Deep Learning for Computer Vision (Discriminative way)

Andrii Liubonko

2025

About me

- Theoretical Physics background
  - Ukraine & Germany
- Samsung (7 years)
  - mainly Computer Vision projects
- Grammarly (3 years)
  - NLP related projects
- EPAM (current)
  - mix of LLMs & NLP & Computer Vision

reach me on Slack if you have any questions

Course structure

Module 1. Basic image processing - Oles Dobosevych

Module 2. Learning Discriminative Models - Andrii Liubonko

Module 3. Image Segmentation, Image Correspondence - Maksym Davydov

Module 4. Metric Learning, Representation Learning and Context Importance, Tracking - Igor Krashenyi

Module 5. Learning Generative Models - Andrii Liubonko

Logistics (this module)

6 lectures

homework:

- assignment 1 : tutorials notebooks 50%
- assignment 2 : mini-project 50%

deadline date: *TBD*

course repo:
https://github.com/lyubonko/ucu2025cv

Overview of the module

Lecture I Intro, big picture

Lecture II Essential architectures (CNN)

Lecture III Essential architectures (Transformers)

Lecture IV CV Foundational models in depth (DINO, SAM)

Lecture V Multimodal Foundational models (GEMMA, APIs)

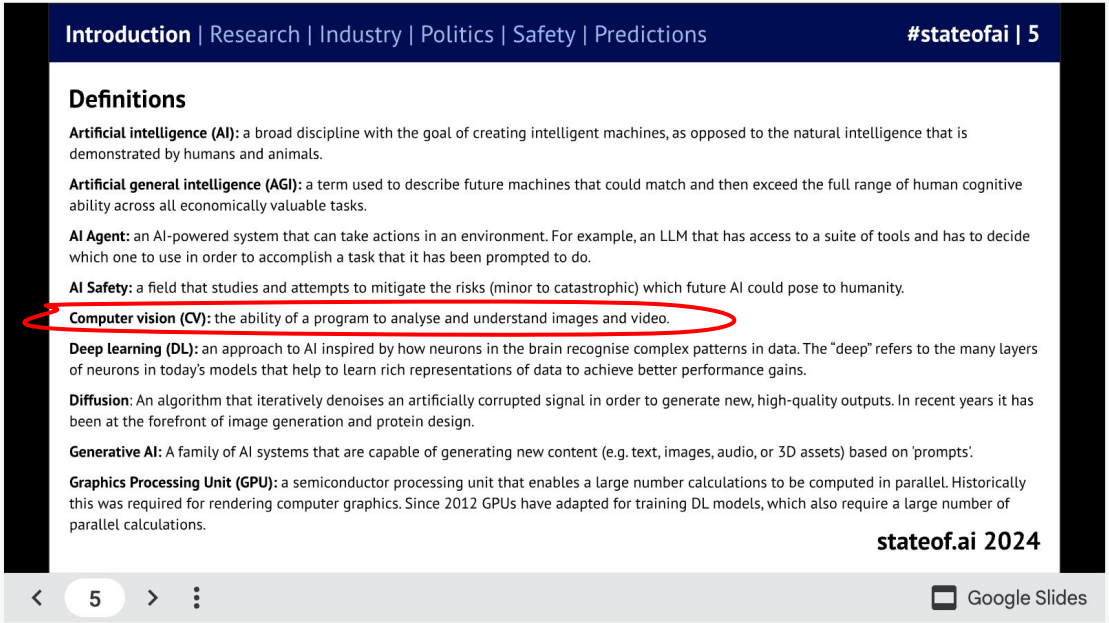
Lecture VI Object Detection

Goals of the Course

- Get big picture of the deep learning for CV
- Working knowledge of essential elements/blocks of **Convolutional Neural Networks [CNNs]**
- Working knowledge of essential elements/blocks of **Transformers** and their use in CV
- Get flavor of CV-related **Foundational models**
- Get deeper with one particular problem (**Object Detection**)

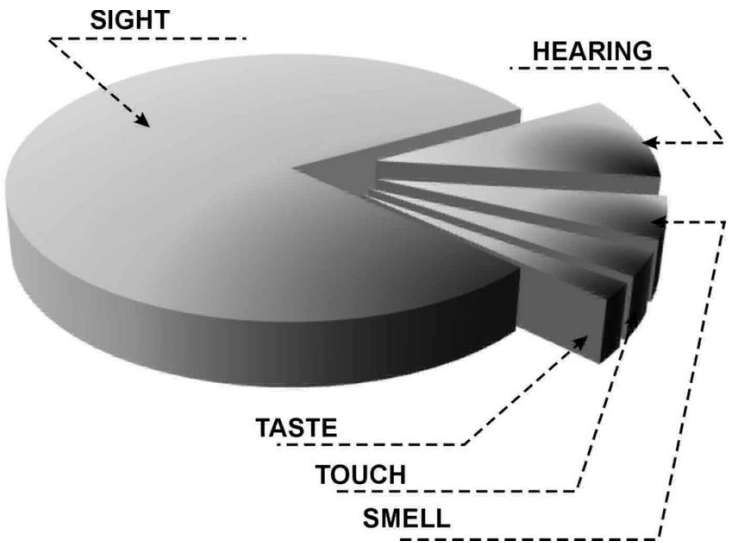
Intro

The **State of AI Report** analyses the most interesting developments in AI. We aim to trigger an informed conversation about the state of AI and its implication for the future. The Report is produced by AI investor **Nathan Benaich** and **Air Street Capital**.



https://www.stateof.ai/

Intro



Computer Vision *aims*

- ➔ *to extract useful information from visual input* **(Discriminative Models)**
- ➔ *to generate novel visual content* **(Generative Models)**

➔ *to extract useful information from visual input* **(Discriminative Models)**

- indoor/outdoor? [image classification]
- Where are the objects? [object detection]
- How far is the object ? [depth estimation]
- What people are doing? [activity recognition]
- Is the state of the environment normal? [anomaly detection]
- ...

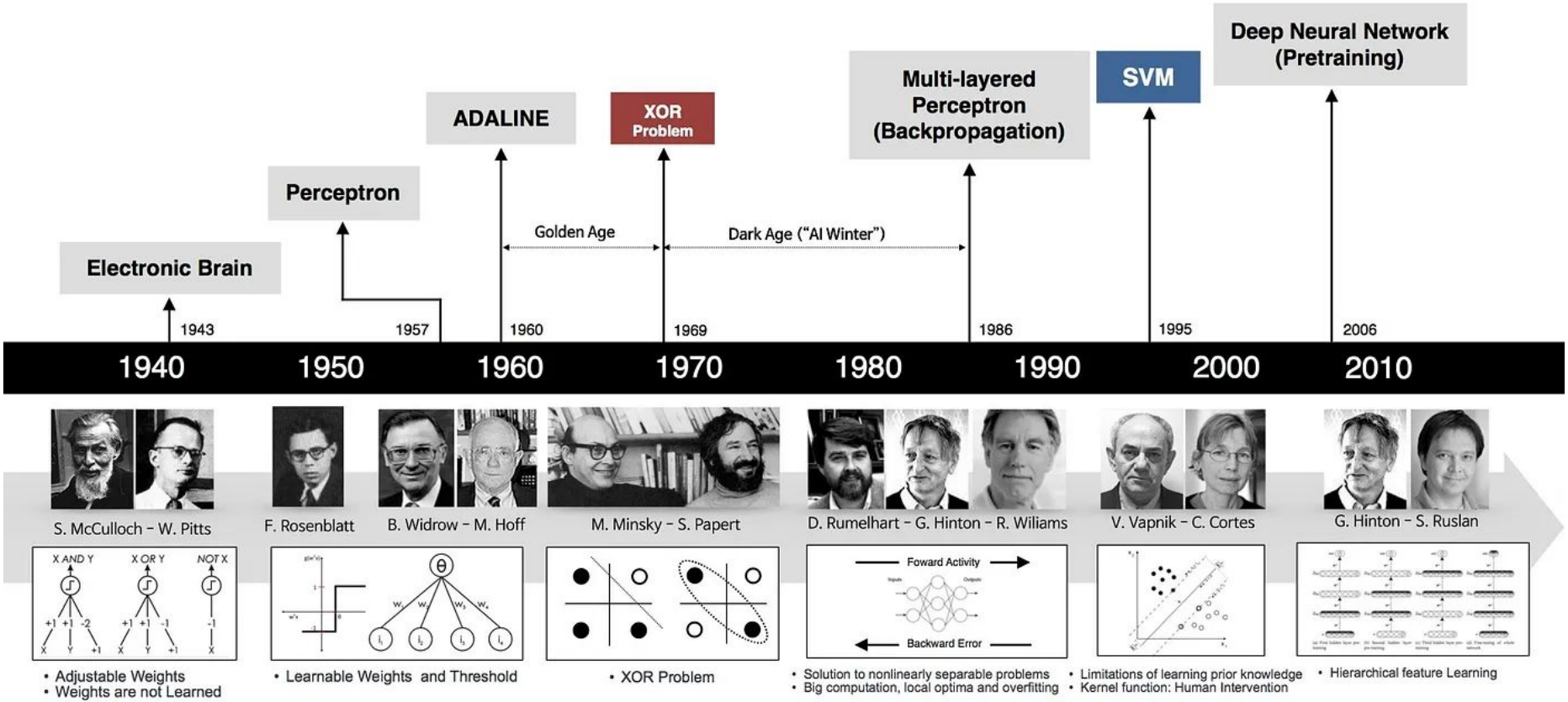
➔ *to generate novel visual content* **(Generative Models)**

