Self-Supervised Learning (SSL)
AI 835
Aug-Dec 2023

Introduction
3rd Aug 2023

Dr. V. Ramasubramanian
Professor
International Institute of Information Technology - Bangalore (IIIT-B)
Bangalore, India

Introduction

Module – 1

Basic learning paradigms (supervised, unsupervised, semi-supervised, self-supervised) definitions and settings.

Origins of Self-Supervised Learning (SSL) – early unsupervised learning paradigms (RBM, AE, Word2vec, AR, etc.).

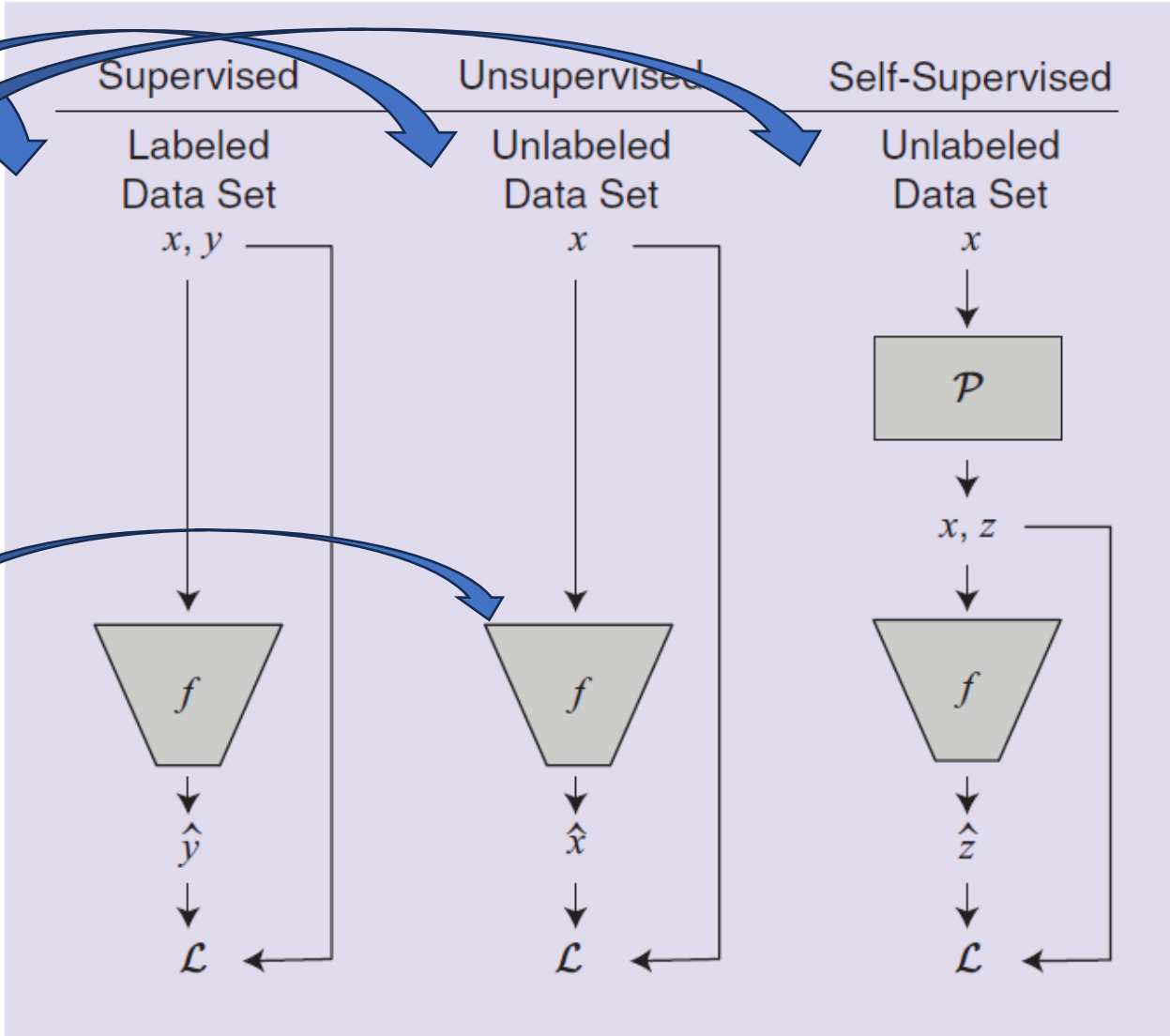


FIGURE 1. Contrasting supervised, unsupervised and self-supervised learning paradigms for training a model f using raw data x, labels y, and loss function $\mathcal{L}$. Self-supervision methods introduce pretext tasks $\mathcal{P}$ that generate pseudolabels z for discriminative training of f.

Module – 2

Basics of Foundation Models, SSL formalisms, definitions and examples of

- **Pretext tasks**
- **Losses**
- **Downstream adaptations**

in a domain-agnostic setting

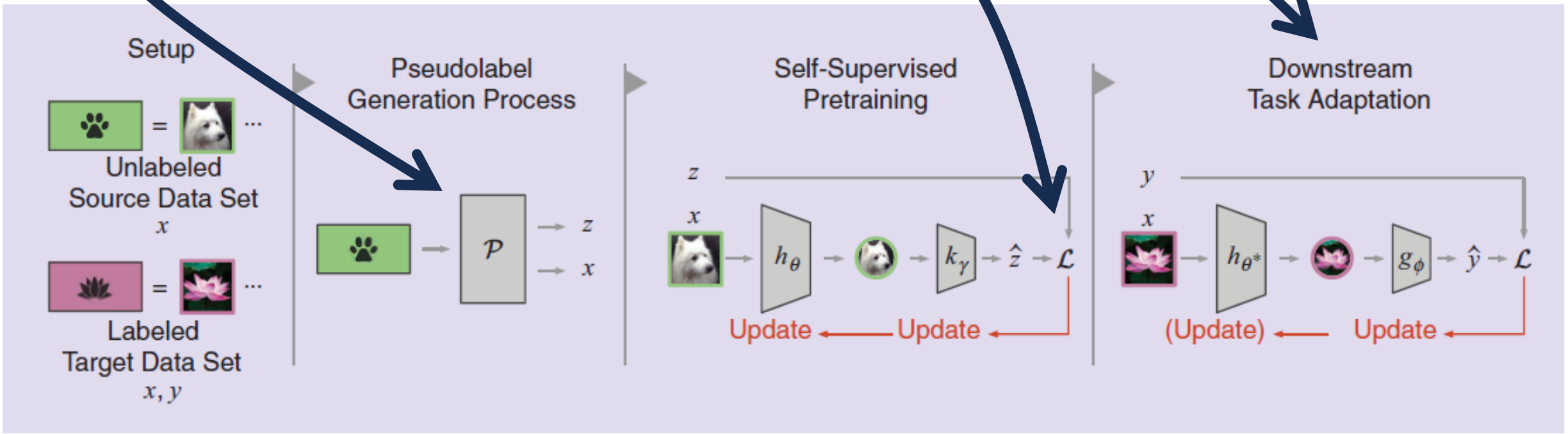


FIGURE 2. The self-supervised workflow starts with an unlabeled source data set and a labeled target data set. As defined by the pretext task, pseudolabels are programmatically generated from the unlabeled set. The resulting inputs, x and pseudolabels z, are used to pretrain the model $k_\gamma(h_\theta(\cdot))$—composed of feature extractor h_θ and output k_γ modules—to solve the pretext task. After pretraining is complete, the learned weights θ^* of the feature extractor h_{θ^*} are transferred and used together with a new output module g_ϕ to solve the downstream target task.

Yann LeCun's Cake Analogy

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Credit by Yann LeCun

Unsupervised Learning vs. Self-supervised Learning

> *I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term. …*
>
> *Self-supervised learning uses way more supervisory signals than supervised learning, and enormously more than reinforcement learning. That's why calling it "unsupervised" is totally misleading.*
>
> by Yann LeCun (2019. 04. 30)

- How to define "unsupervised learning" term? (there is no answer …)
  - **Q)** We need an objective (or loss) for learning; is the objective not a (self-)supervision?
  - **Q)** Unsupervised learning ⊇ self-supervised learning?
  - **Q)** What are the purely unsupervised learning methods?
    - In classic ML, clustering, grouping and dimensionality reduction …

- *Supervised learning* – learning with labeled data
  - Approach: collect a large dataset, manually label the data, train a model, deploy
  - It is the dominant form of ML at present
  - Learned feature representations on large datasets are often transferred via pre-trained models to smaller domain-specific datasets
- *Unsupervised learning* – learning with unlabeled data
  - Approach: discover patterns in data either via clustering similar instances, or density estimation, or dimensionality reduction …
- *Self-supervised learning* – representation learning with unlabeled data
  - Learn useful feature representations from unlabeled data through pretext tasks
  - The term "self-supervised" refers to creating its own supervision (i.e., without supervision, without labels)
  - Self-supervised learning is one category of unsupervised learning

A small note on terminology...

