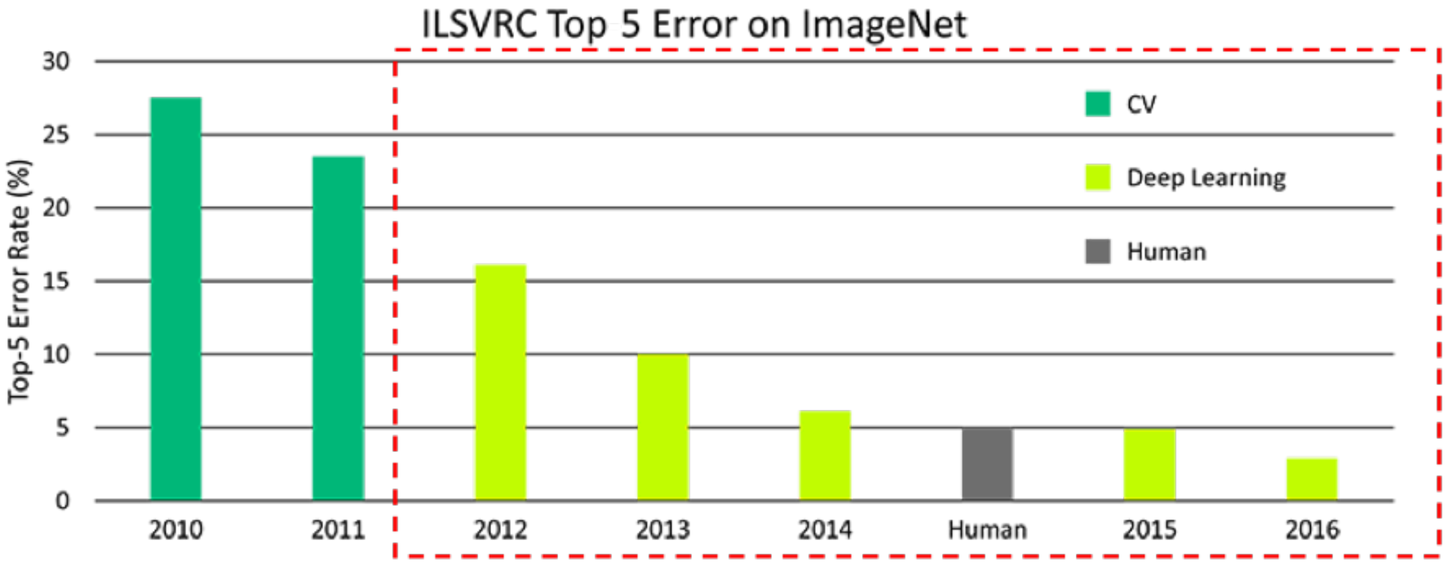


The introduction of Deep Learning techniques drove performance on image categorization from 30% error rates in 2010, down to <2% in 2017

StyleGAN (2019)

The Era of Deep Learning

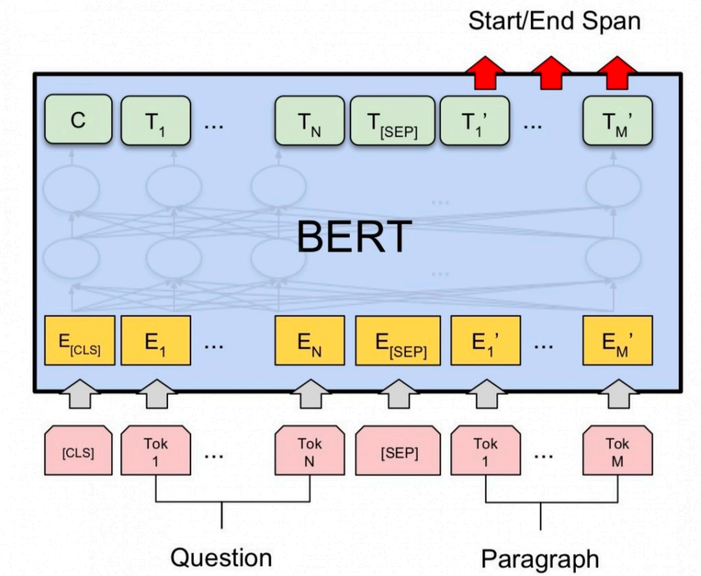
- What is Deep Learning?
- Why do people pay attention to deep learning today?



The introduction of Deep Learning techniques drove performance on image categorization from 30% error rates in 2010, down to <2% in 2017

SQuAD2.0
The Stanford Question Answering Dataset

The Era of Deep Learning



Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 12, 2020	ALBERT + DAAF + Verifier *ensemble*	**90.386**	**92.777**
2 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
3 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
4 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180

The Era of Deep Learning

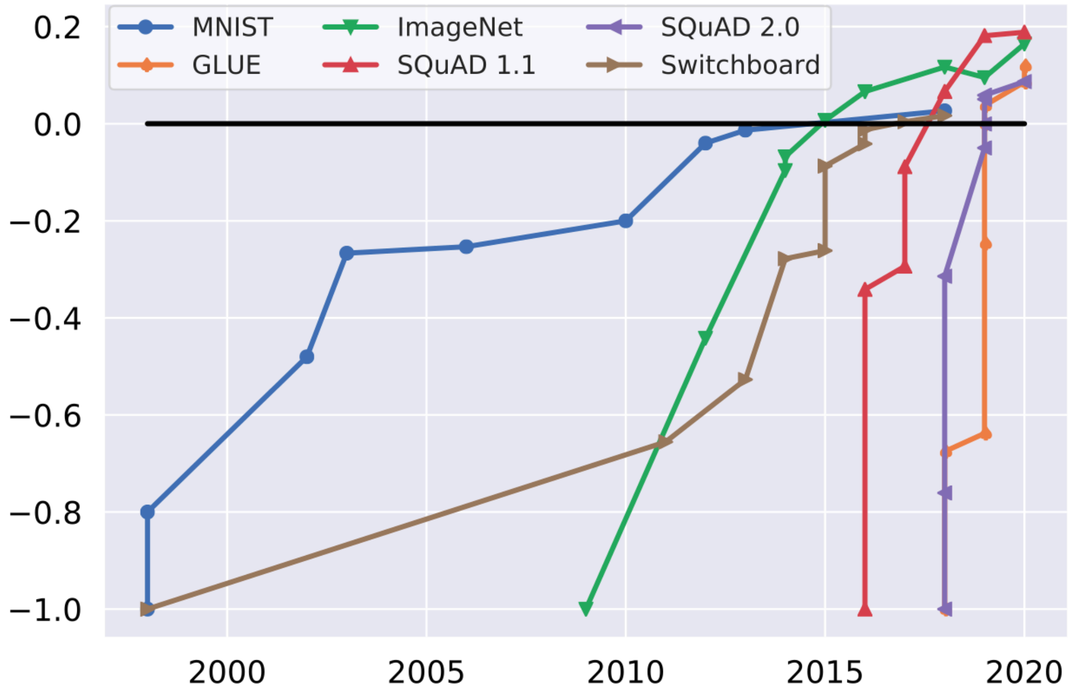
카카오 AI, 한국어 독해 능력 평가서 인간보다 높은 점수로 1위

발행일 : 2019.02.01 16:16

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

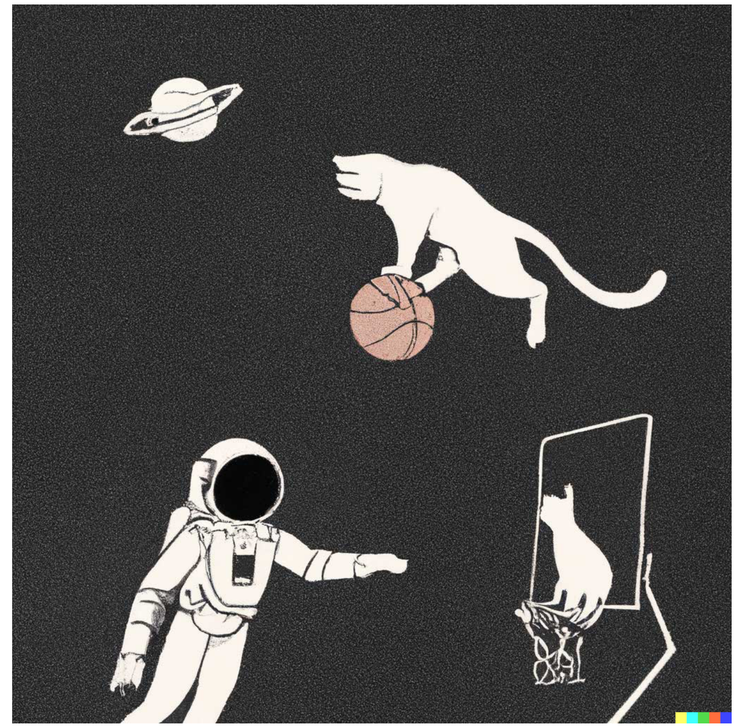
Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean



The Era of Deep Learning

An astronaut riding a horse
in a photorealistic style

An astronaut lounging in a
tropical resort in space as pixel art



An astronaut playing basketball
with cats in space in a minimalist style

Hierarchical Text-Conditional Image Generation with CLIP Latents, Ramesh et al., 2022

The Era of Deep Learning

Fig. 1: We propose a novel text-based editing approach for talking-head video. Given an edited transcript, our approach produces a realistic output video in which the dialogue of the speaker has been modified and the resulting video maintains a seamless audio-visual flow (i.e. no jump cuts).

Text-based Editing of Talking-head Video, Fried et al., ***SIGGRAPH*** 2019

The Era of Deep Learning

medical AI

autonomous driving

protein prediction

What are we going to learn in this class?