2011.09055

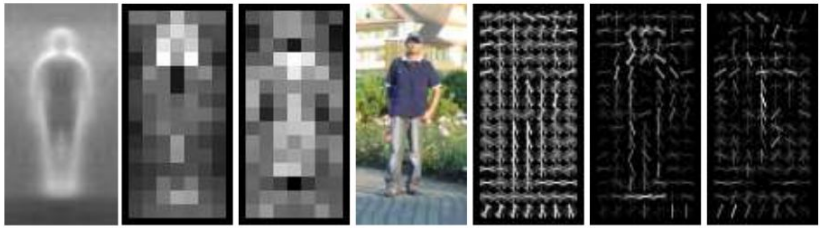
SORA (openAI)
VEO (Google)

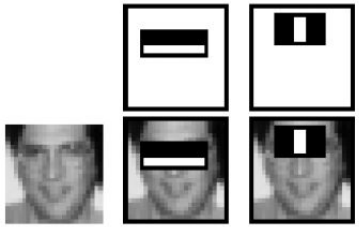
A (very) Brief History

History till 2012

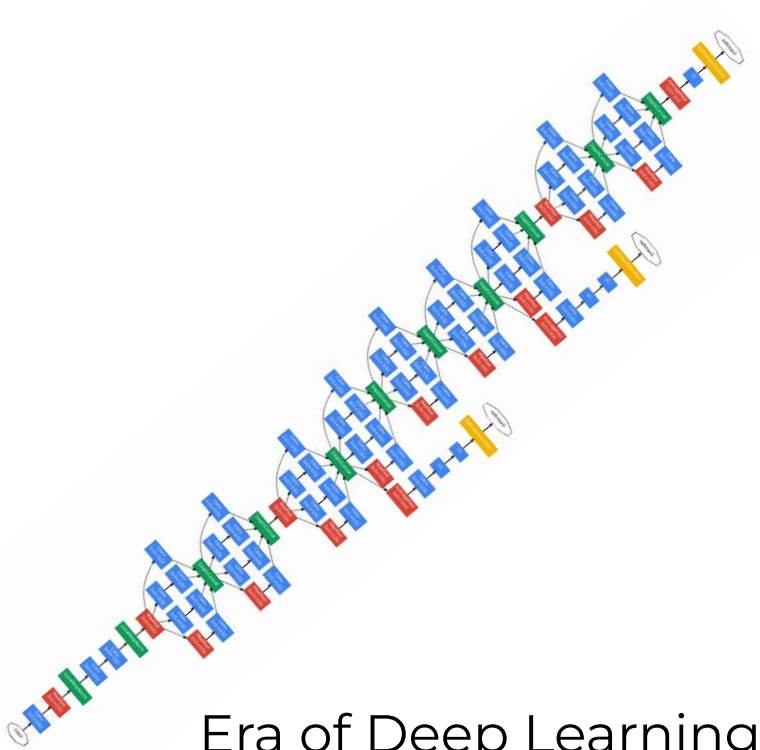


Intro





Era of Human-Crafter Features



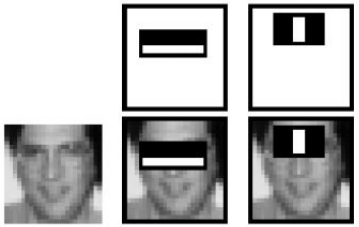
Era of Deep Learning

1986
BackProp

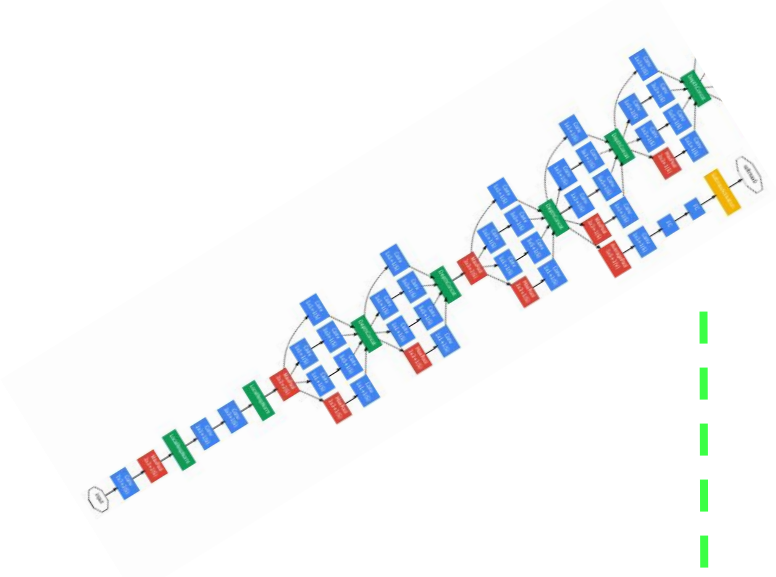
1998
LeNet

2012
AlexNet

Intro



Era of
Human-Crafter
Features



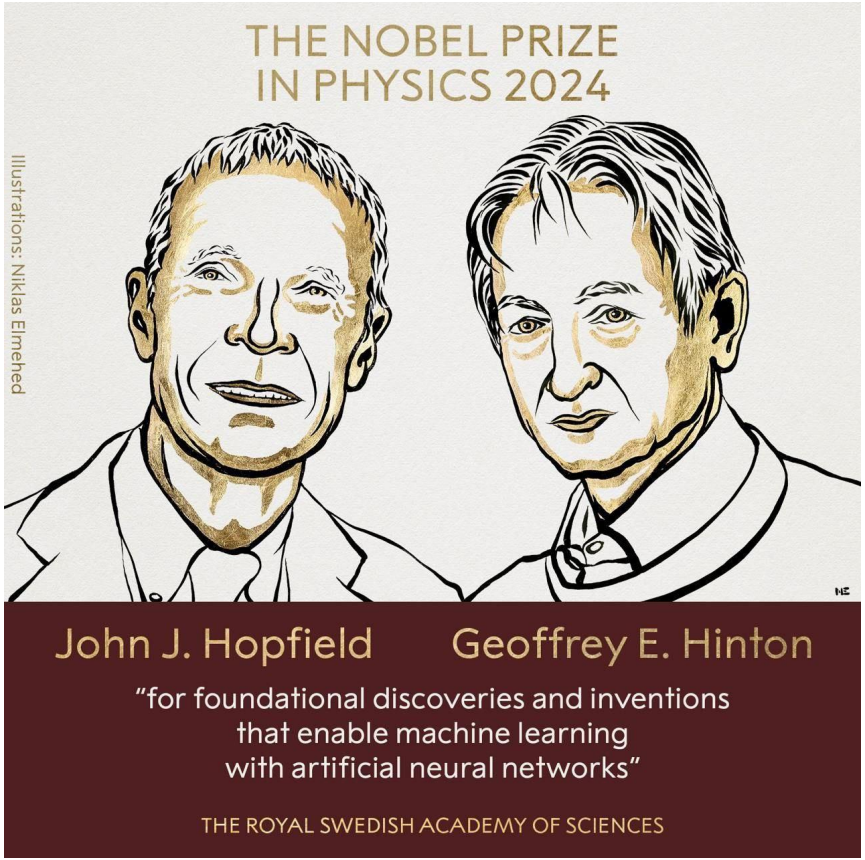
Era of
Deep Learning

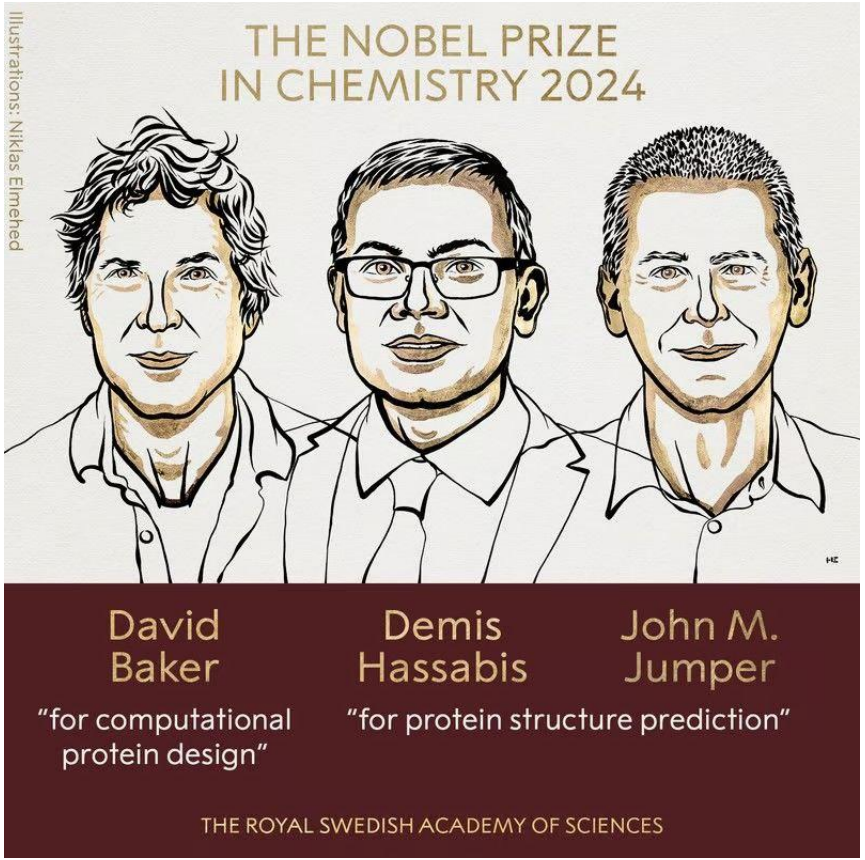
Era of
LLMs

2012
AlexNet

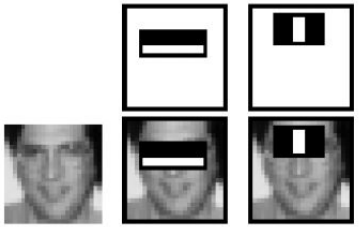
2022
ChatGPT

Intro

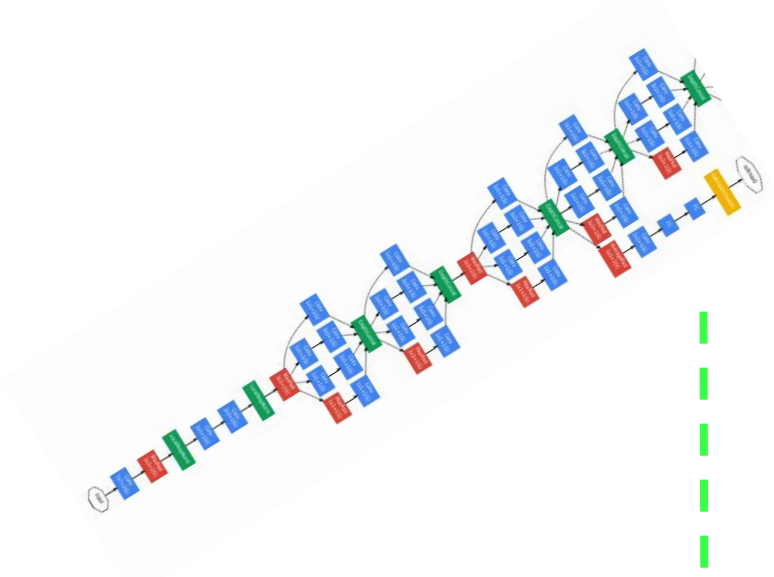

THE NOBEL PRIZE
IN PHYSICS 2024
Illustrations: Niklas Elmehed