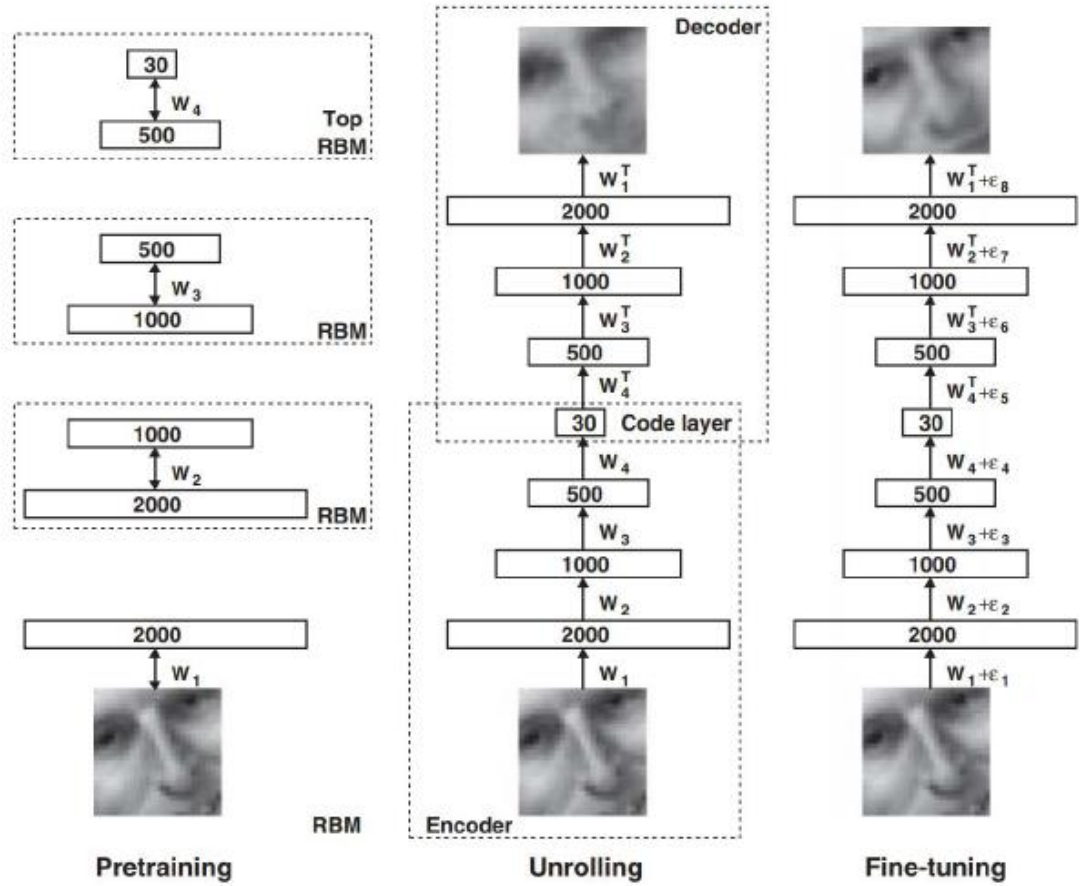
- The distinction between **unsupervised** versus **self-supervised** learning can be blurry sometimes
- Roughly it is this:
  - **Unsupervised learning** attempts to learn representations without labels by *not using any targets of any sort during training*, e.g. by using correlations in activity between units
  - **Self-supervised learning** attempts to learn representations without labels by *using the data itself to generate targets*, e.g. generating targets using the next word in a sentence
  - Put another way, self-supervised learning looks a lot like supervised learning in code, but there is a big difference related to the following question: *do you as a machine learning researcher have to actually ask someone to label the data or not?*

SSL Opened Deep Learning

Science, 2006

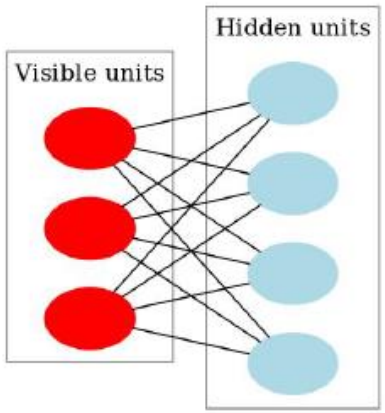
Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

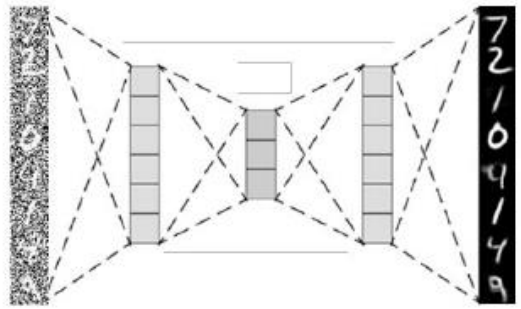


Early Work: Connecting the Dots

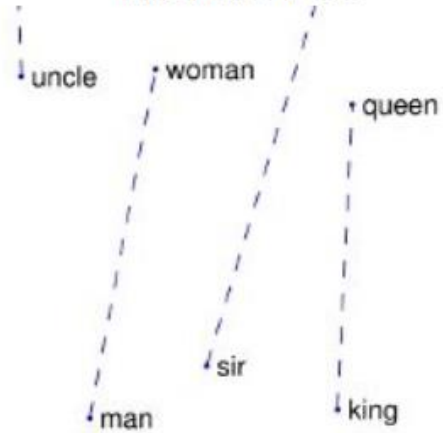
Restricted Boltzmann Machines



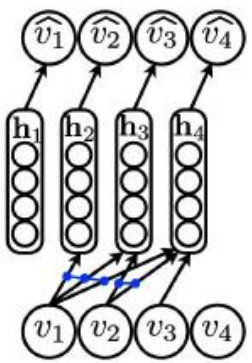
Autoencoders



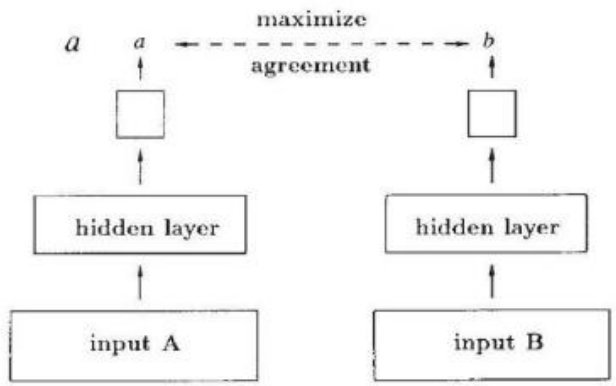
Word2Vec



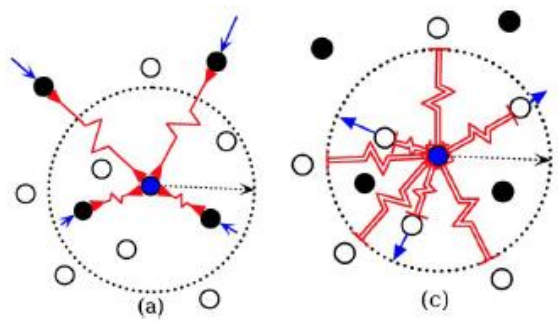
Autoregressive Modeling



Siamese networks



Multiple Instance/Metric Learning



Why Self-Supervision?

1. Expense of producing a new dataset for each new task

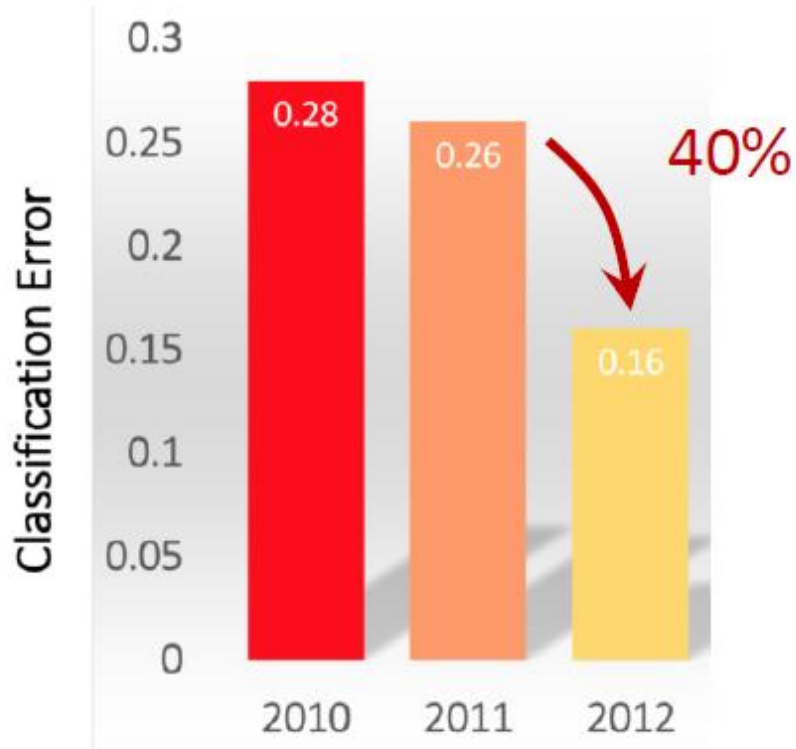
2. Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation

3. Untapped/availability of vast numbers of unlabelled images/videos
   - Facebook: one billion images uploaded per day
   - 300 hours of video are uploaded to YouTube every minute

Burst of Deep Learning in Computer Vision

- Supervised learning using AlexNet (NeurIPS'2012)