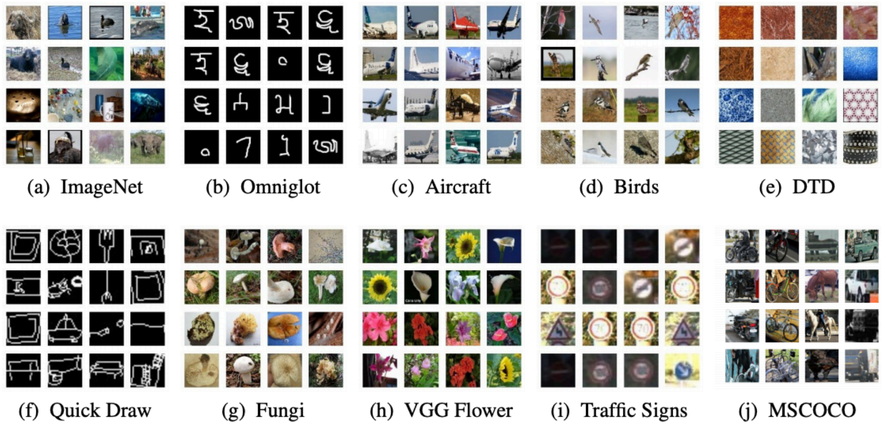
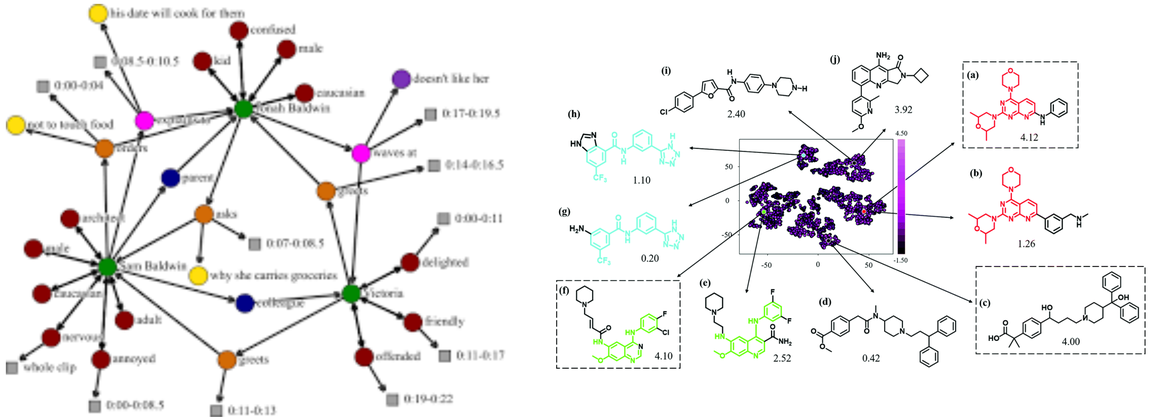
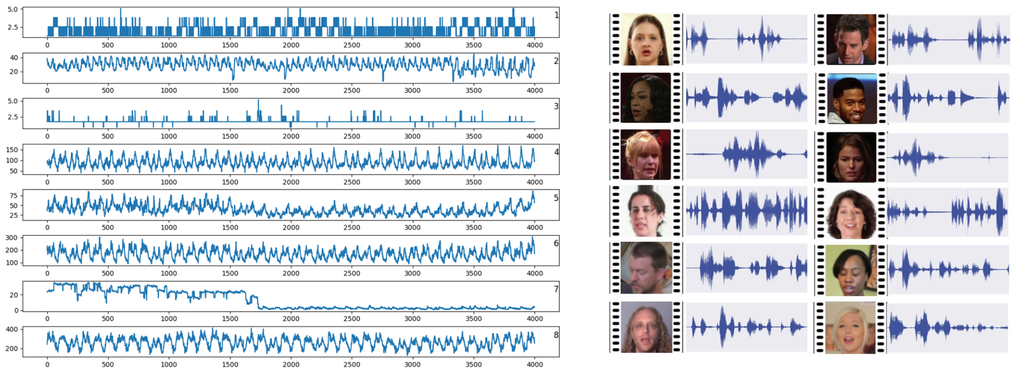


Image Dataset



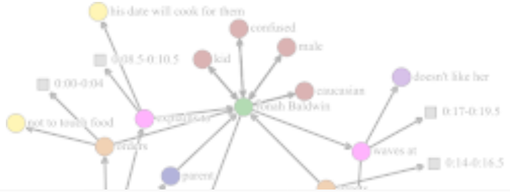
Graph Dataset

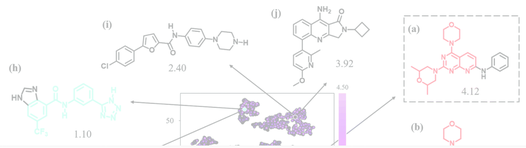
Text Dataset



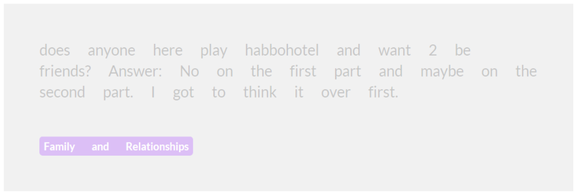
Time Series Dataset

What are we going to learn in this class?

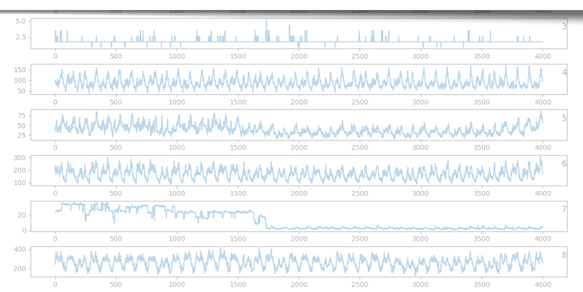




Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	-
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

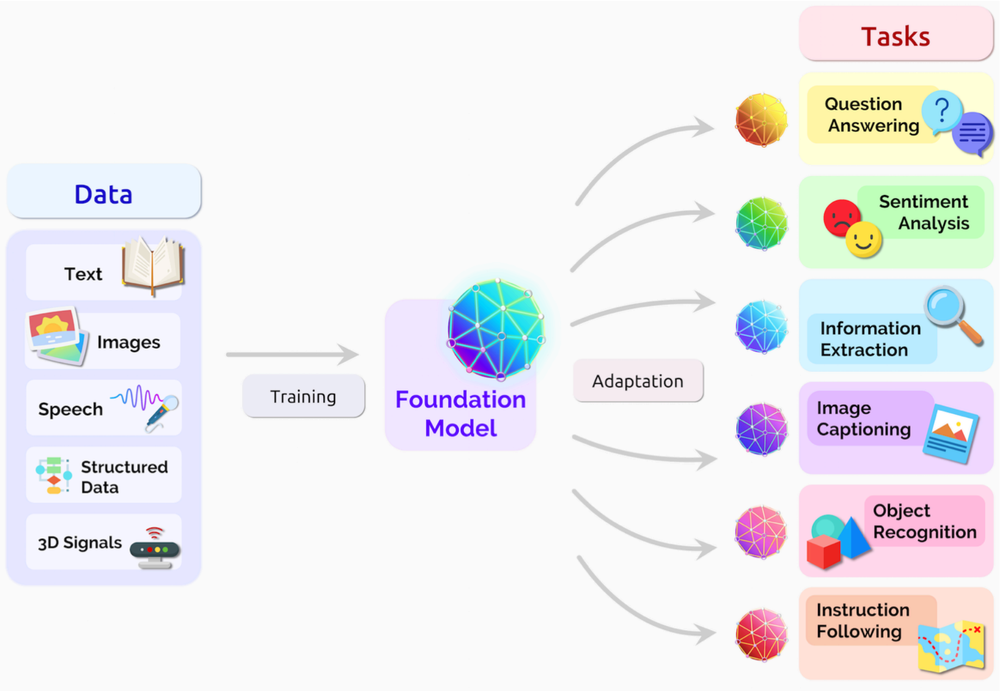


Text Dataset

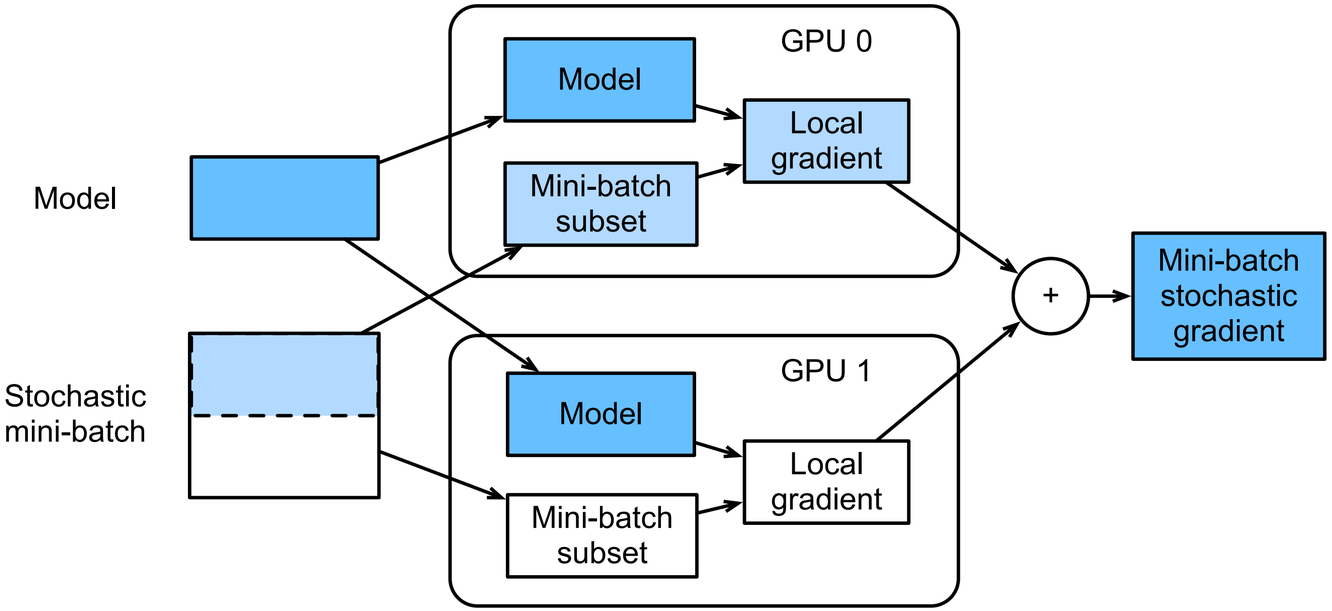


Time Series Dataset

What are we going to learn in this class?