John J. Hopfield
Geoffrey E. Hinton
"for foundational discoveries and inventions
that enable machine learning
with artificial neural networks"
THE ROYAL SWEDISH ACADEMY OF SCIENCES


THE NOBEL PRIZE
IN CHEMISTRY 2024
Illustrations: Niklas Elmehed
David Baker
Demis Hassabis
John M. Jumper
"for computational protein design"
"for protein structure prediction"
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Intro



Era of
Human-Crafter
Features



Era of
Deep Learning

Era of
LLMs

2012
AlexNet

2022
ChatGPT

Intro

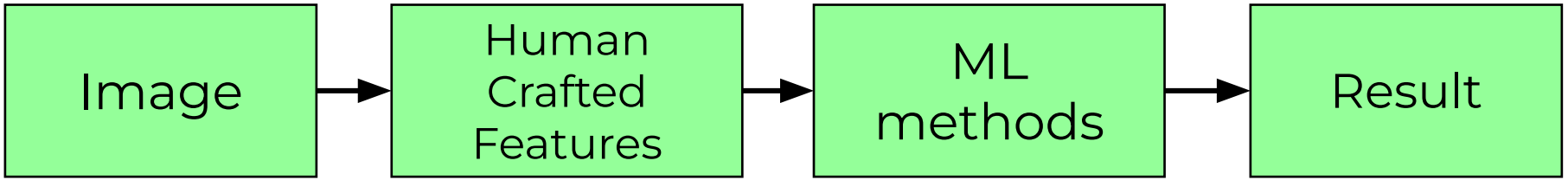
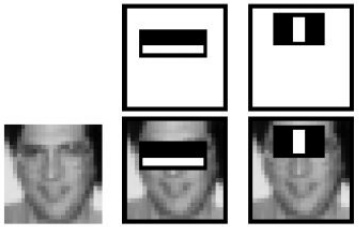
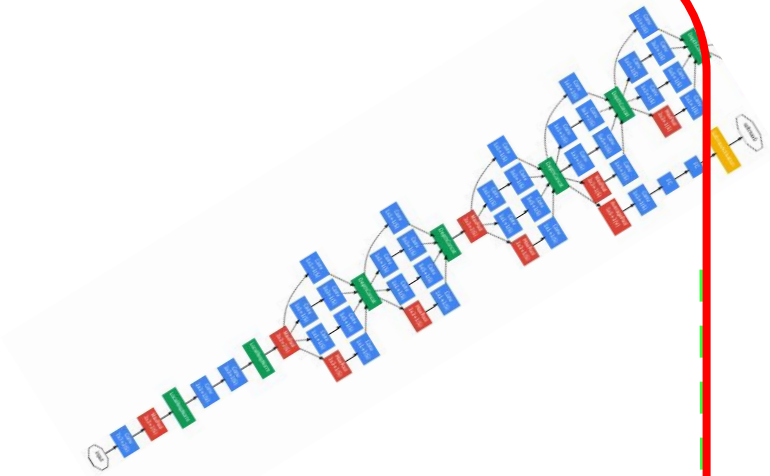


Image
Human Crafted Features
ML methods
Result
Descriptors: SIFT, SURF, etc.
Linear, SVM, XGBoost, ...