ImageNet Challenge



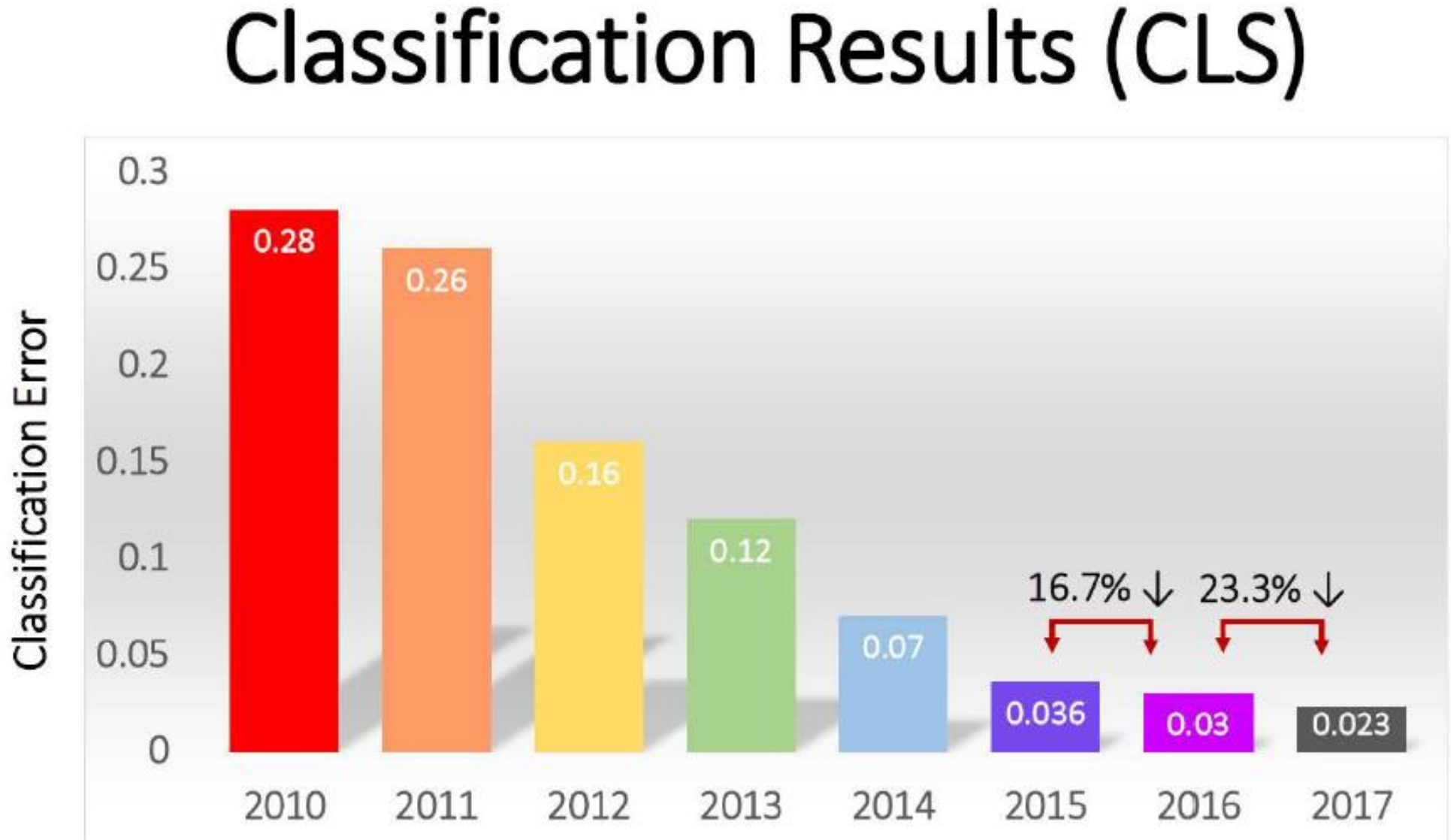
The ImageNet Challenge Story …

IMAGENET

1000 categories

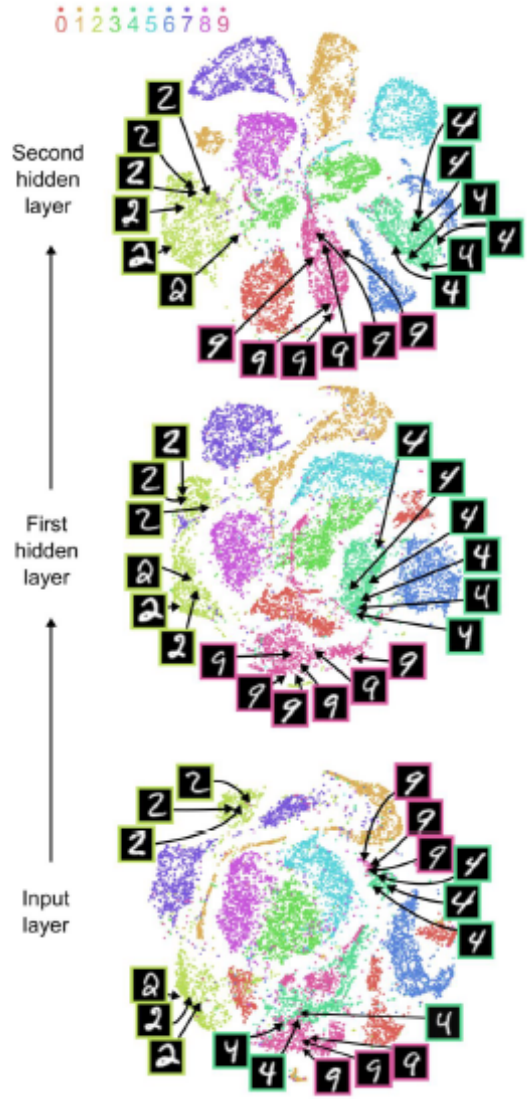
- Training: 1000 images for each category
- Testing: 100k images

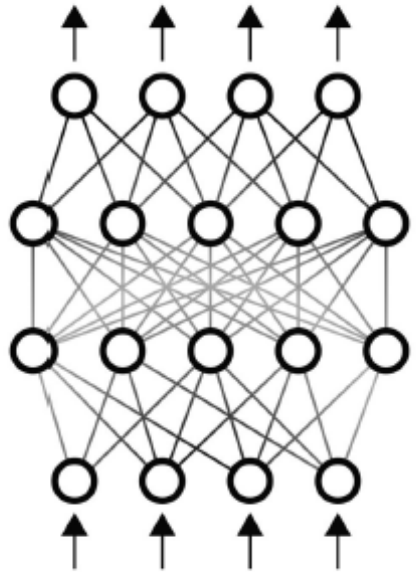
The ImageNet Challenge Story … strong supervision



Supervised deep learning induces progressively more useful representations

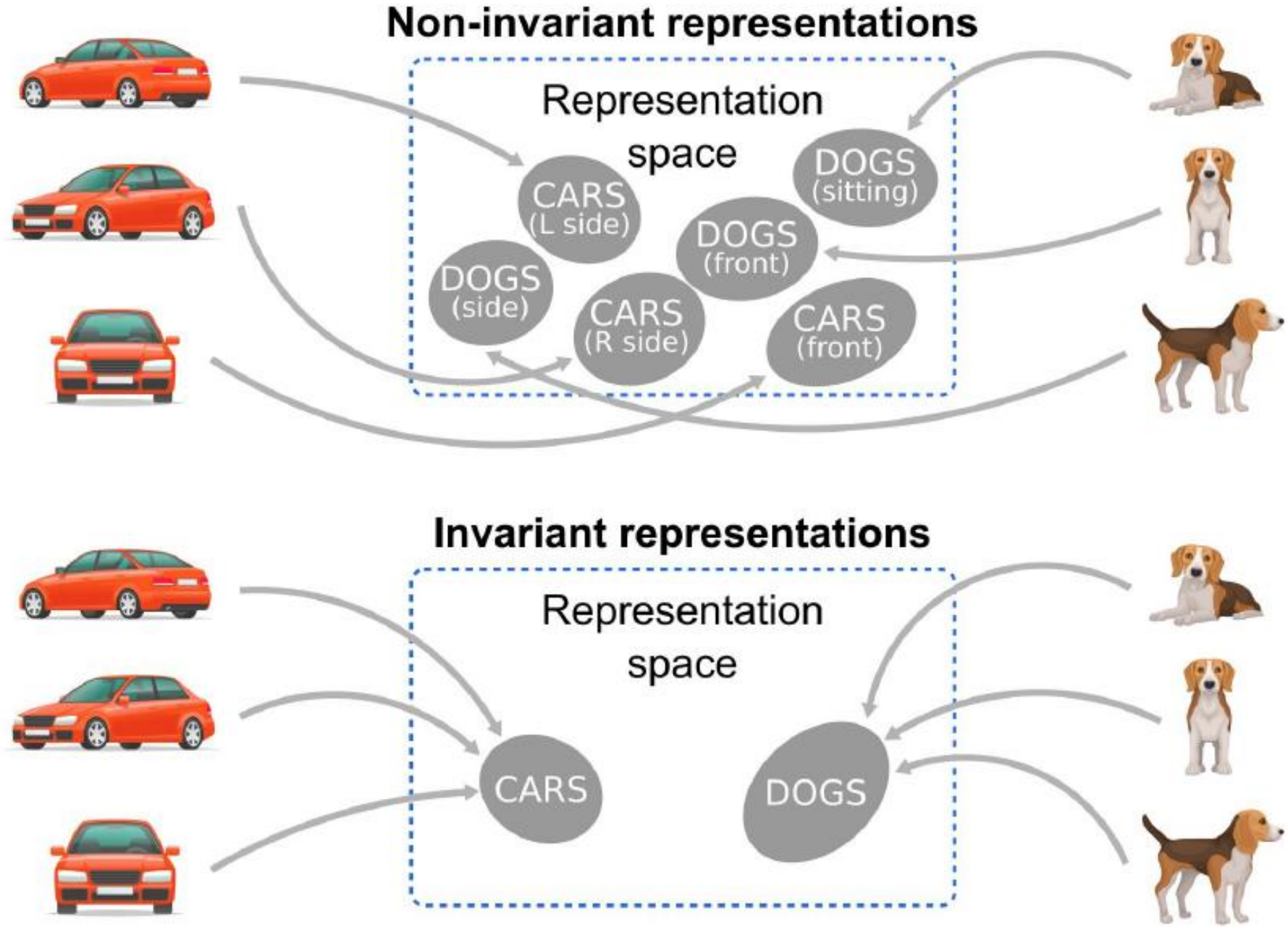
When we do supervised training of a neural network on some task, e.g. image categorization, the advantage of deep learning is precisely that it allows us to learn representations in the hidden layers that make the task easier at the final output layer