Foundation Model



Multi-GPU Training

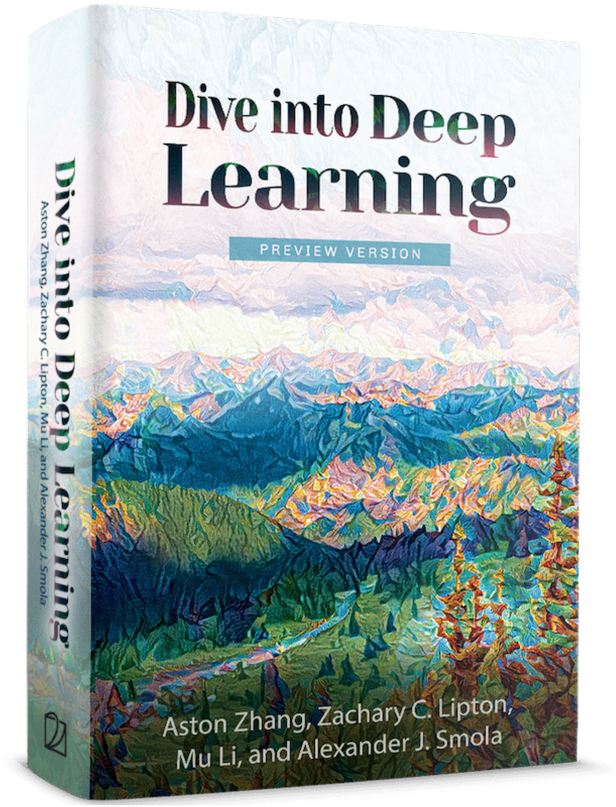
Course Guideline

Contents

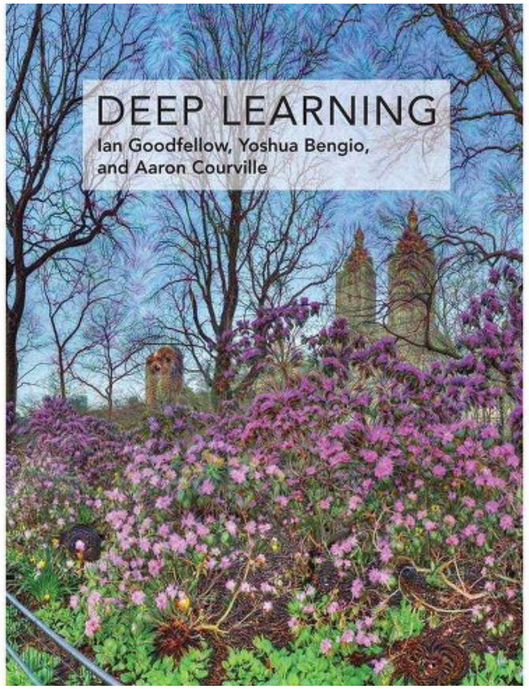
- Multilayer Perceptron (MLP)
- Convolutional Networks (CNN) for Image Data
- Graph Neural Networks (GNN) for Graph Data
- Recurrent Neural Networks (RNN) for Sequential Data
- Attention Modules & Transformer for Text Data
- Self-Supervised Training
- Multi-GPU Training
- Advanced Topics: NeurIPS 2022 paper reading

we will use PyTorch

Textbook?



Dive into Deep Learning



Deep Learning

HF Course

huggingface.co/course

DEEP GRAPH LIBRARY

Open Sources

Grading

- Attendance **(5 points)**
  - COVID-19: report to administration
- Individual Assignments **(70 points)**
  - will be given after ***practice*** class (9/5, 9/19, 9/26, 10/24, 10/31, 11/7, 11/14)
  - submit your code via GitHub classroom
- ***No*** Mid-term Exam
- Final Exam **(25 points)**

Notice

- Students must know the following as preliminaries:
  - AI Programming I : ITP107 → **Python, NumPy**
  - AI Programming II: ITP117 → **Elementary ML, SGD, Colab**
  - Applied Linear Algebra: MTH203 → **Vector, Matrix, Eigenvalue**
- Students must be familiar with Python programming
  - this class is designed for senior/graduate students
  - this class is ***not*** designed for beginners in ML & programming

Three more things…

- Do ***NOT*** distribute the course materials (slides, assignments, tests, etc.)
- No excuse for cheating/plagiarism
  - ***F*** grade, report to administration
  - Enforcement Decree of the Improper Solicitation and Graft Act
    부정청탁 및 금품등 수수의 금지에 관한 법률 (aka 김영란법)
- If you have any question, please contact to **official email**
  - requests via private email will be discarded ai502deeplearning@gmail.com
  - transparency, fairness, and security

GPU Resources?

jupyter

CO

GitHub

Elements of Machine Learning

Principles of Deep Learning (AI502/IE408/IE511)

Sungbin Lim (UNIST AIGS & IE)

What is Machine Learning?

- Mitchell (1997)
  - 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 measure by P, improves with experience E

What is Machine Learning?

- Mitchell (1997)
  - 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 measure by P, improves with experience E
- Task $T \rightarrow$ classification, regression, machine translation, robot control, …

What is Machine Learning?

- Mitchell (1997)
  - 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 measure by P, improves with experience E
- Task $T \rightarrow$ classification, regression, machine translation, robot control, …
- Performance measure $P \rightarrow$ accuracy, error, perplexity, log-likelihood, …

What is Machine Learning?

- Mitchell (1997)
  - 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 measure by P, improves with experience E
- Task $T \rightarrow$ classification, regression, machine translation, robot control, …
- Performance measure $P \rightarrow$ accuracy, error, perplexity, log-likelihood, …
- Experience $E \rightarrow$ dataset, supervised, unsupervised, reinforcement, …

Three ingredients of ML Algorithms

- Data $\mathcal{D} \rightarrow$ image, graph, text, video, …
  - dimension
  - data structure

Three ingredients of ML Algorithms

- Data $\mathcal{D} \rightarrow$ image, graph, text, video, …
  - dimension
  - data structure
- Model $\mathcal{M} \rightarrow$ linear model, SVM, decision tree, neural networks, …
  - input/output
  - function with parameters

Three ingredients of ML Algorithms

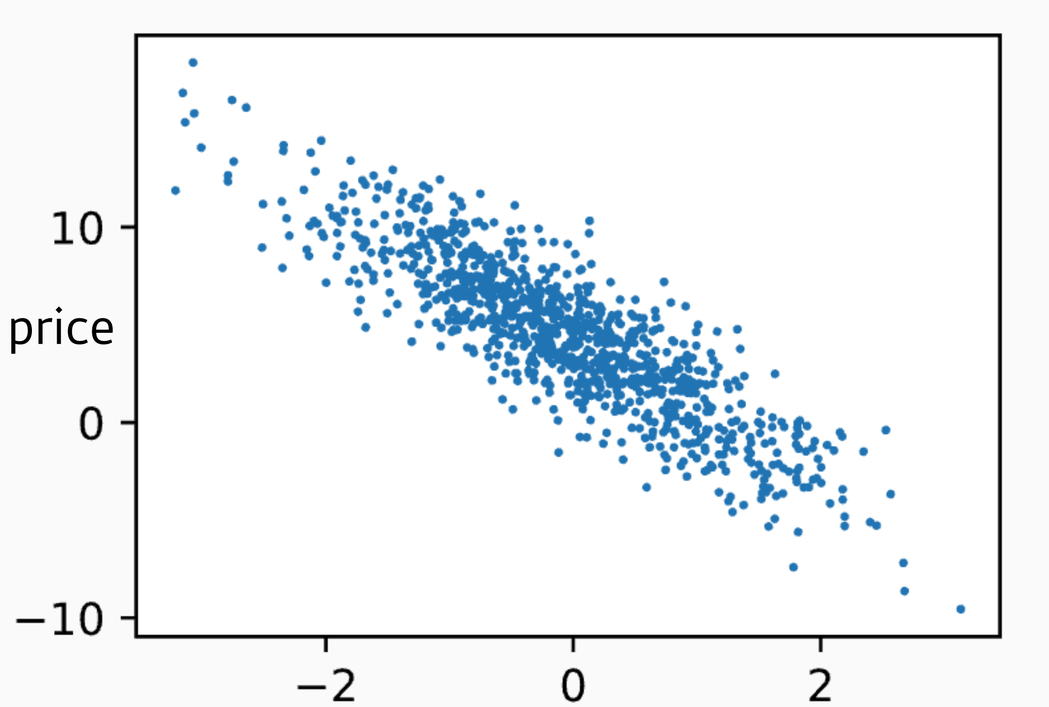
- Data $\mathcal{D} \rightarrow$ image, graph, text, video, …
  - dimension
  - data structure
- Model $\mathcal{M} \rightarrow$ linear model, SVM, decision tree, neural networks, …
  - input/output
  - function with parameters
- Loss function $\mathcal{L} \rightarrow$ L1-loss, L2-loss, cross-entropy, …
  - optimal points must coincide desired performance