Intro



Era of
Human-Crafter
Features



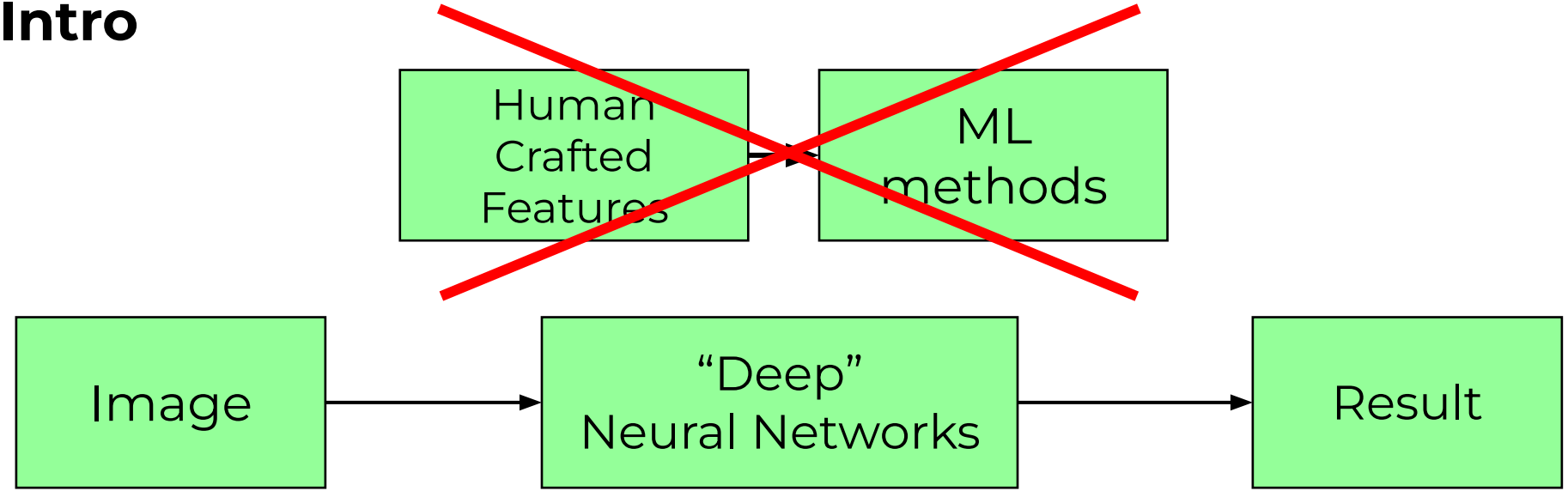
Era of
Deep Learning

Era of
LLMs

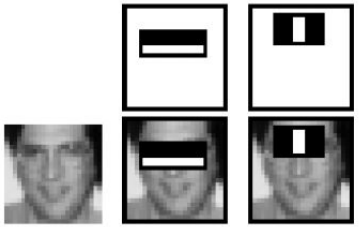
2012
AlexNet

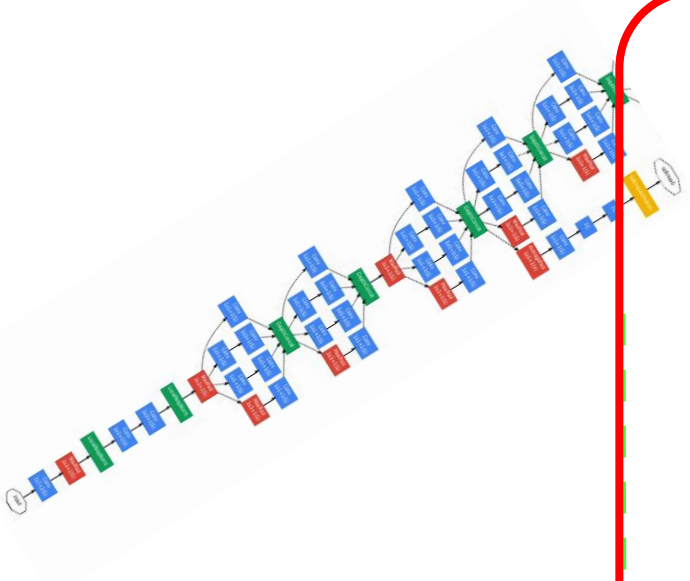
2022
ChatGPT

Intro


Human Crafted Features
ML methods
Image
“Deep” Neural Networks
Result

Intro





Era of
Human-Crafter
Features

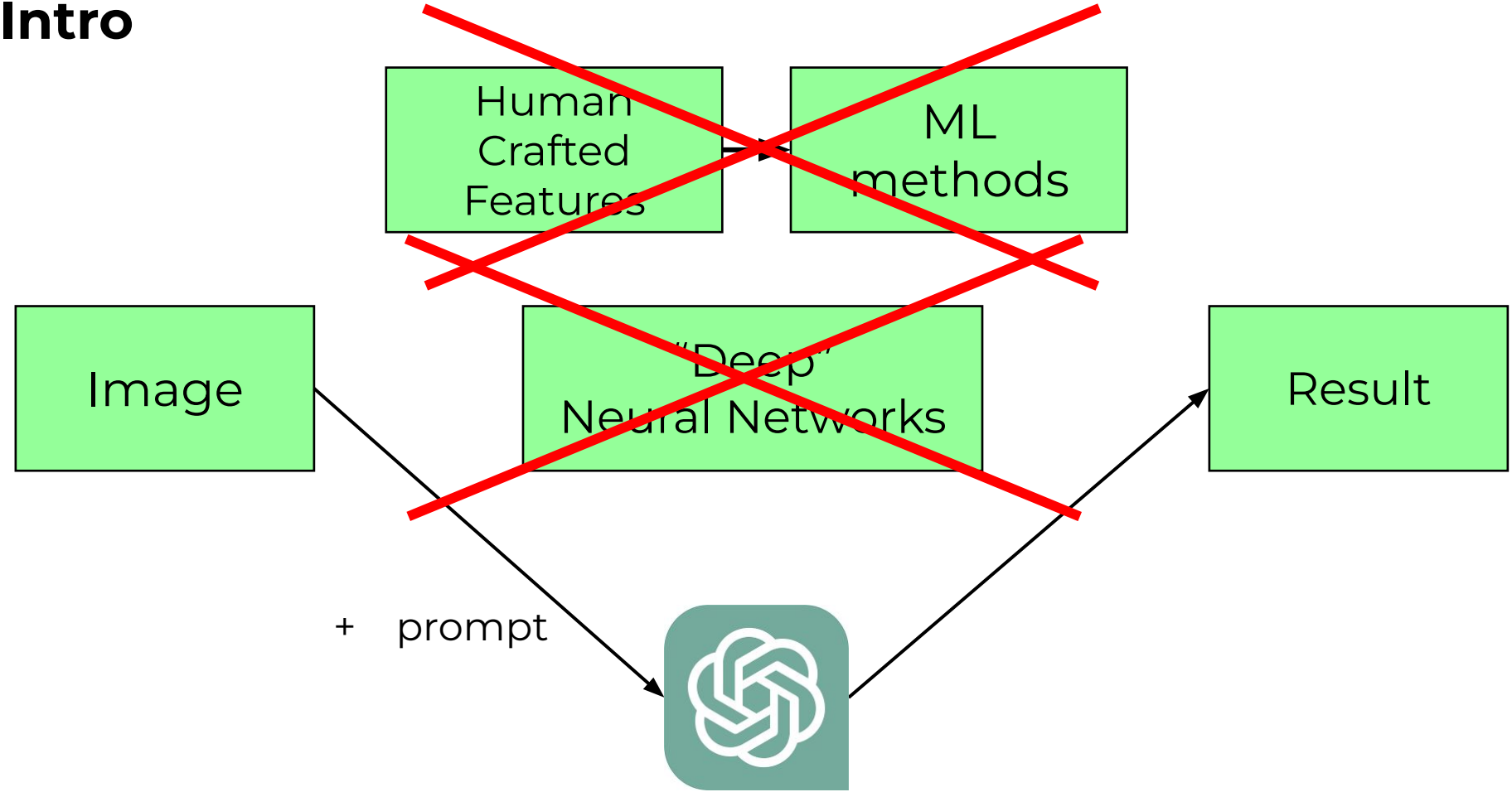
Era of
Deep Learning

Era of