What are the key properties of the representations induced by supervised learning? **Invariance!**

- One of the key properties of the emergent representations from supervised learning is **invariance**
- This means that you get a similar representation for the same object (or object class) regardless of the specific viewpoint, lighting, placement in the image, etc.
- Invariant representations make it much easier for a downstream circuit to recognize a given object regardless of the conditions under which it is viewed


Non-invariant representations
Representation space
CARS (L side)
DOGS (sitting)
DOGS (front)
DOGS (side)
CARS (R side)
CARS (front)
Invariant representations
Representation space
CARS
DOGS

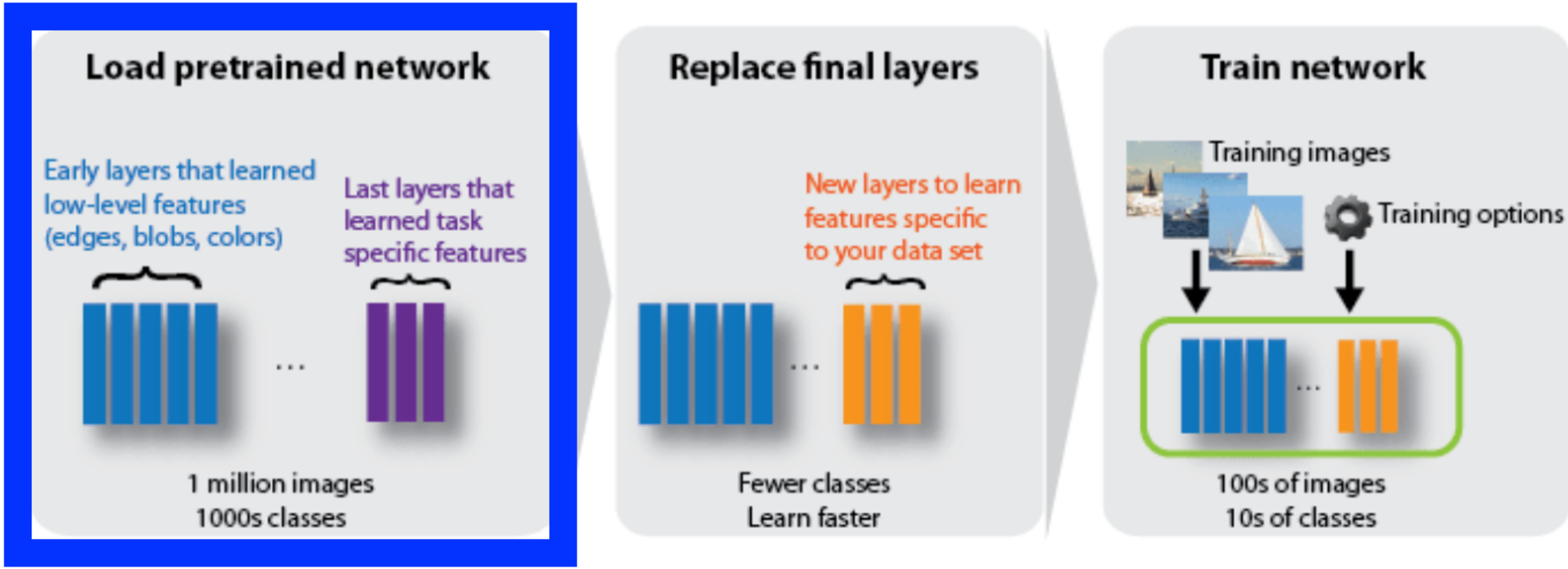
The ImageNet Challenge Story … outcomes

Strong supervision:

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification

Fine-Tuning (aka, Transfer Learning)

Key observation: features from a pretrained network can be useful for other datasets/tasks



ImageNet (1.2M images)

Flowers102 (2k images)

→ → Pretraining (Supervised)

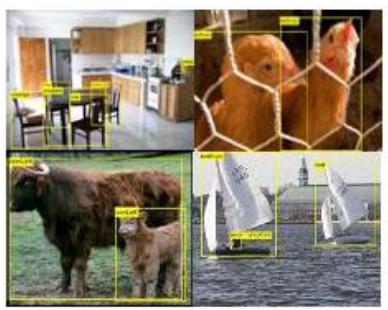
↓ Transfer (initialization)