Example: Regression

T : regression

P : mean-squared error

E : supervised learning

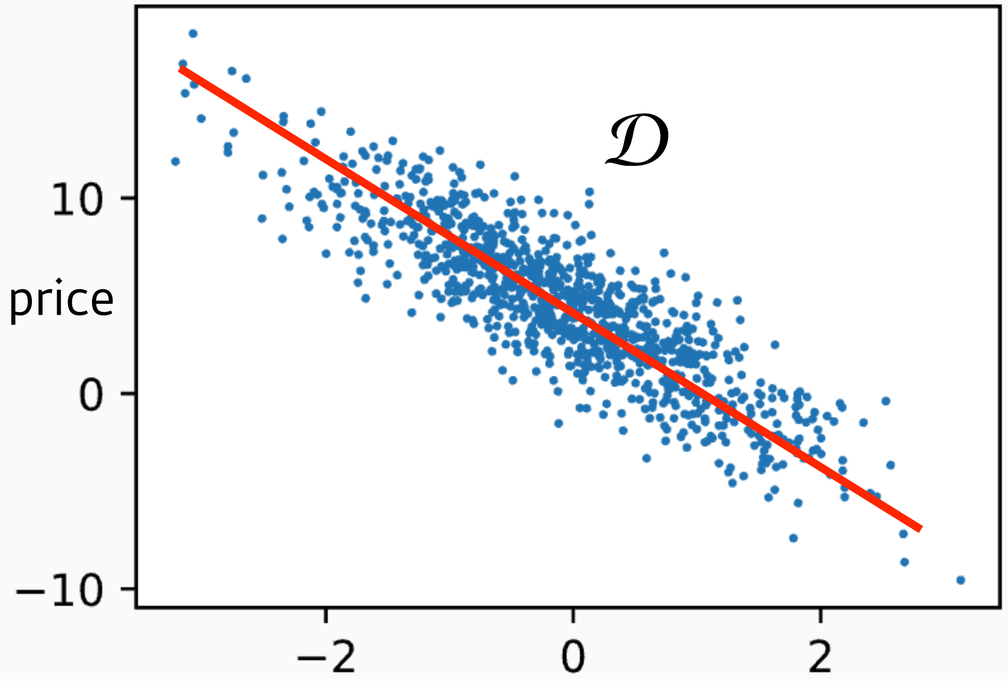


Example: Regression

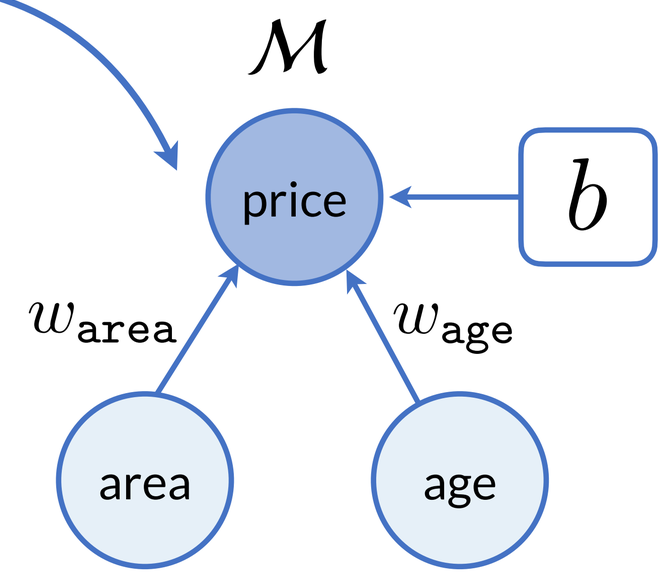
T: regression
P: mean-squared error
E: supervised learning

$\mathcal{L}$ MSE Loss $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}|y_i^{(d)} - \hat{y}_i^{(d)}|^2$

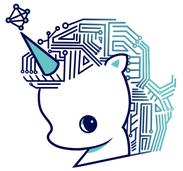
$$\text{price} = w_{\text{area}} \cdot \text{area} + w_{\text{age}} \cdot \text{age} + b.$$



graphical model



Example: Image Classification



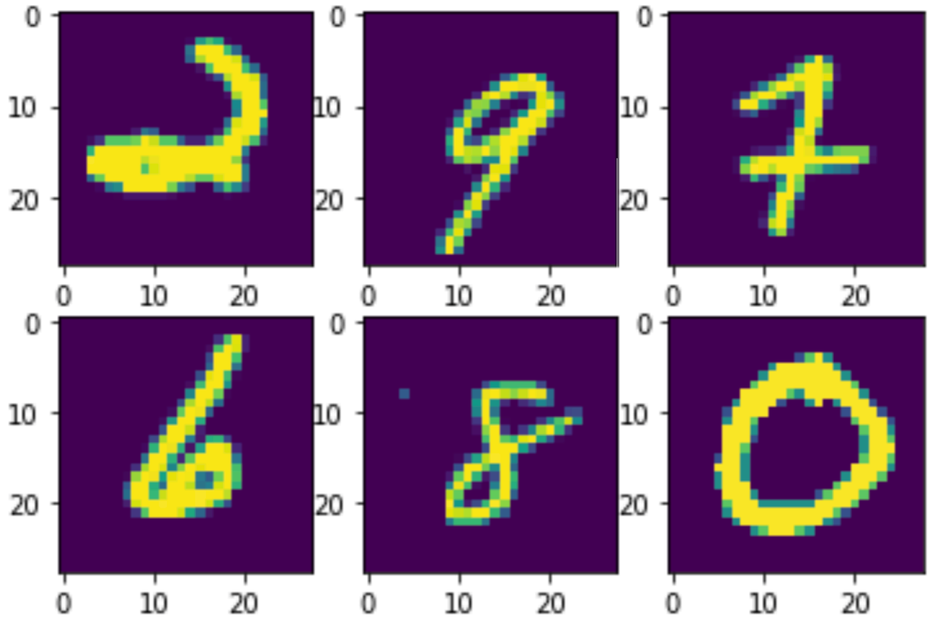
we want to minimize the information gap

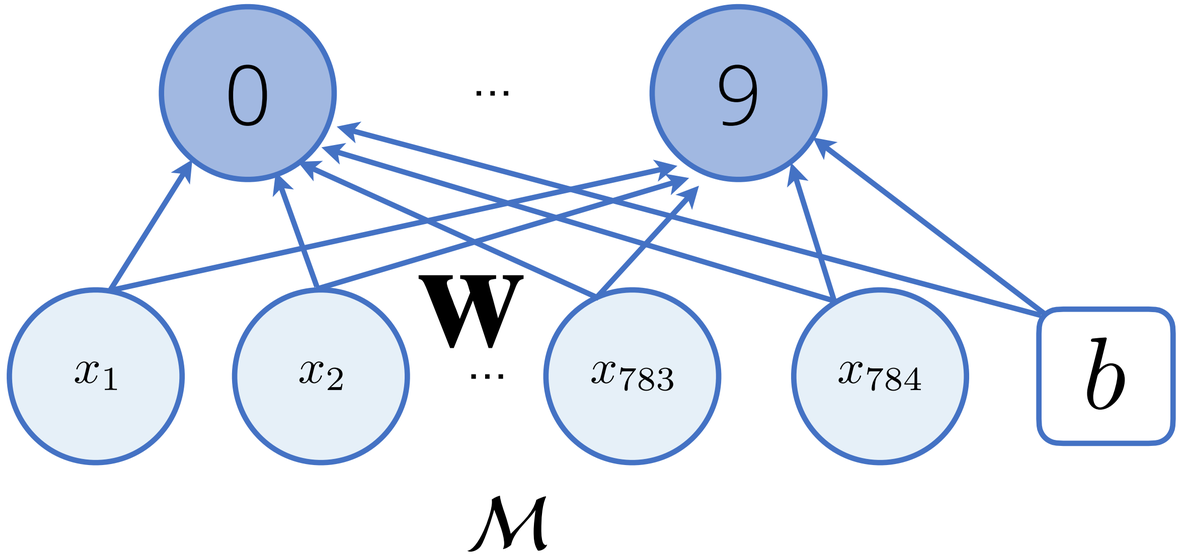
T : classification
P : accuracy
E : supervised learning

$\mathcal{L}$ Cross-Entropy Loss $\frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{y}_i, \log \hat{\mathbf{y}}_i \rangle = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} y_i^{(d)} \log \hat{y}_i^{(d)}$

```
torch.Size([6, 1, 28, 28])
tensor([2, 9, 7, 6, 8, 0])
```

$\mathcal{D}$



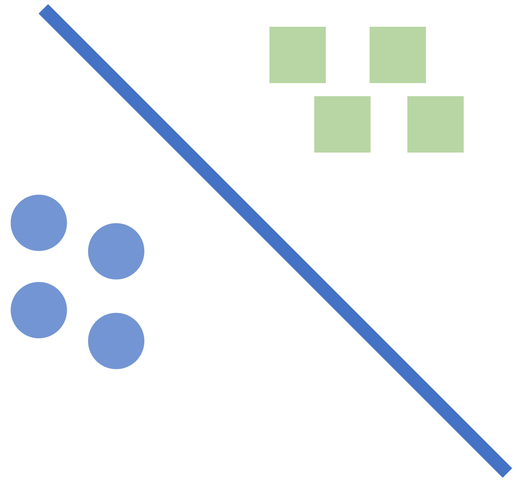


$\mathcal{M}$

Beyond Linear Function

not linear models!

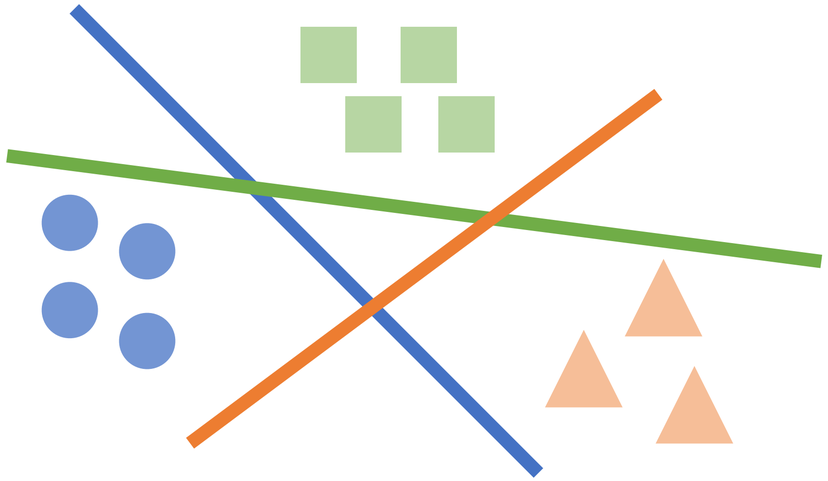
- Let’s consider a binary classification problem
- According to the learning theory, classification is equivalent to the finding a hyperplane between two classes



Beyond Linear Function

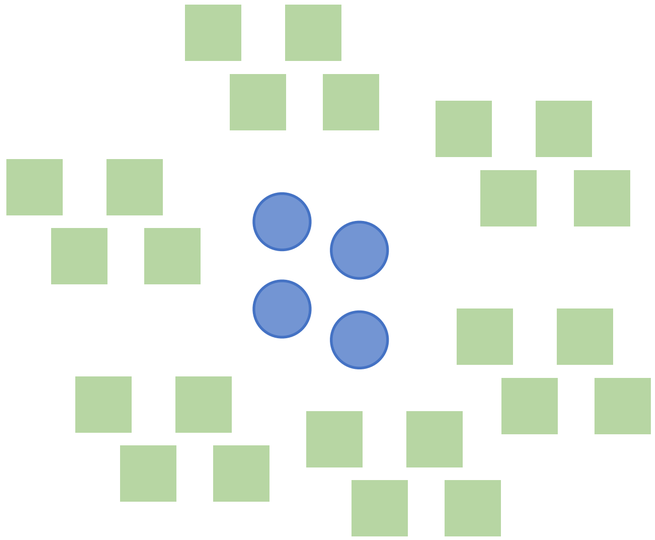
not linear models!