LLMs

2012
AlexNet

2022
ChatGPT

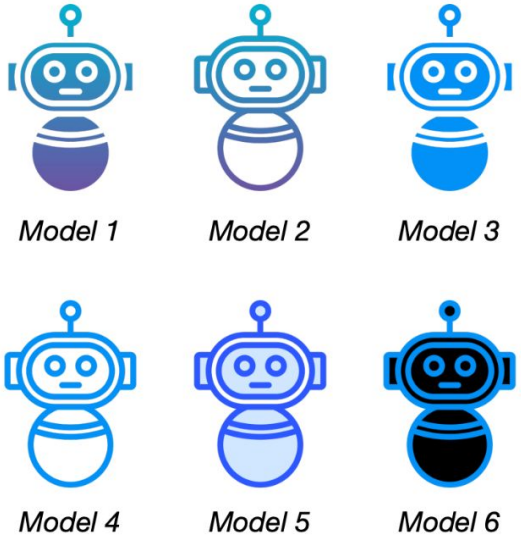
Intro


Human Crafted Features
ML methods
Image
"Deep" Neural Networks
Result
+ prompt

Intro

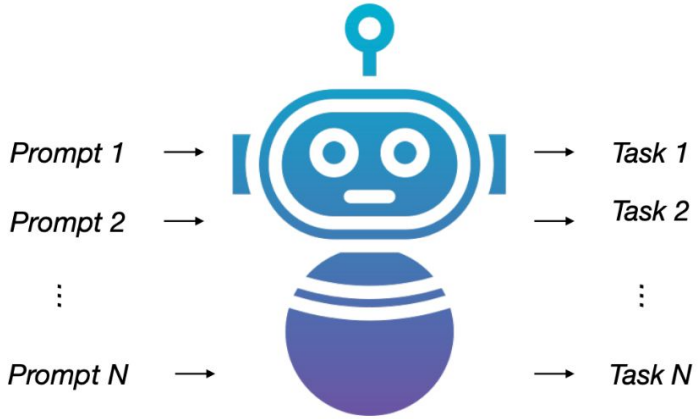
traditional ML

Old days: one model for one purpose



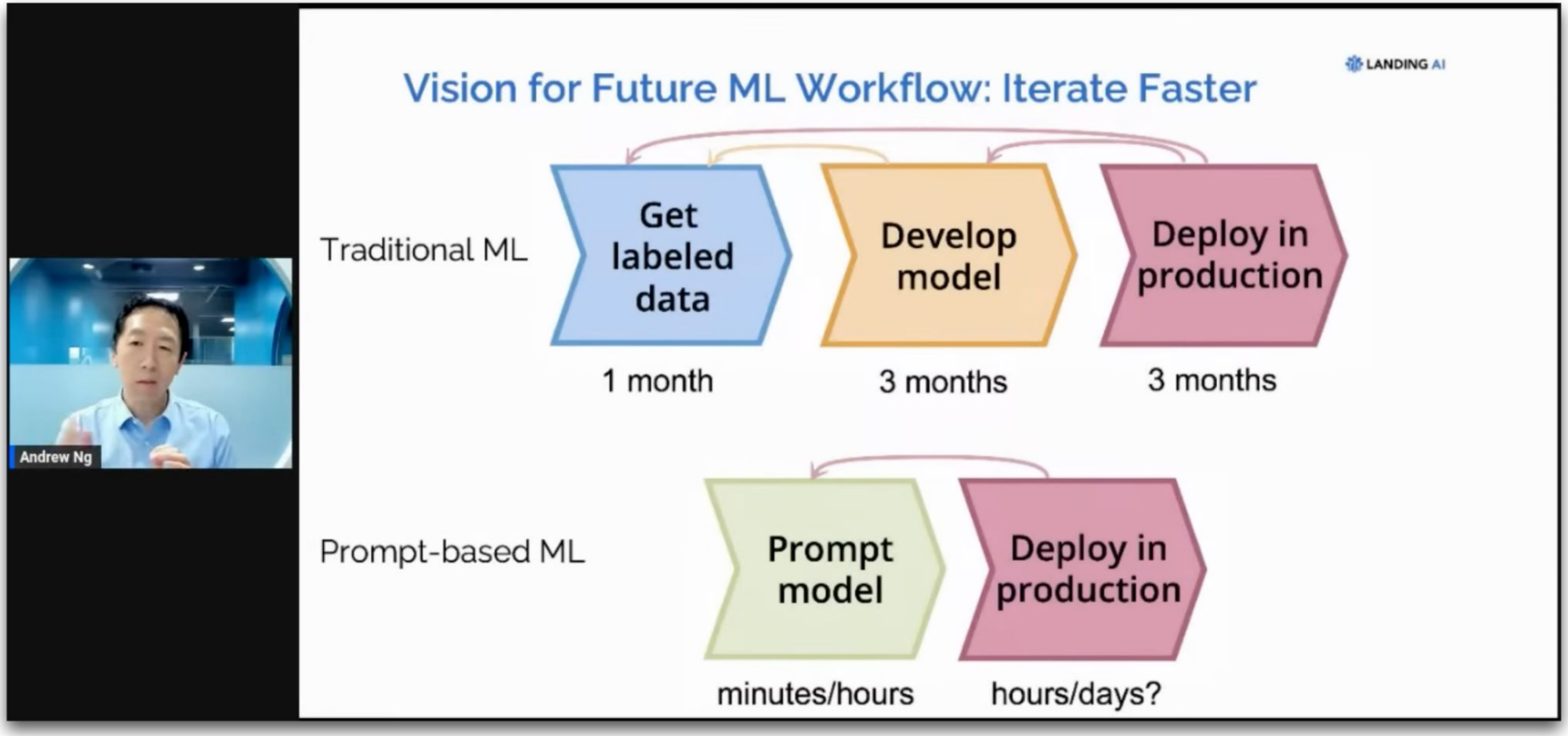
prompt-based ML

Now: one model for multiple purposes



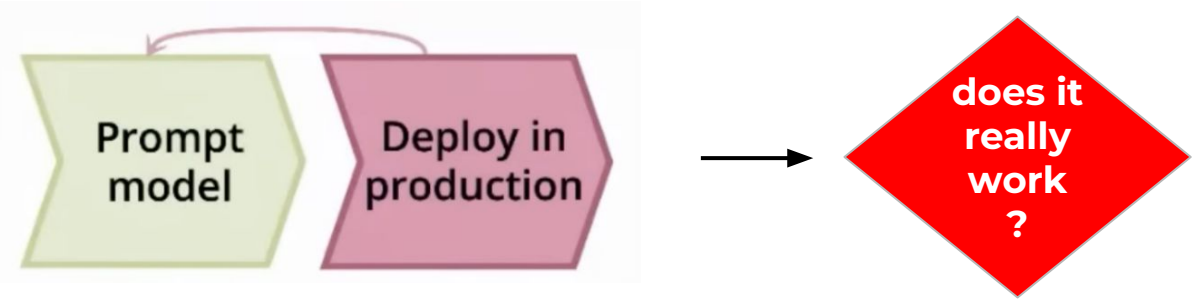
"Foundation model"