→ → Linear evaluation or Fine-tuning

Supervised Pre-training + Fine-tuning

Pretraining on ImageNet Classification

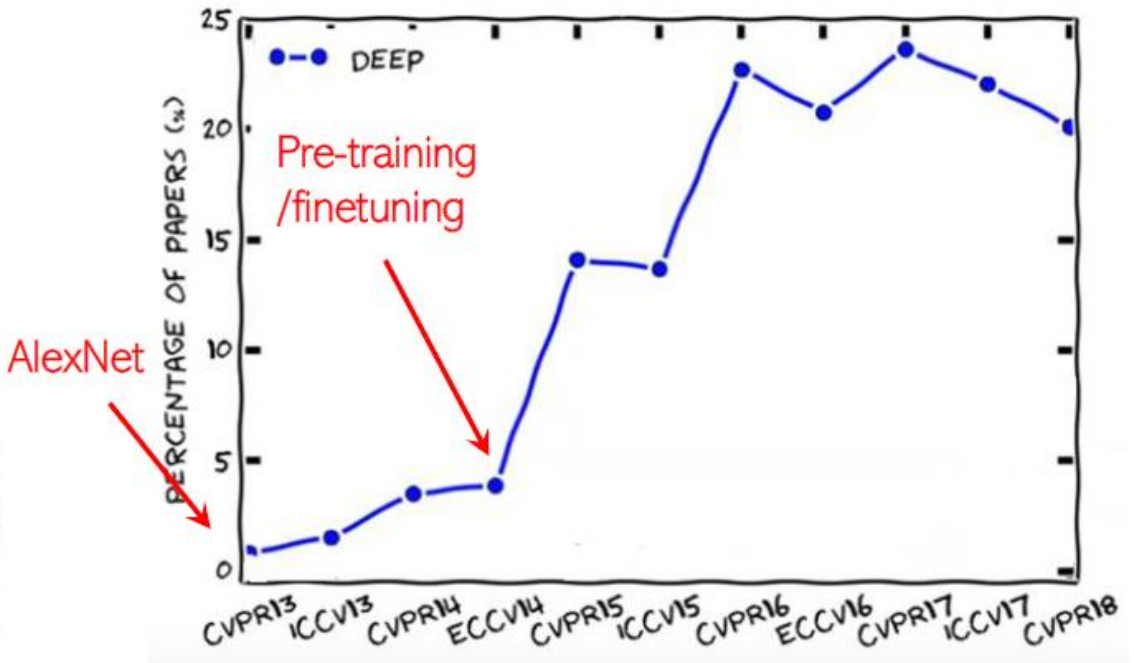
Finetuning



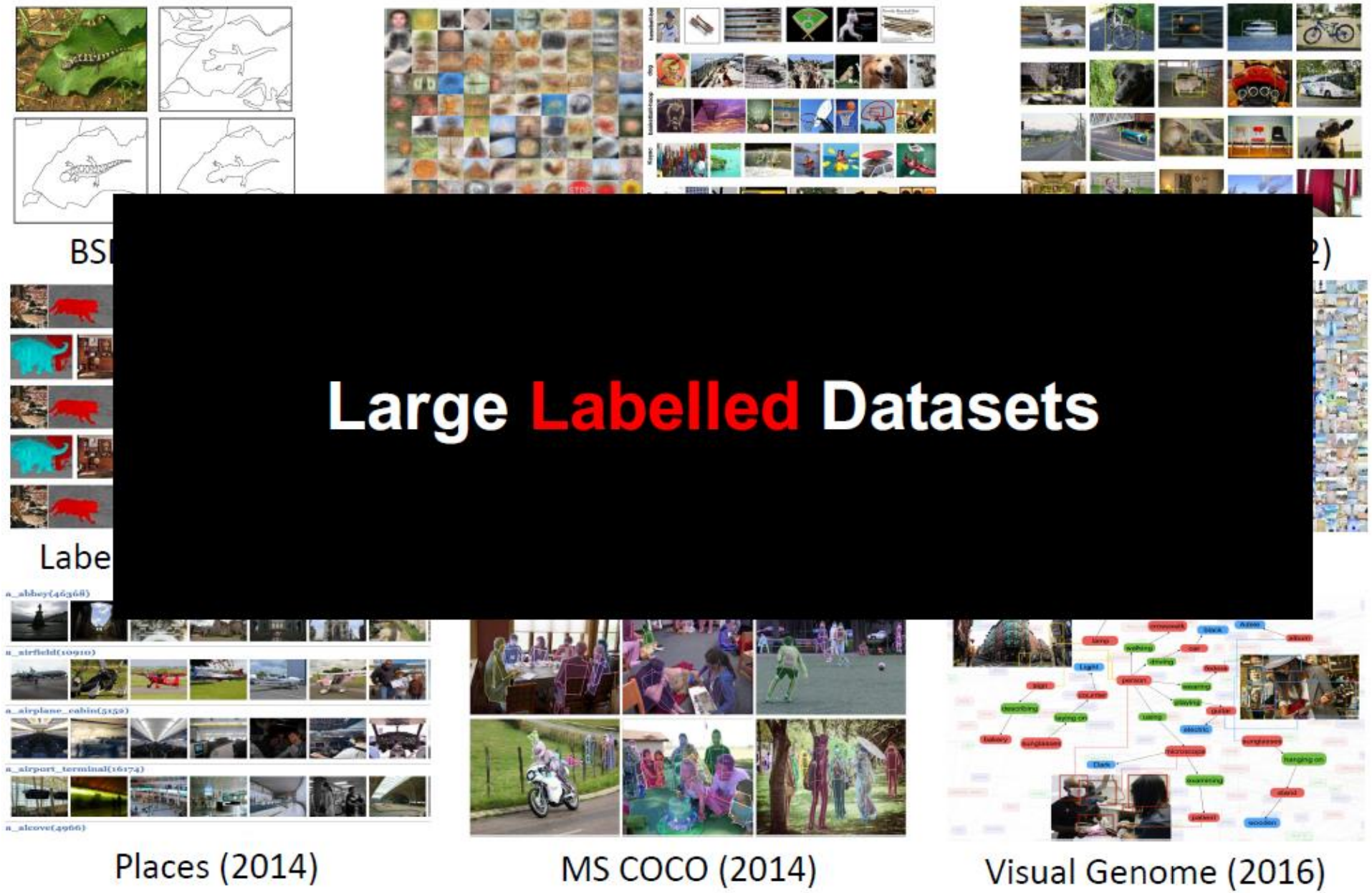
Object Detection

Semantic Segmentation

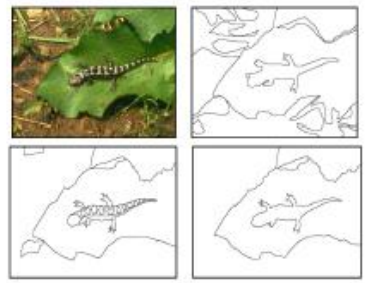
Fine-grained Classification



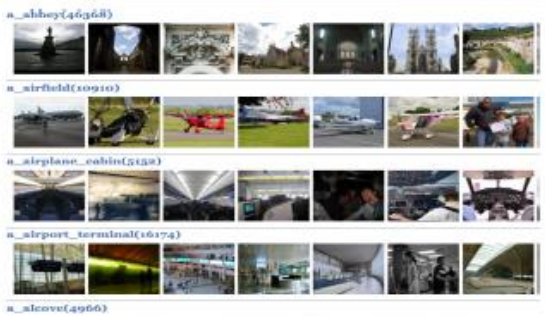
How Have Pretrained Networks Learned



Why Not Rely On Large Labelled Datasets?



- Expensive
- Relatively Slow to Build Dataset



Places (2014)

MS COCO (2014)

Visual Genome (2016)

The ImageNet Challenge Story … outcomes

Strong supervision:

- Features from networks trained on ImageNet can be used for other visual tasks, e.g. detection, segmentation, action recognition, fine grained visual classification

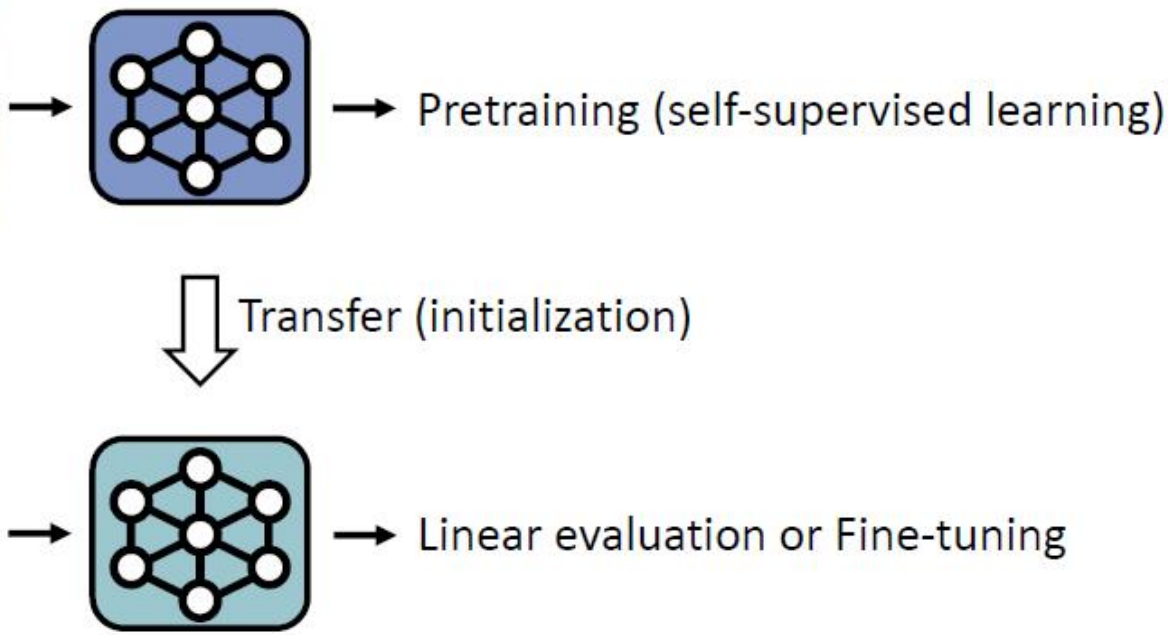
- Are there alternatives to strong supervision for training? Self-Supervised learning ….

Evaluating Self-supervised Representation

- **How to evaluate the quality of self-supervision?**
  1. Self-supervised learning in a large-scale dataset (e.g., ImageNet)
  2. Transfer the pretrained network to various downstream tasks
     - Linear probing: freeze the network and training only the linear classifier
       ⇒ **it directly evaluates the learned representation qualities**
     - Fine-tuning whole parameters

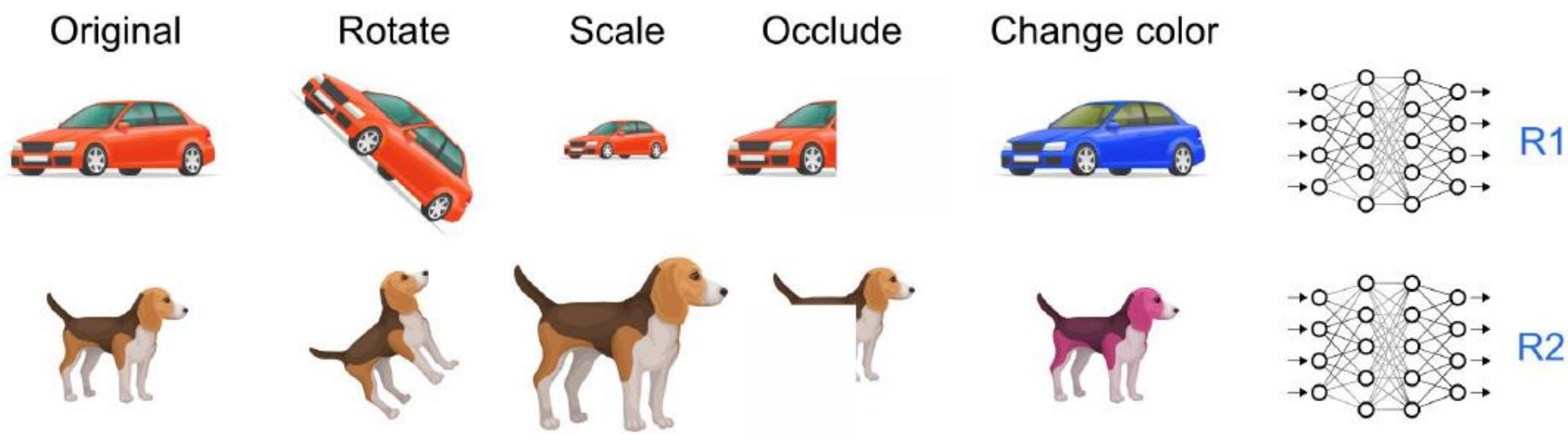
ImageNet (1.2M images)

Flowers102 (2k images)