- Let’s consider a binary classification problem
- According to the learning theory, classification is equivalent to the finding a hyperplane between two classes
- One-hot vector represents the output of multi-dimensional binary classification



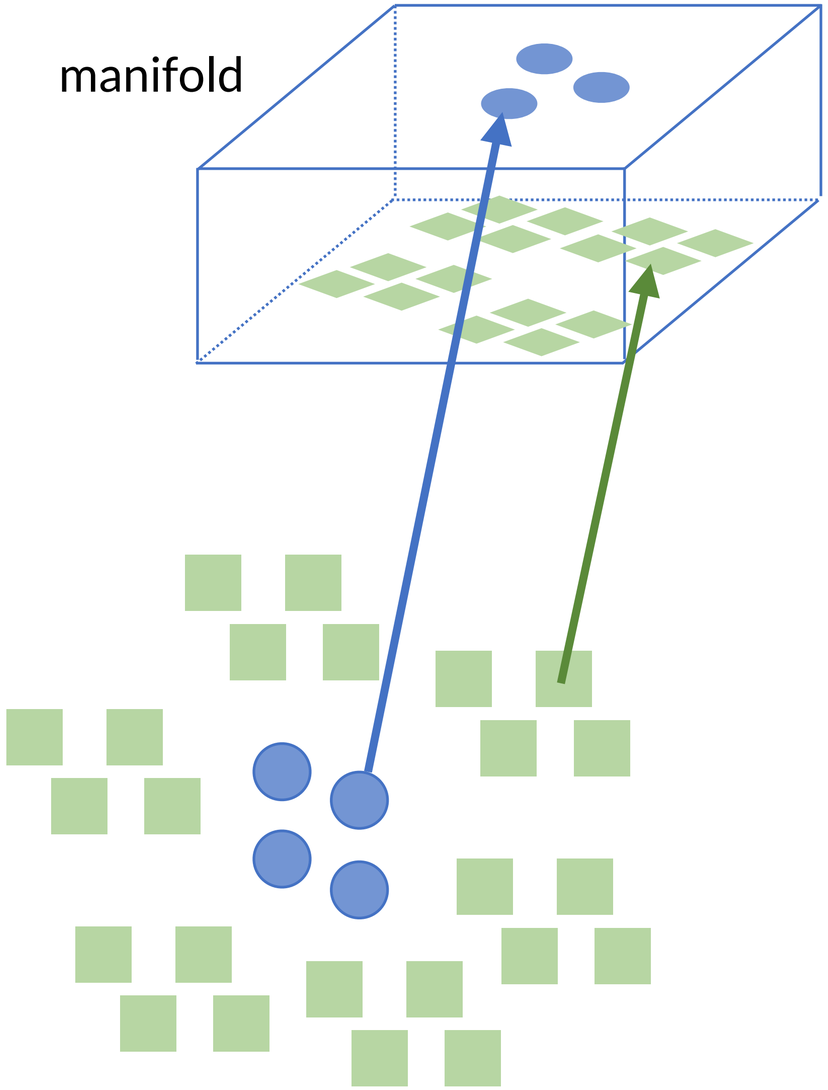
Beyond Linear Function

- If the decision boundary is nonlinear, then we have to find a continuous mapping which transforms the features from non-separable space to a separable one.
- But linear function cannot do this!
  - linear transform = affine transform
- Then how?



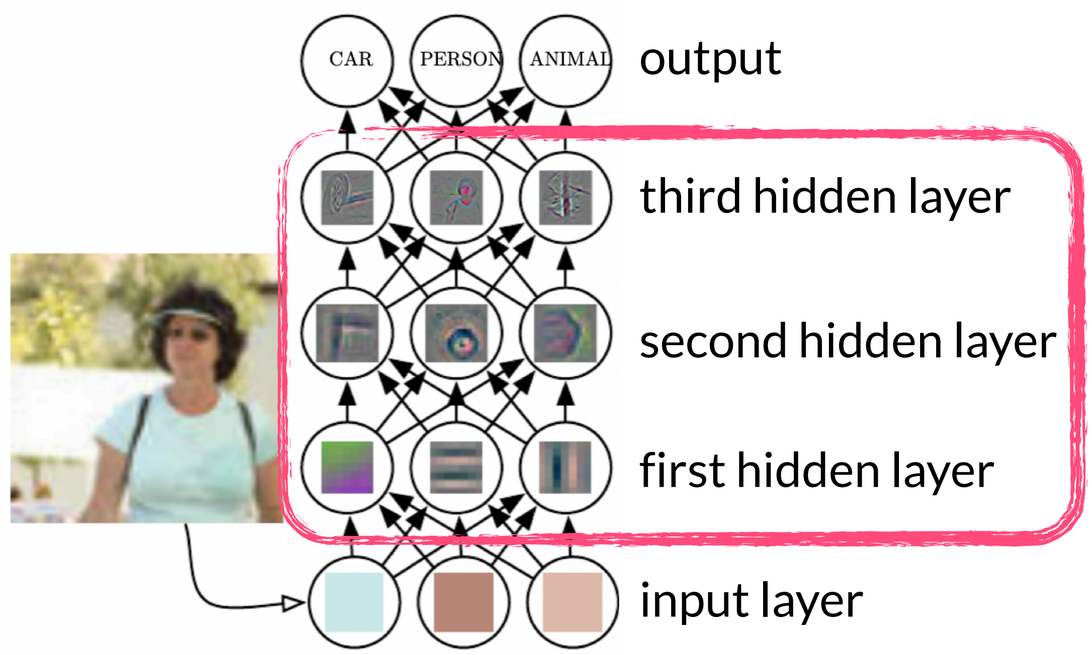
Beyond Linear Function

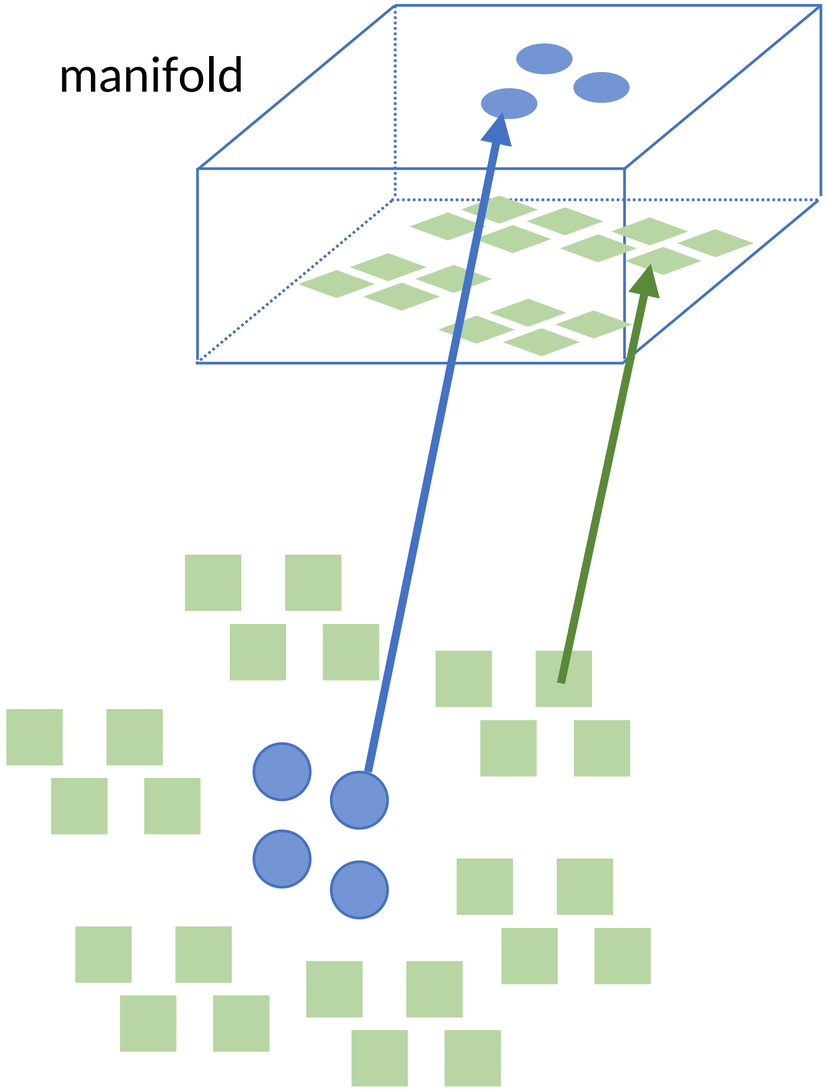
- We find a manifold in some high-dimensional spaces by learning a smooth representation
- This is possible if ***manifold hypothesis*** holds



Beyond Linear Function

- We find a manifold in some high-dimensional spaces by learning a smooth representation
- This is possible if ***manifold hypothesis*** holds