Intro

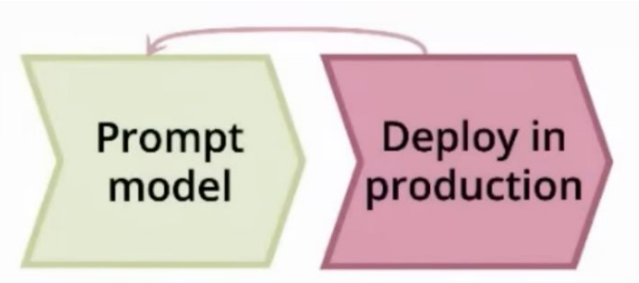


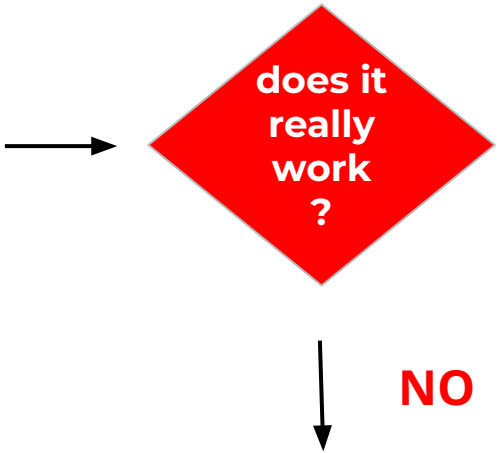
Visual Prompting With Andrew Ng

prompt-based ML

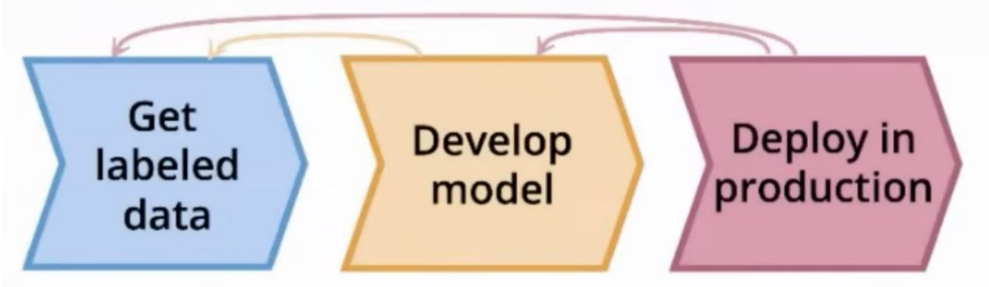


prompt-based ML

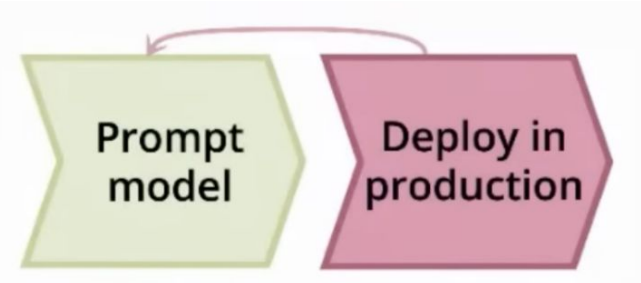


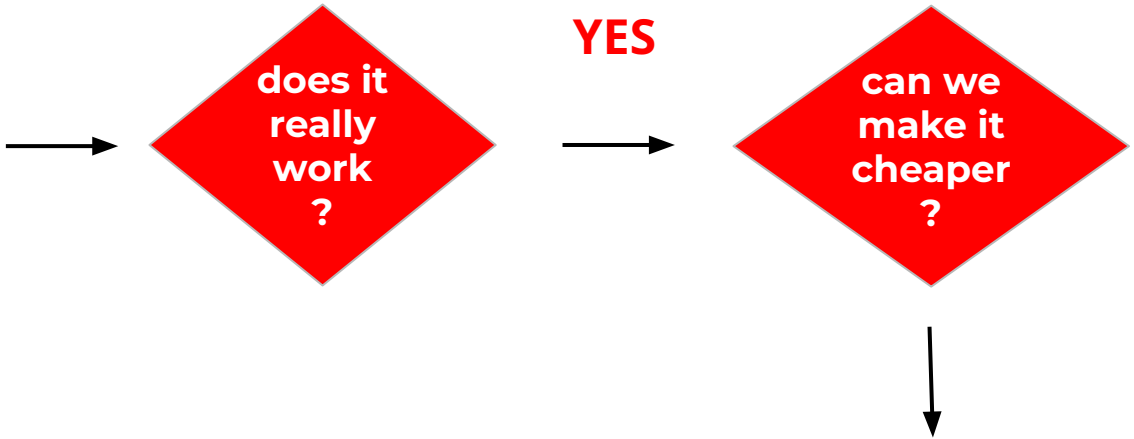


“traditional” ML/DL

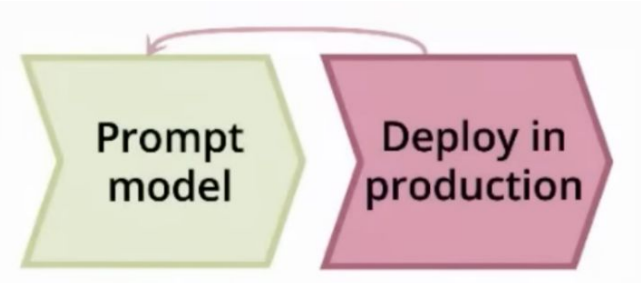


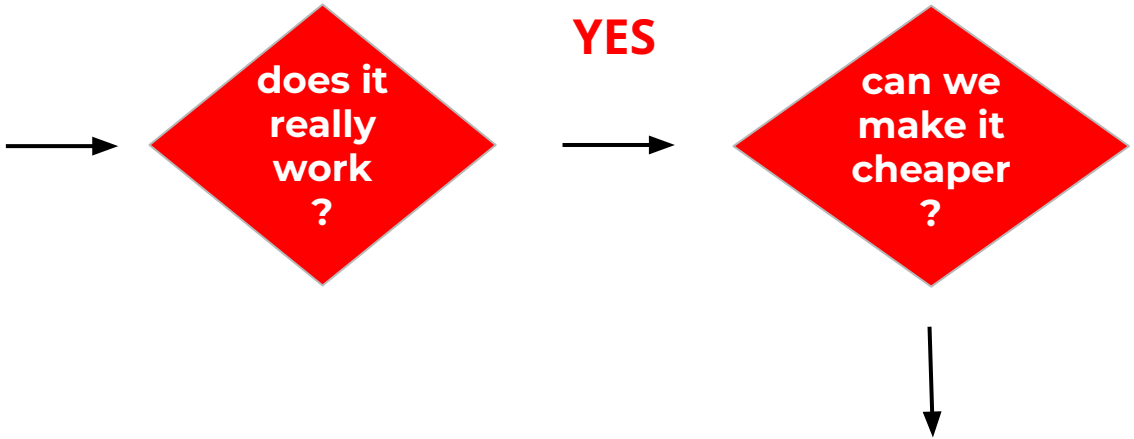
prompt-based ML



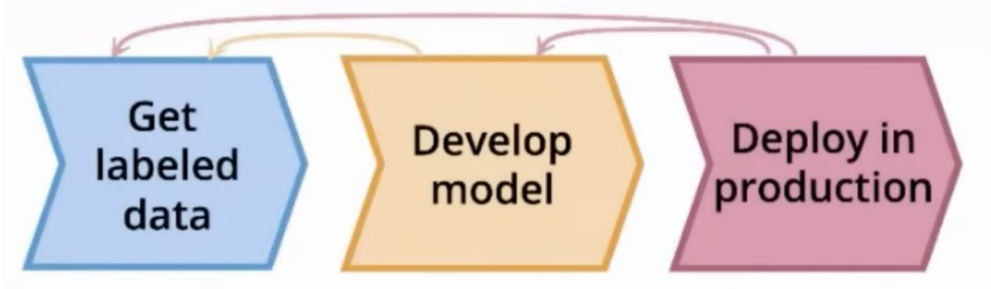


prompt-based ML

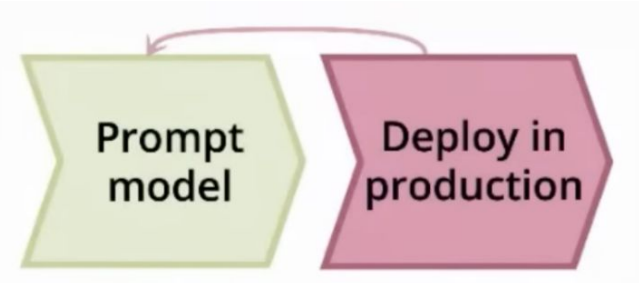


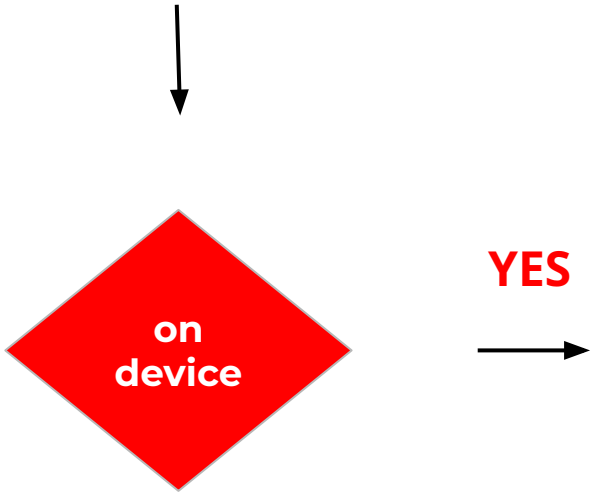


“traditional” ML/DL

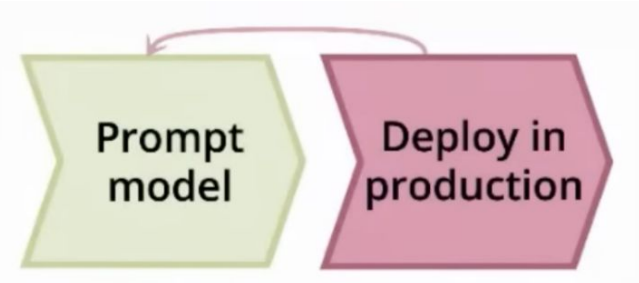


prompt-based ML



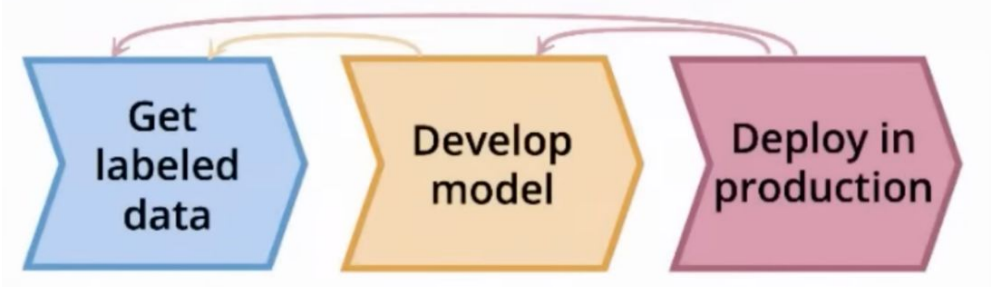


prompt-based ML

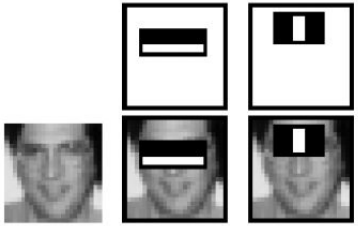


YES

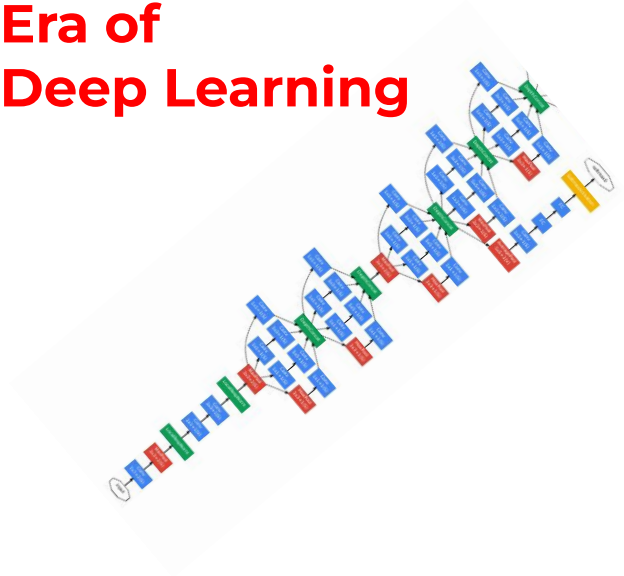
“traditional” ML/DL



Era of Human-Crafter Features



Era of Deep Learning