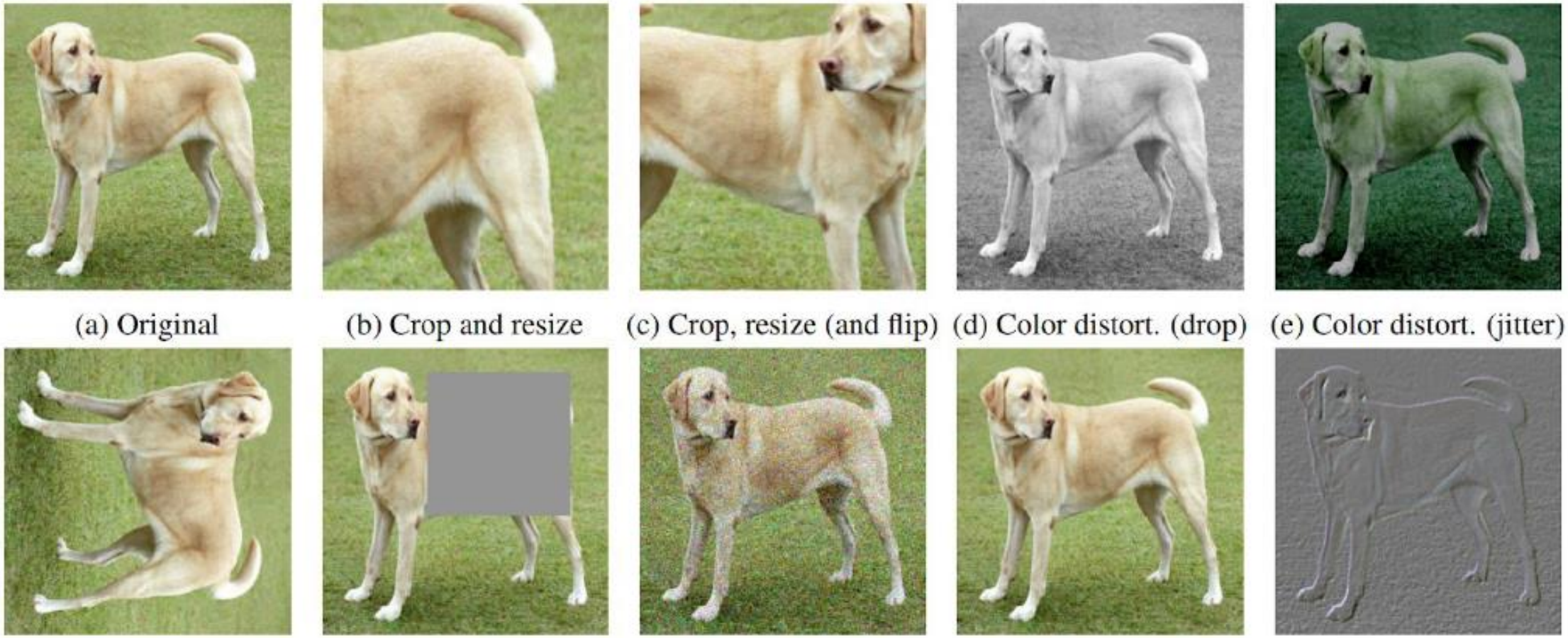


What if we aim for invariance to transformations?

- If our goal is invariant representations, then maybe we could just train directly towards that goal using data transformations/augmentations?



The transformations/augmentations used are critical for achieving good results



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Self-Supervised Learning

A form of unsupervised learning, where the data itself serves as supervision

What is Self-Supervised Learning?

Self-Supervised Learning (SSL) is a special type of representation learning that enables learning good data representation from unlabelled dataset.

It is motivated by the idea of *constructing supervised learning tasks out of unsupervised datasets*.

Why?

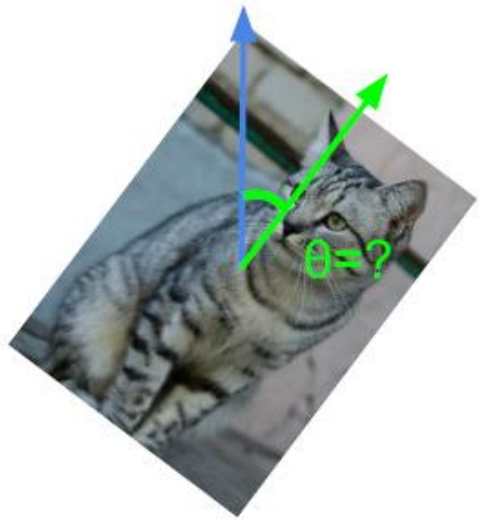
1. Data labeling is expensive and thus high-quality labeled dataset is limited.
2. Learning good representation makes it easier to transfer useful information to a variety of downstream tasks.
    - e.g. A downstream task has only a few examples.
    - e.g. Zero-shot transfer to new tasks.

Self-supervised learning tasks are also known as ***pretext tasks***.

Self-supervised pretext tasks

Example: learn to predict image transformations / complete corrupted images

image completion



rotation prediction

“jigsaw puzzle”

colorization

1. Solving the pretext tasks allow the model to learn good features.
2. We can automatically generate labels for the pretext tasks.

Methods

- Self-prediction
- Contrastive learning

Self-Prediction

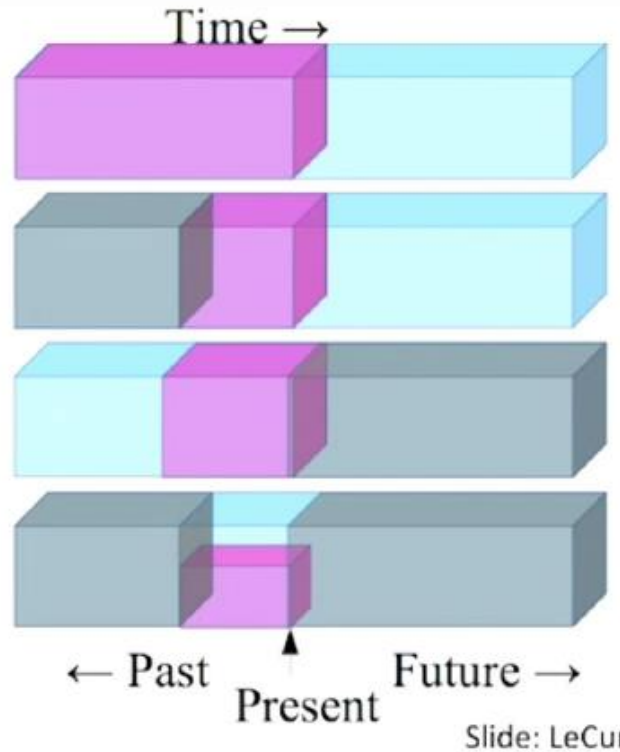
Self-prediction construct prediction tasks within every individual data sample: to predict a part of the data from the rest while pretending we don't know that part.

1. Autoregressive generation
2. Masked generation
3. Innate relationship prediction
4. Hybrid self-prediction

Self-Prediction

Self-prediction construct prediction tasks within every individual data sample: to predict a part of the data from the rest while pretending we don't know that part.

- **Predict any part of the input from any other part.**
- **Predict the future from the past.**
- **Predict the future from the recent past.**
- **Predict the past from the present.**
- **Predict the top from the bottom.**
- **Predict the occluded from the visible**
- **Pretend there is a part of the input you don't know and predict that.**



(Famous illustration from Yann LeCun)

Methods

- Self-prediction
- Contrastive learning

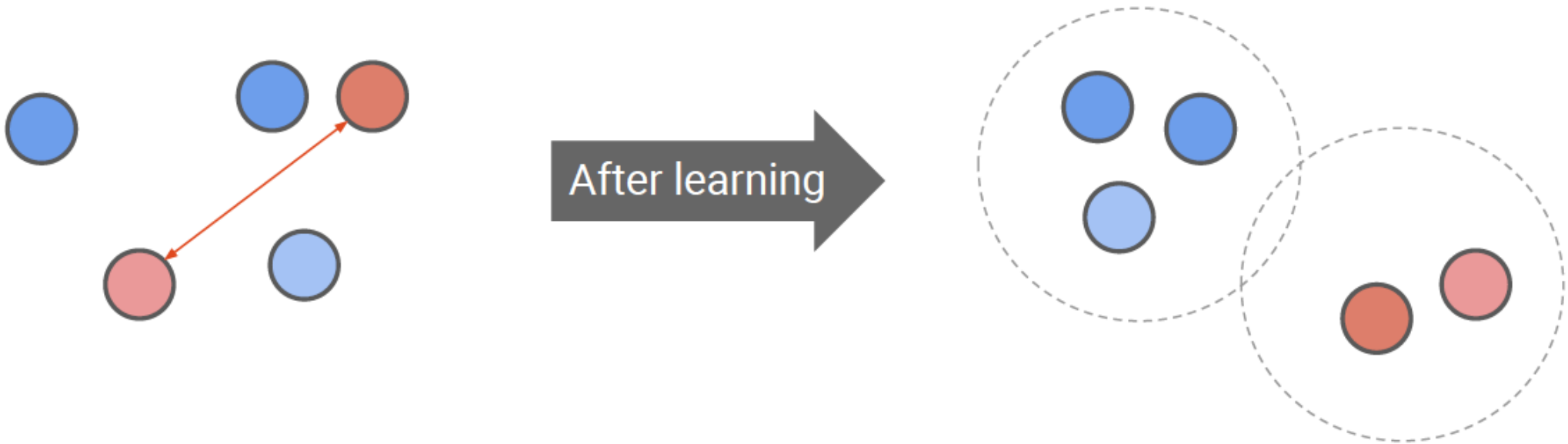
Contrastive Learning

The goal of contrastive representation learning is to learn such an embedding space in which *similar* sample pairs stay *close* to each other while *dissimilar* ones are *far apart*.

1. Inter-sample classification
2. Feature clustering
3. Multiview coding

Contrastive Learning

The goal of contrastive representation learning is to learn such an embedding space in which *similar* sample pairs stay *close* to each other while *dissimilar* ones are *far apart*.