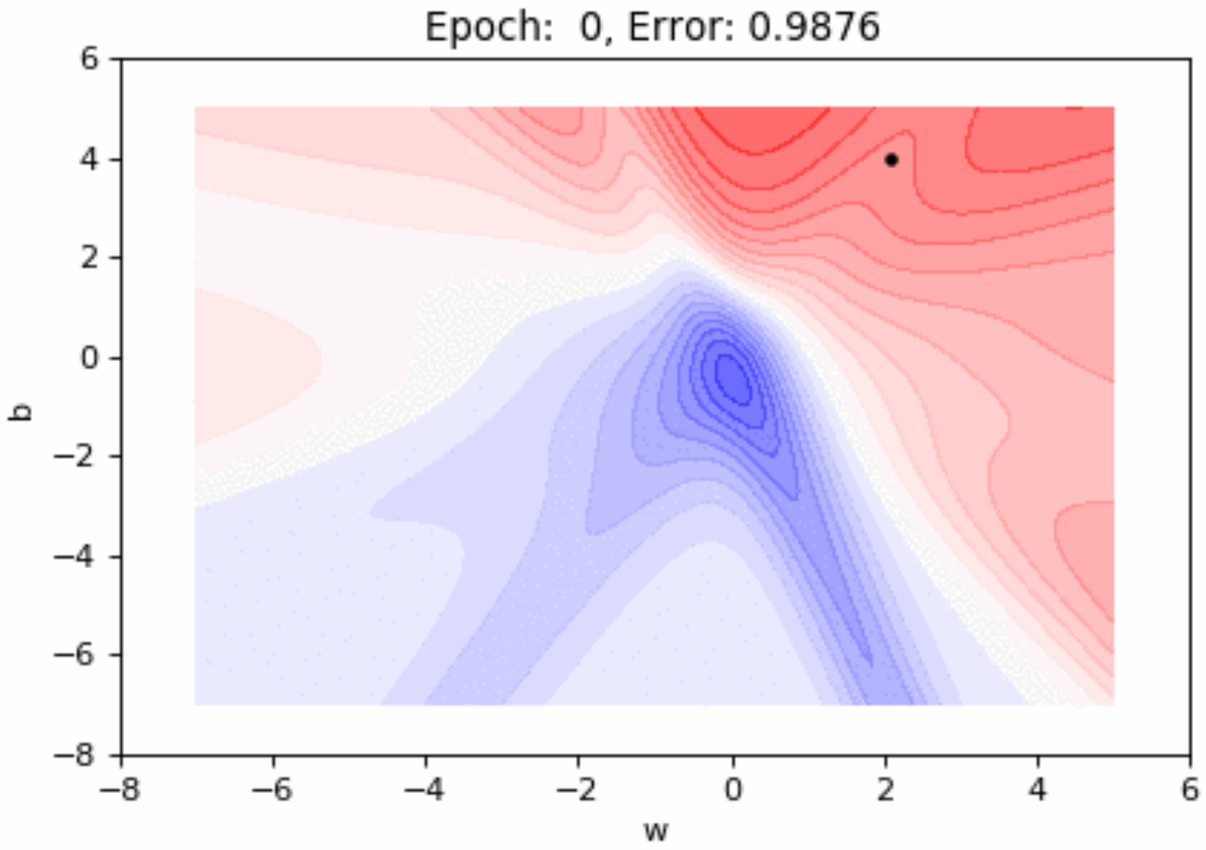
SGD Optimization

hence we need to consider differentiable loss functions

iterative update

$$(\mathbf{w}, b) \leftarrow (\mathbf{w}, b) - \frac{\eta}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \partial_{(\mathbf{w},b)} l^{(i)}(\mathbf{w}, b).$$

partial derivative



SGD Optimization

hence we need to consider differentiable loss functions

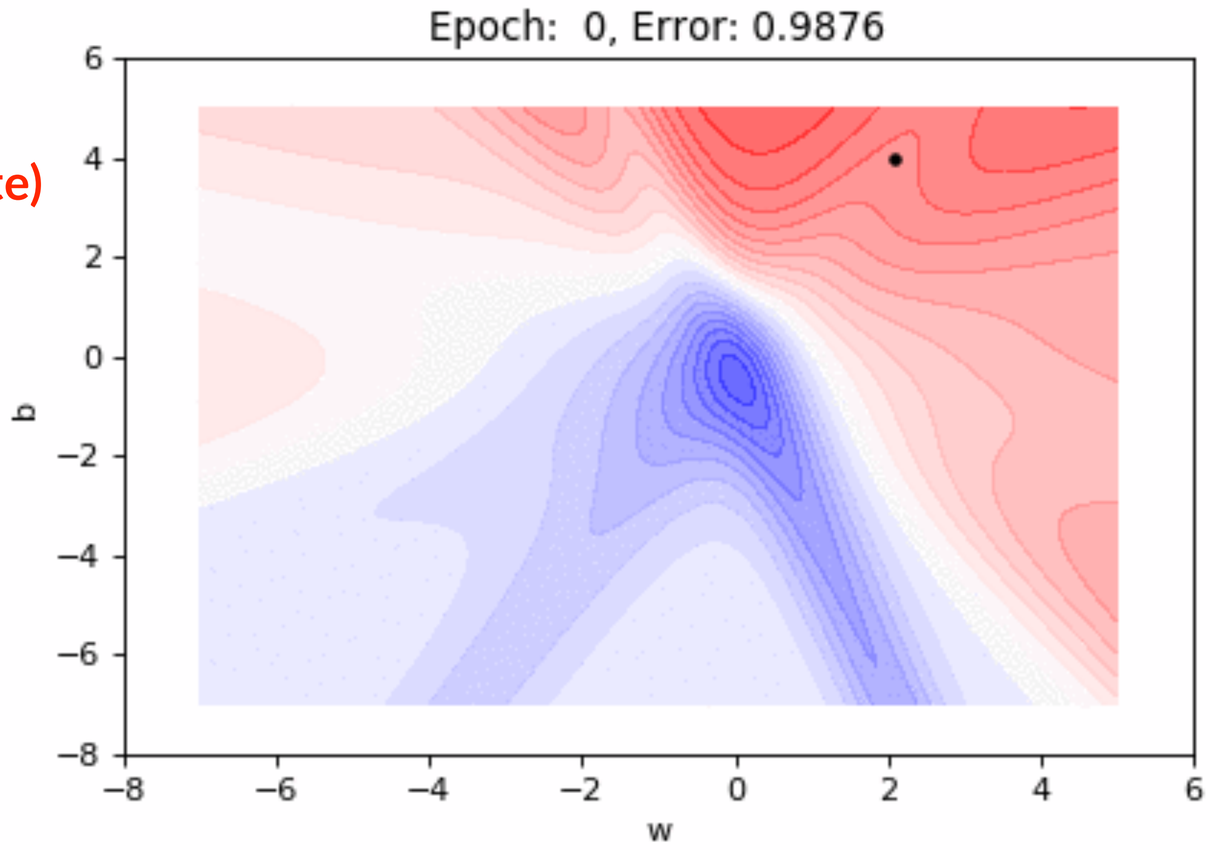
iterative update

step size (learning rate)

$$(\mathbf{w}, b) \leftarrow (\mathbf{w}, b) - \frac{\eta}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \partial_{(\mathbf{w},b)} l^{(i)}(\mathbf{w}, b).$$

number of batches

partial derivative



Fundamental Problem of ML

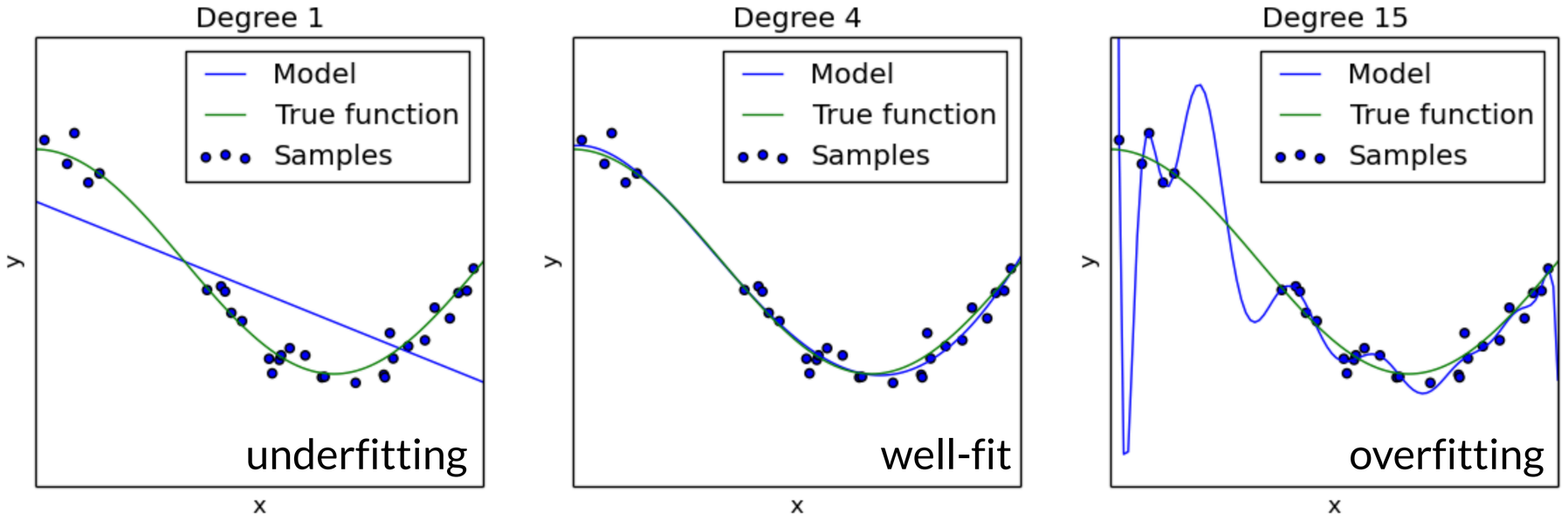
- Overfitting vs Underfitting
  - make the training error small
  - make the gap between training and test error small

optimization

$$\min_{\mathbf{w}} \mathcal{L}(\mathcal{M}(\mathbf{w}), \mathcal{D}_{\text{train}})$$