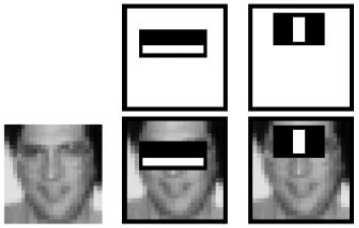


Era of LLMs

2012
AlexNet

2022
ChatGPT

Era of Human-Crafter Features



Era of Deep Learning

Era of LLMs

2012
AlexNet

2022
ChatGPT

APIs multimodal LLMs

Local multimodal LLMs

CV foundational models

“traditional” CV DL

“classical” CV

math

APIs multimodal LLMs

Local multimodal LLMs

CV DL foundational models

CV DL fundamentals

“classical” CV

math

APIs

Gemini

OpenAI

Claude

Grok

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

APIs + Fine-tuning

Image fine-tuning

Introducing vision to the fine-tuning API | OpenAI

APIs multimodal LLMs

Local multimodal LLMs

DL foundational models

CV DL fundamentals

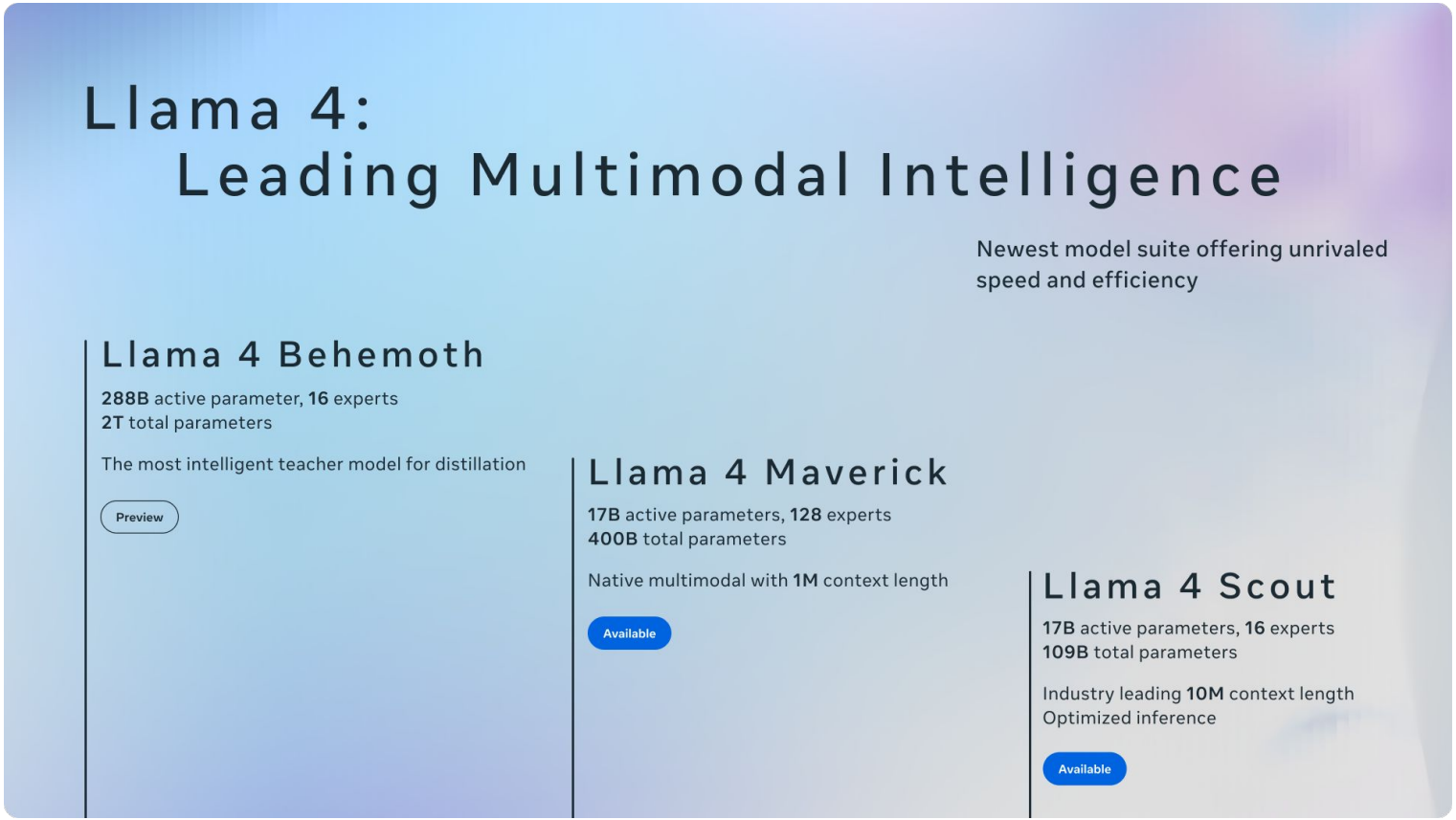
“classical” CV

math

Gemma

Gemma explained: What's new in Gemma 3 - Google Developers Blog
2310.07707

LLAMA / DeepSeek / ...