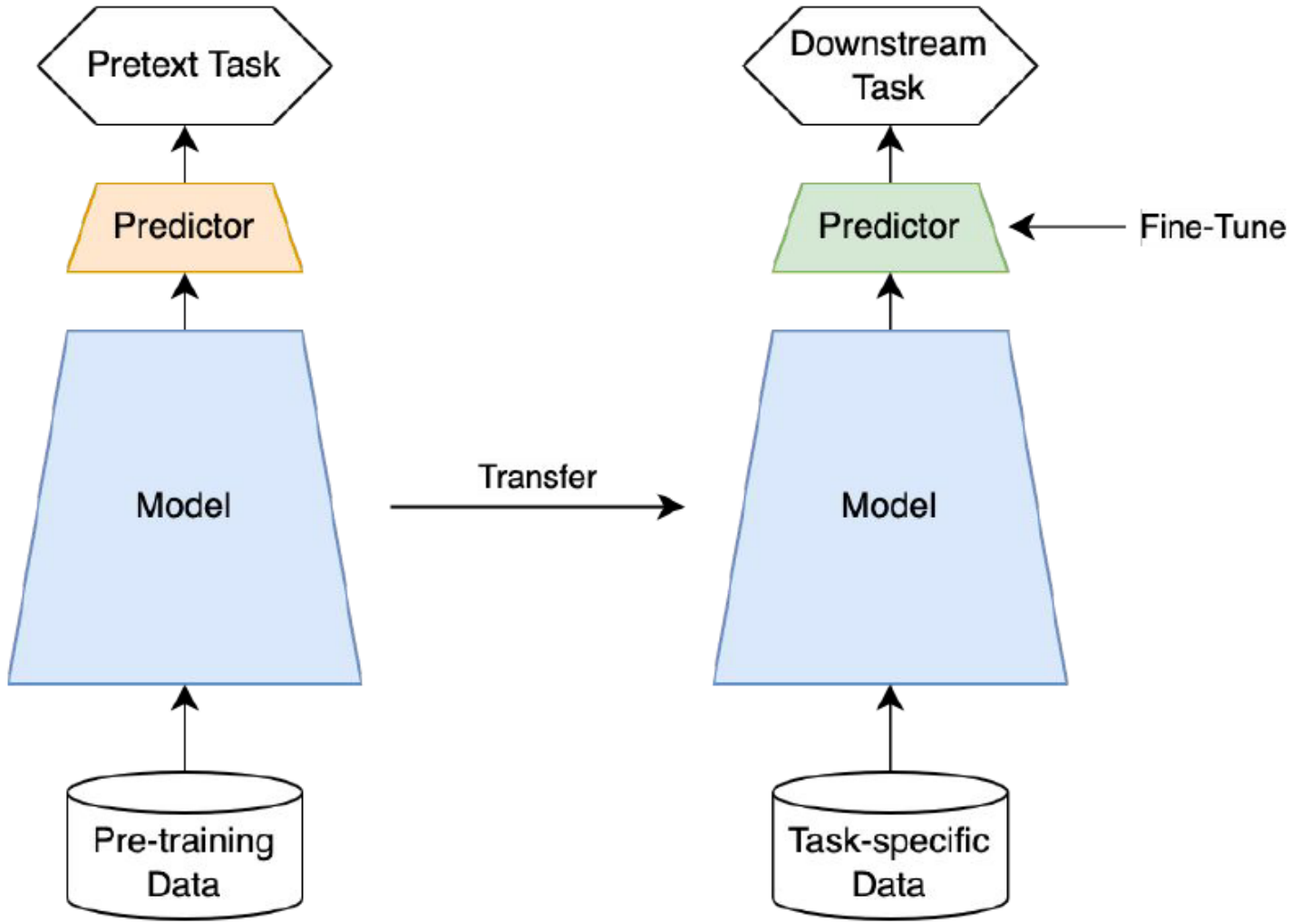
Pretext Tasks

- Vision
- Video
- Audio
- Multimodal
- Language

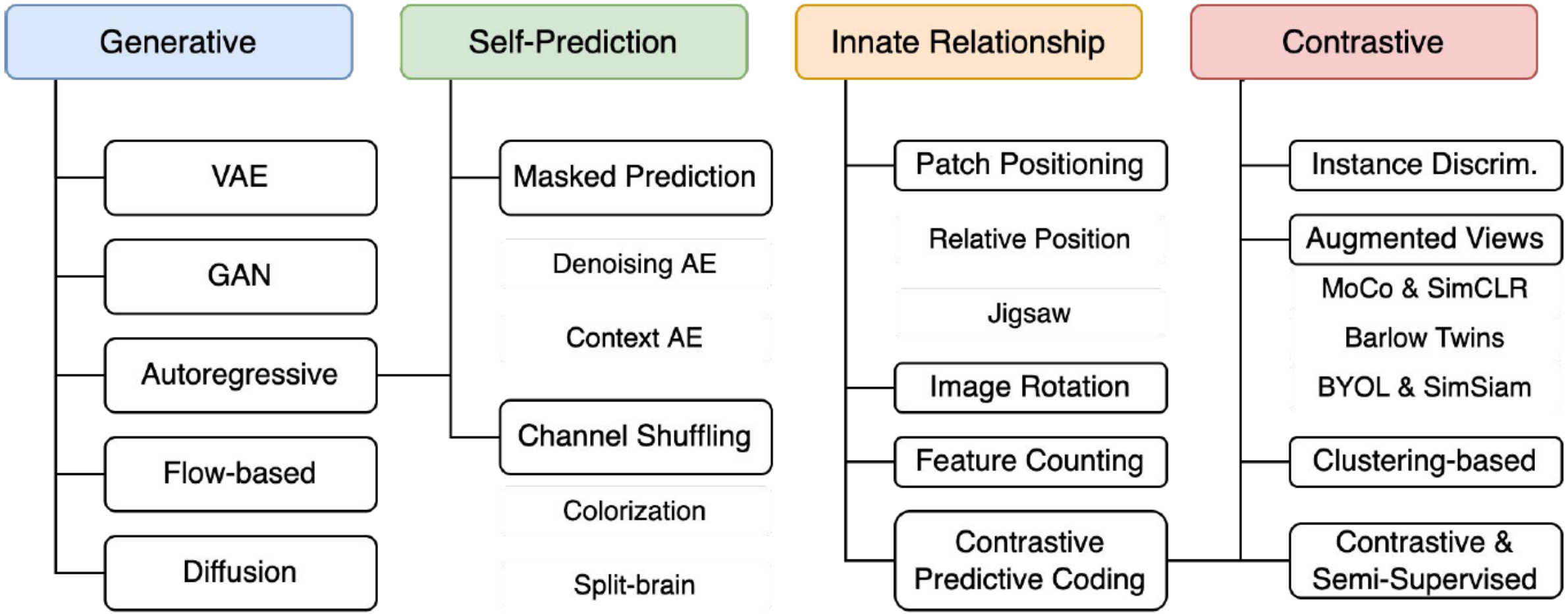
Recap: Pretext Tasks

Step 1: Pre-train a model for a pretext task

Step 2: Transfer to applications



Pretext Tasks: Taxonomy



On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Self-Supervised Learning (SSL)

- What makes Foundation Models (FMs) tick?
- What is “under-the-hood” of FMs?