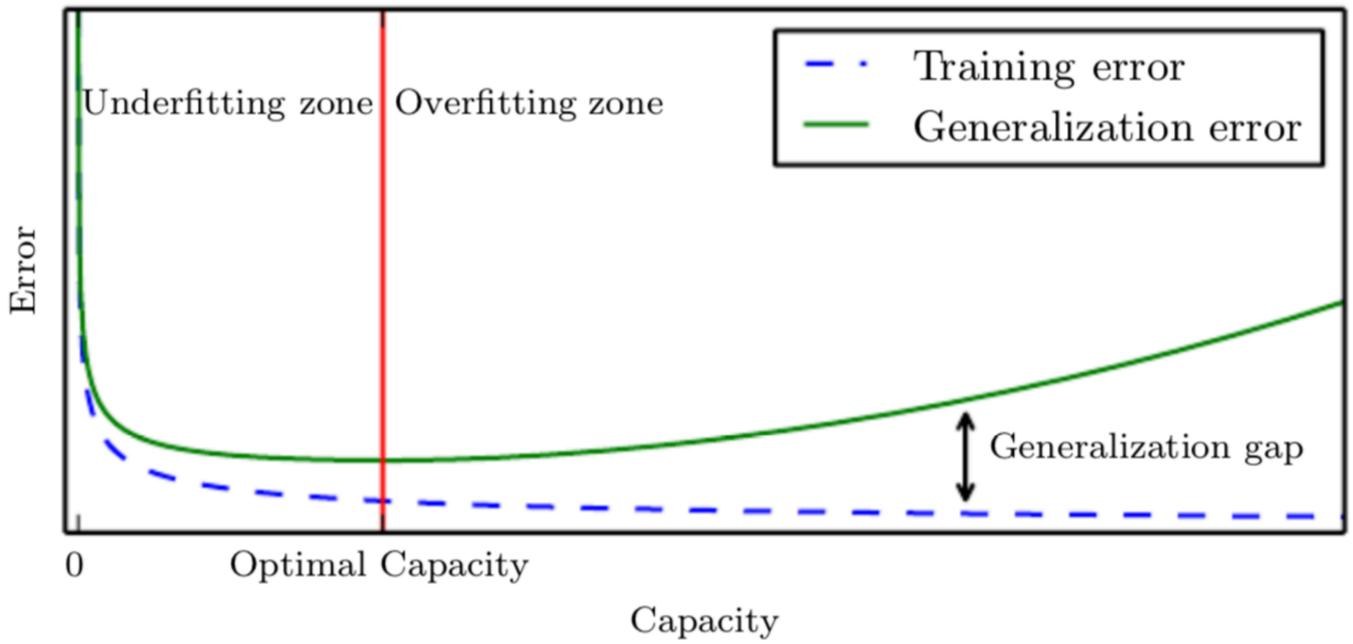
capacity

$$\min_{\mathbf{w}} \mathcal{L}(\mathcal{M}(\mathbf{w}), \mathcal{D}_{\text{test}})$$



Generalization Performance

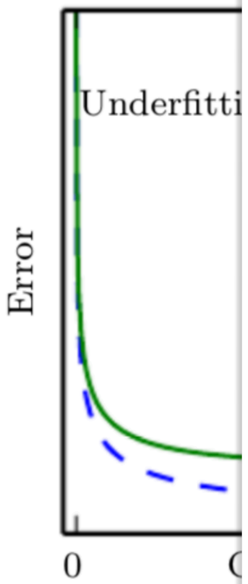
- Generalization gap: test error - training error
- We need to determine optimal capacity via generalization gap

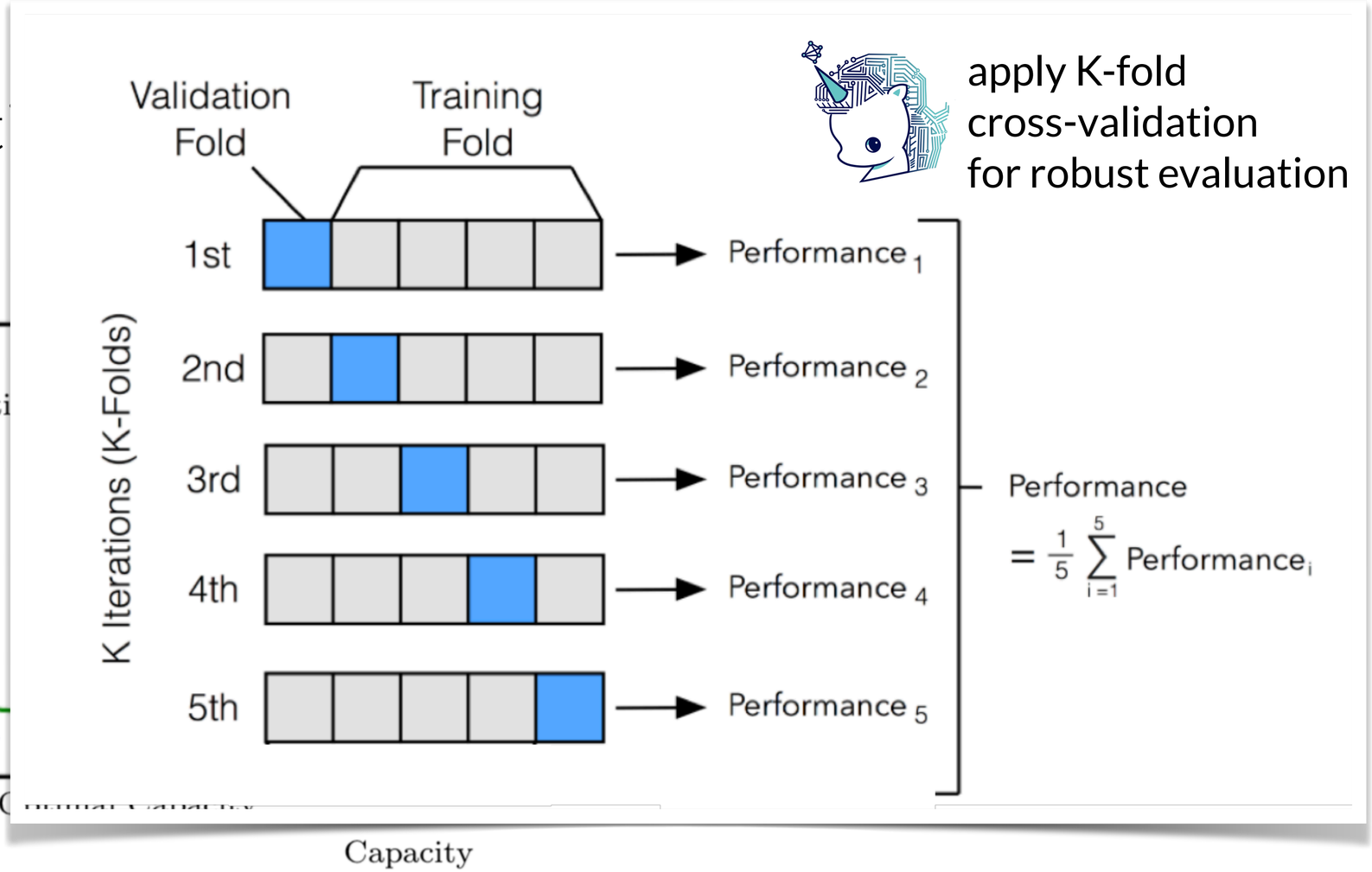


but this control does not help reducing overfitting from mismatch between train & test (data generating process)

Generalization Performance

- Generalizat
- We need to

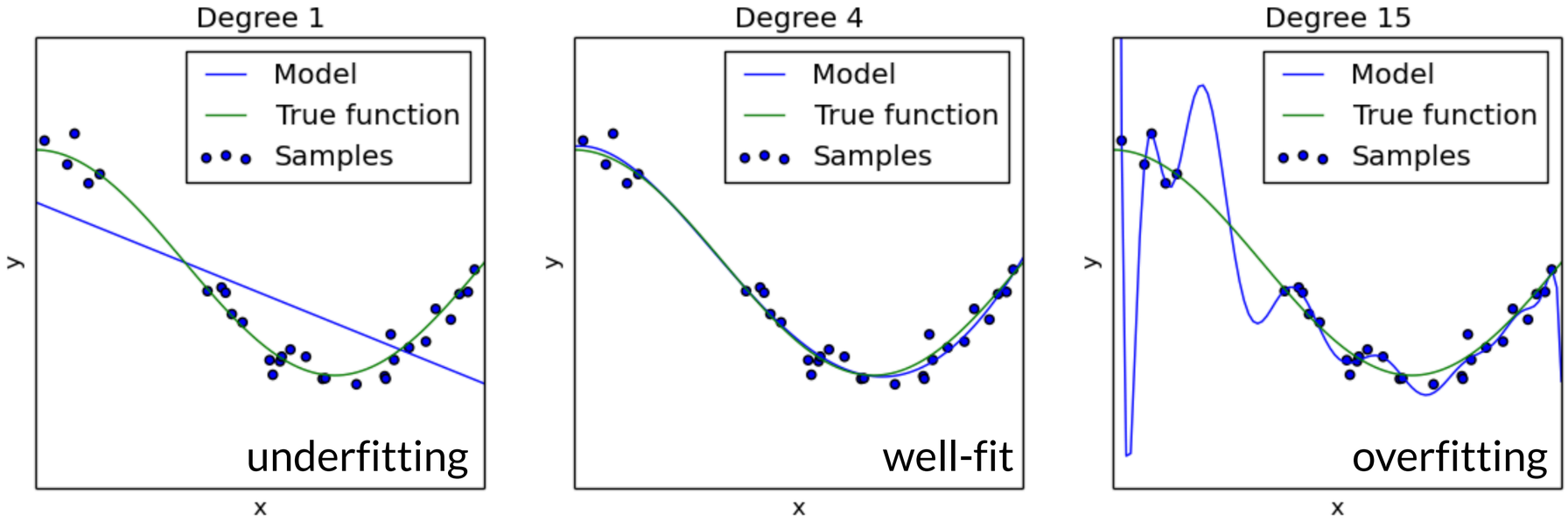




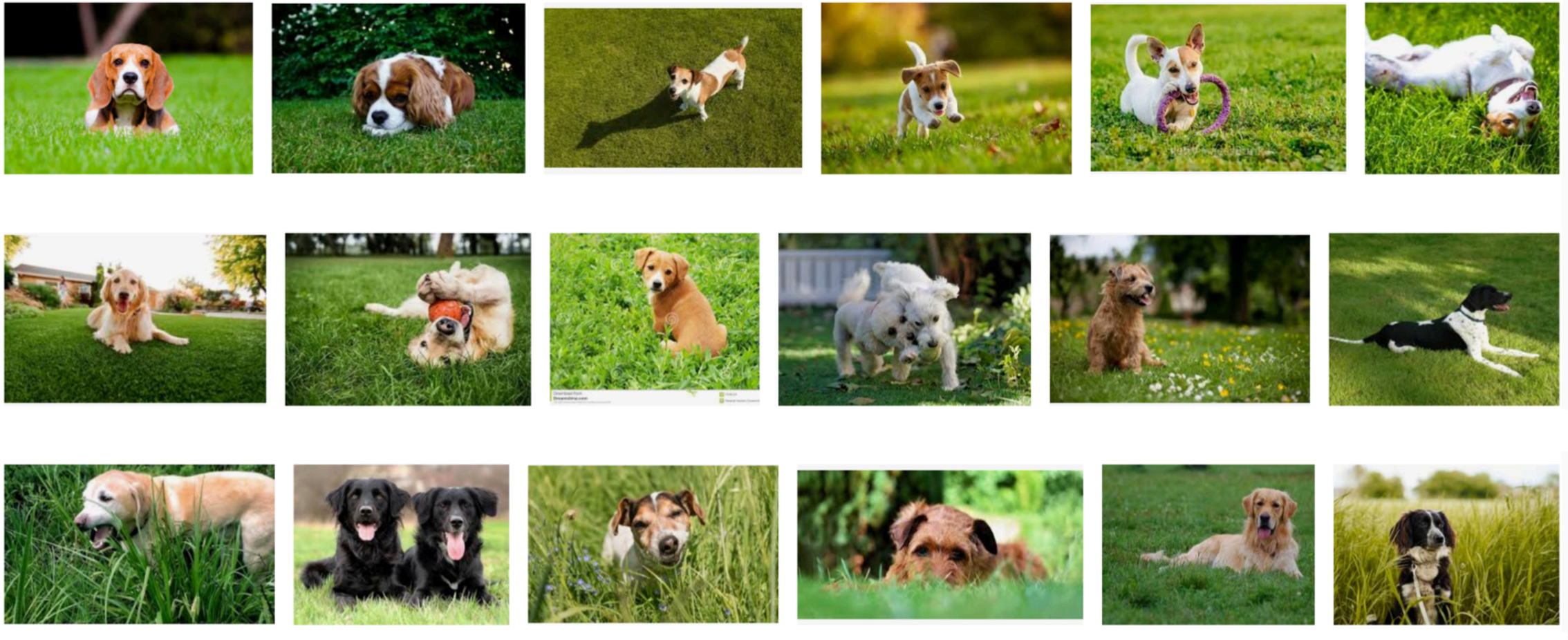
apply K-fold
cross-validation
for robust evaluation

Fundamental Problem of ML

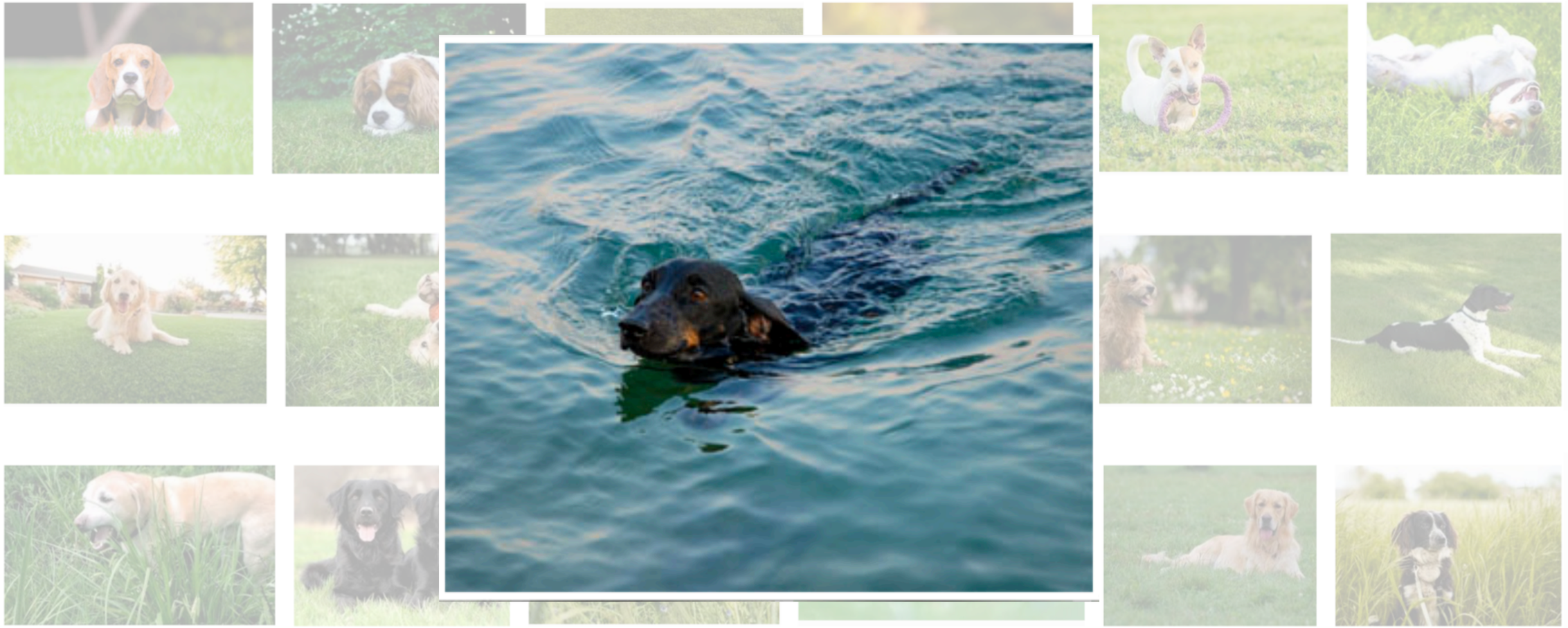
- Overfitting vs Underfitting
  - make the training error small → optimization $\min_{\mathbf{w}} \mathcal{L}(\mathcal{M}(\mathbf{w}), \mathcal{D}_{\text{train}})$
  - make the gap between training and test error small → capacity $\min_{\mathbf{w}} \mathcal{L}(\mathcal{M}(\mathbf{w}), \mathcal{D}_{\text{test}})$ → $p_{\text{train}} \approx p_{\text{test}}$ data generating process



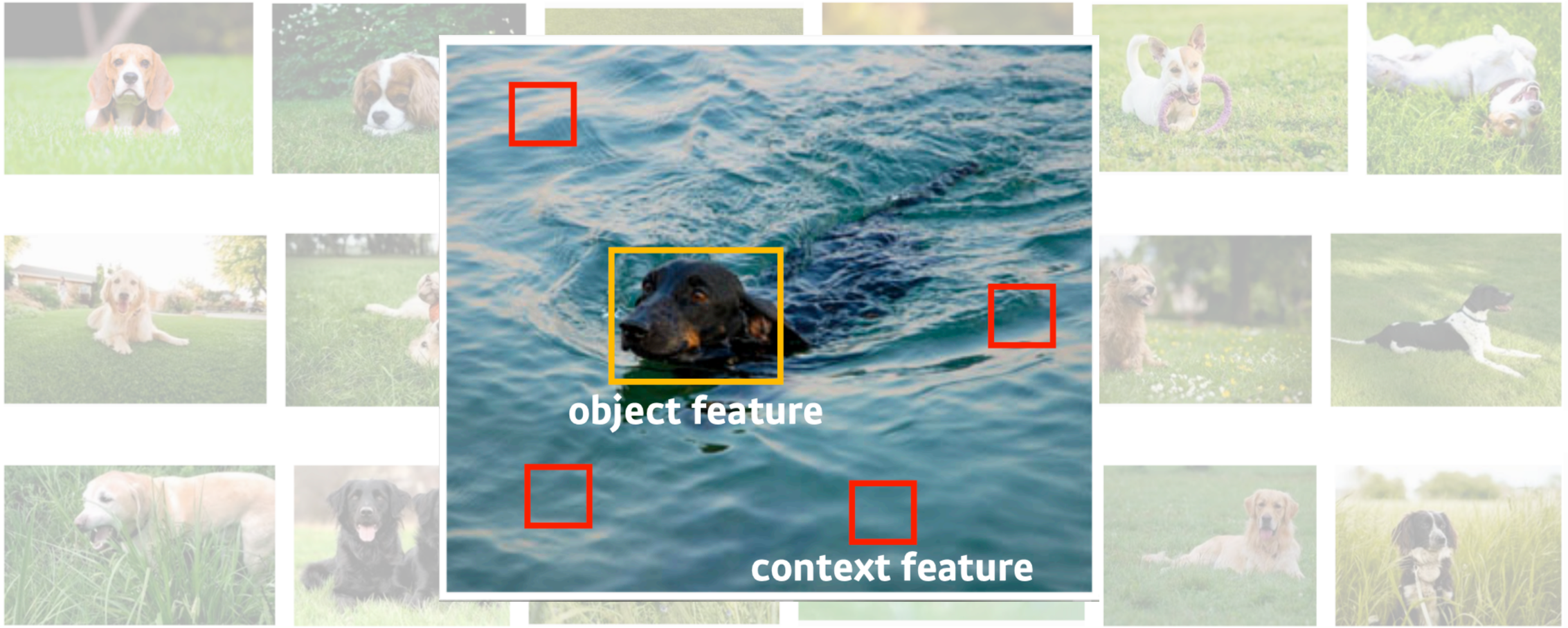
Data Mismatch in ML



Data Mismatch in ML



Data Mismatch in ML



Q&A