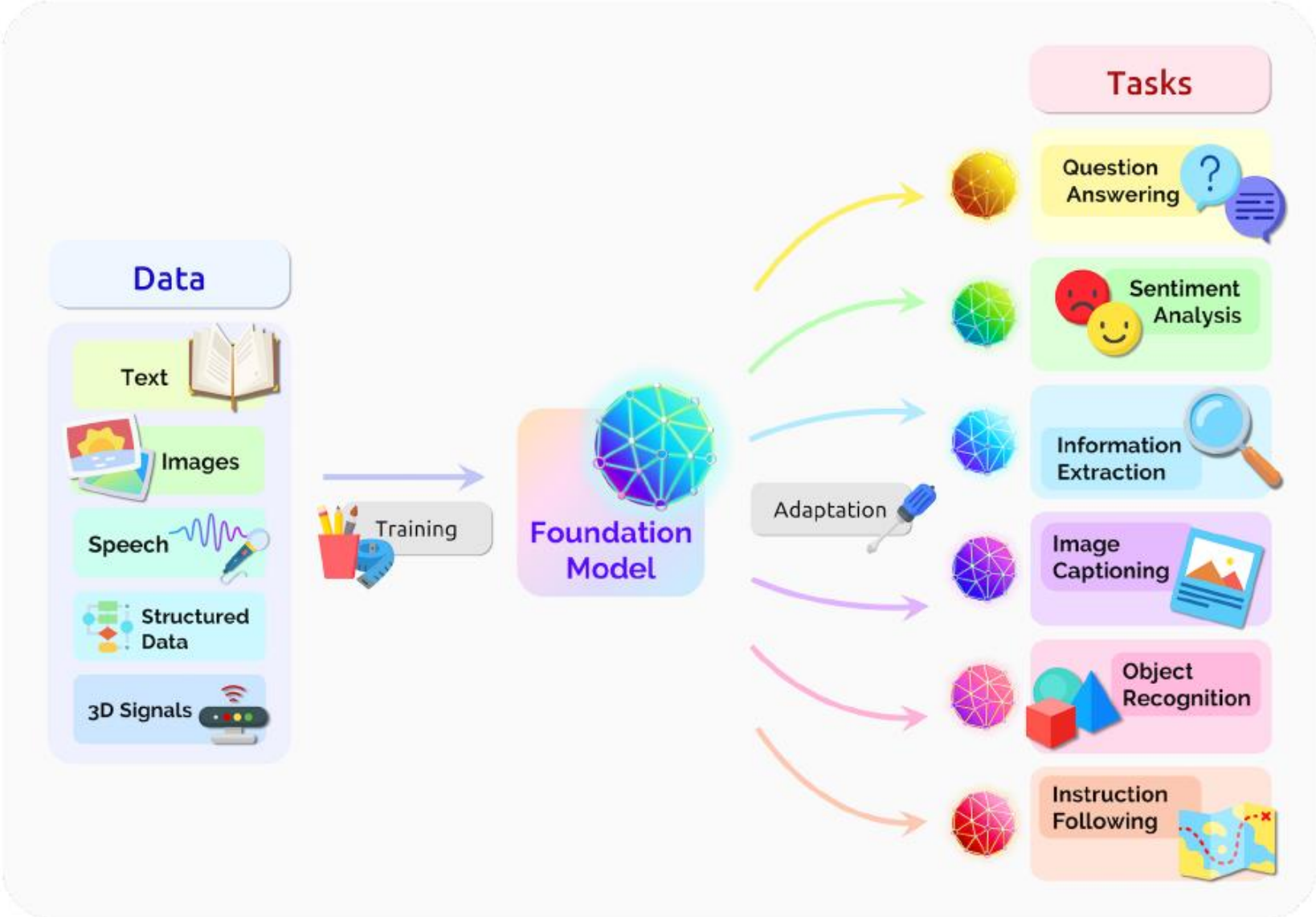


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

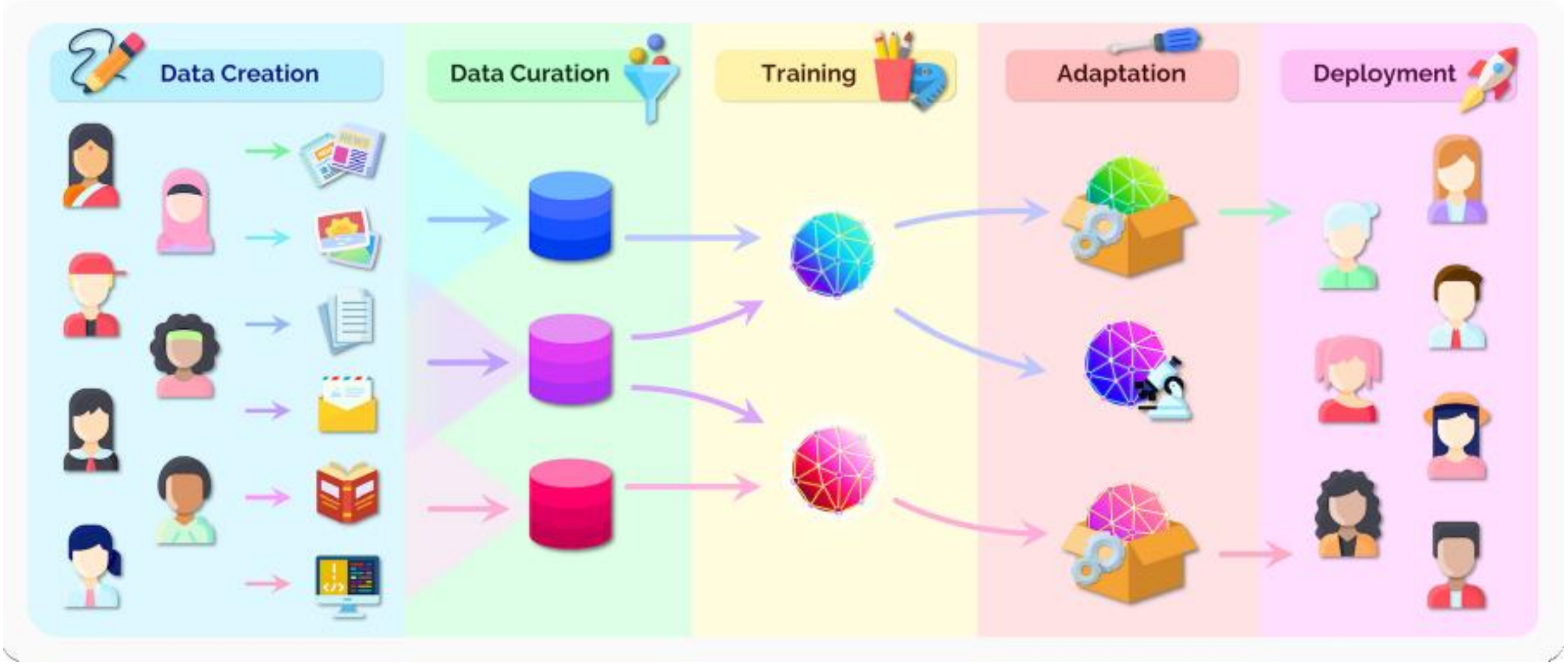


Fig. 3. Before reasoning about the social impact of foundation models, it is important to understand that they are part of a broader ecosystem that stretches from data creation to deployment. At both ends, we highlight the role of people as the ultimate source of data into training of a foundation model, but also as the downstream recipients of any benefits and harms. Thoughtful data curation and adaptation should be part of the responsible development of any AI system. Finally, note that the deployment of adapted foundation models is a decision separate from their construction, which could be for research.

Contents

Paper Roadmap

2. Capabilities

Language
2.1

Vision
2.2

Robotics
2.3

Reasoning
2.4

Interaction
2.5

Philosophy
2.6

3. Applications

Healthcare
3.1

Law
3.2

Education
3.3

4. Technology

Modeling
4.1

Training
4.2

Adaptation
4.3

Evaluation
4.4

Systems
4.5

Data
4.6

Security
4.7

Robustness
4.8

AI Safety
& Alignment
4.9

Theory
4.10

Interpretability
4.11

5. Society

Inequity
5.1

Misuse
5.2

Environment
5.3

Legality
5.4

Economics
5.5

Ethics
5.6

4.10 Theory

Authors: Aditi Raghunathan, Sang Michael Xie, Ananya Kumar, Niladri Chatterji, Rohan Taori, Tatsunori Hashimoto, Tengyu Ma

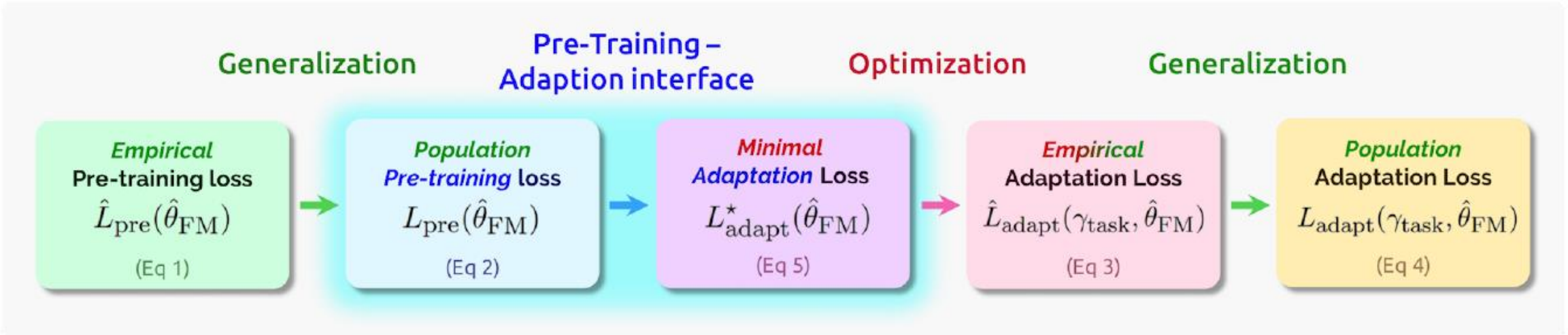
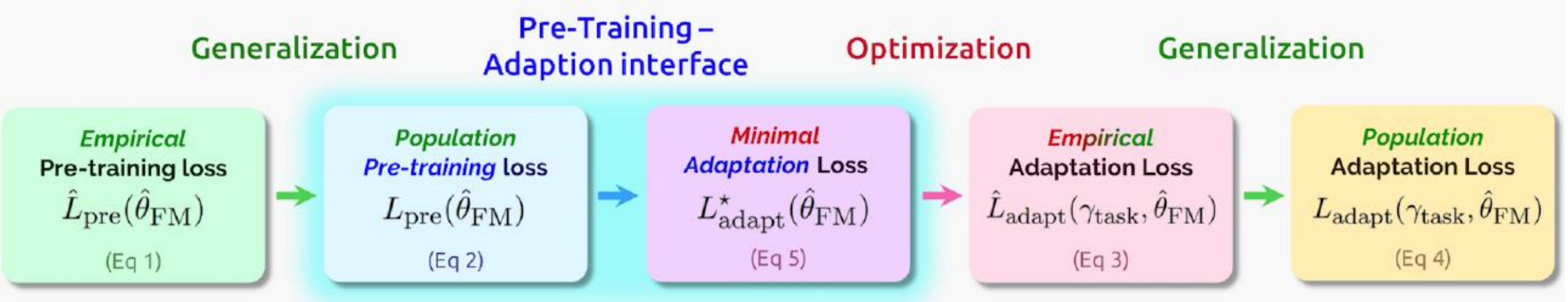


Fig. 22. The analysis of foundation models from pretraining on diverse data to downstream performance on adapted tasks involves capturing the relation between different loss terms as shown above. The main challenge is to analyze the highlighted pretraining-adaptation interface which requires reasoning carefully about the population losses in addition to the model architecture, losses and data distributions of the pretraining and adaptation stages (§4.10.2: THEORY-INTERFACE). Analysis of generalization and optimization largely reduces to their analysis in standard supervised learning.



As shown in Figure 22, the main missing link beyond standard supervised theory is:

Under what conditions does a small population pretraining loss $L_{pre}(\hat{\theta}_{\text{FM}})$ *imply a small minimal adaptation loss* $L^{\star}_{adapt}(\hat{\theta}_{\text{FM}})$ and why?

Thank you !!