The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation

APIs multimodal LLMs

Local multimodal LLMs

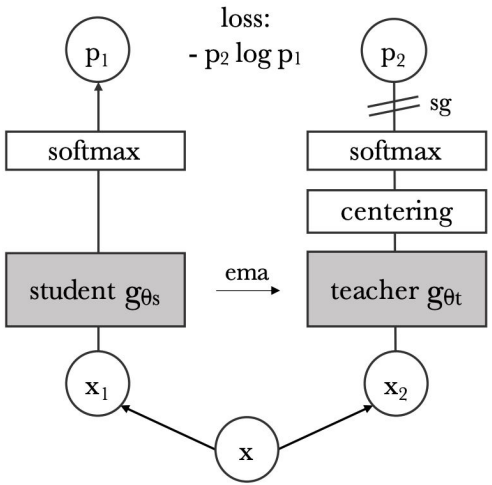
CV foundational models

CV DL fundamentals

“classical” CV

math

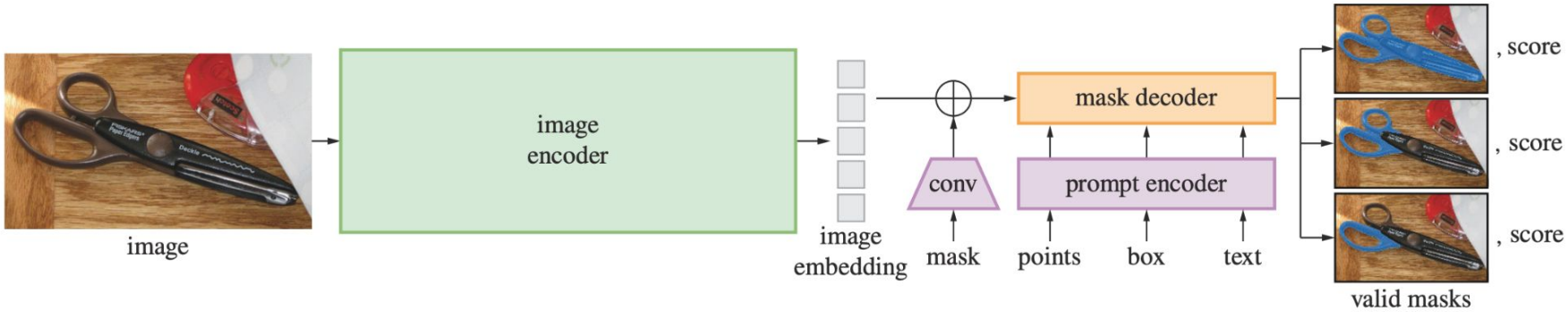
DINO (Self-**Di**stillation with **no** labels)



2104.14294, 2304.07193, 2508.10104

DINOv3

SAM (**S**egment **A**nything)



2304.02643; 2408.00714

Introducing Meta Segment Anything Model 2 (SAM 2)

Segment Anything 3
Coming 2025

APIs multimodal LLMs

Local multimodal LLMs

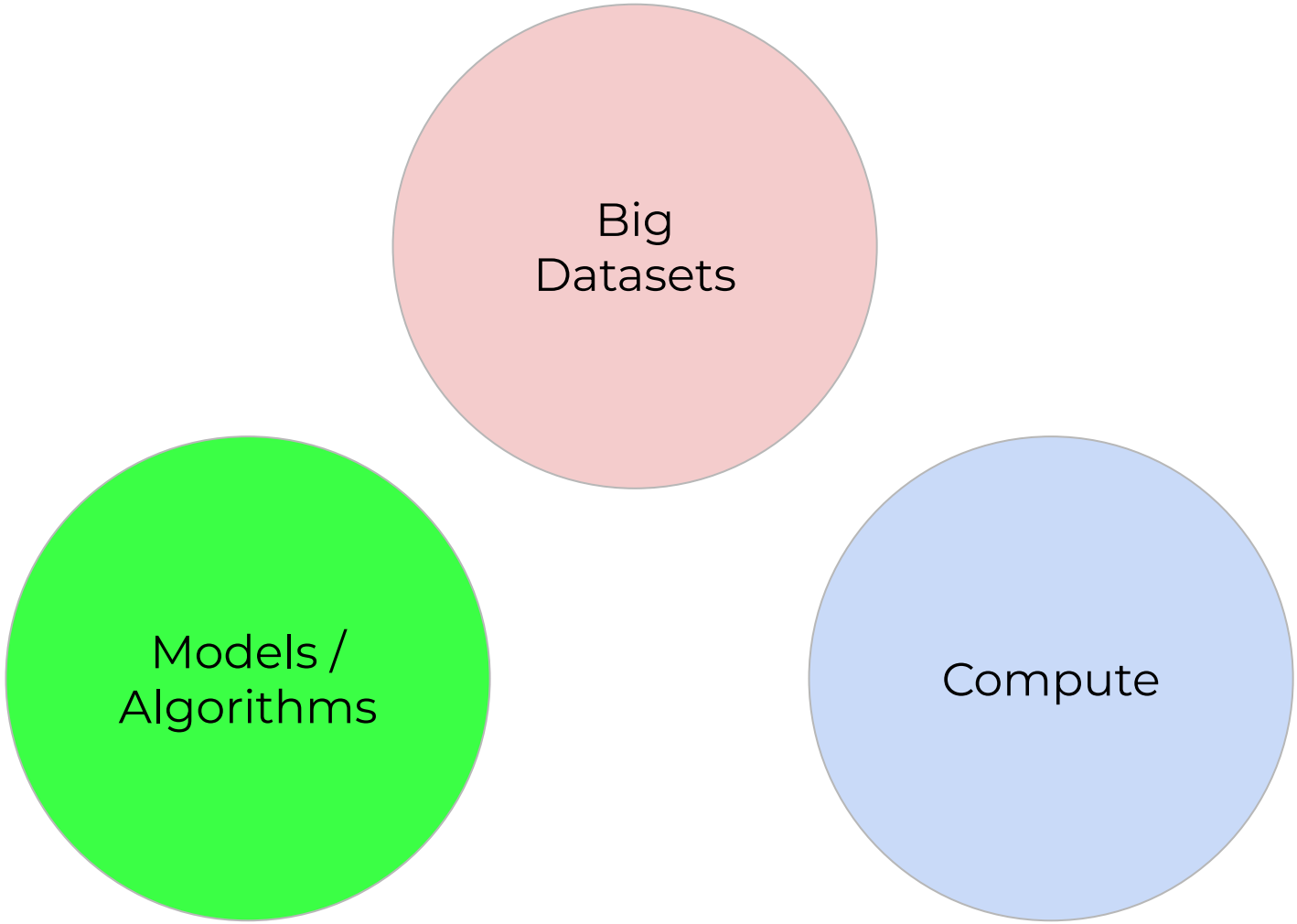
CV DL foundational models

CV DL fundamentals

“classical” CV

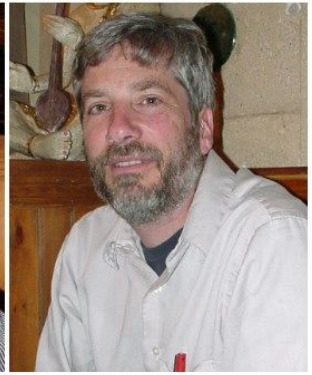
math

Intro



Intro

V. Vapnik
(creator of SVM)



L. Jackel
(proponent of NN)

Y. LeCun

1. Jackel bets (one fancy dinner) that by March 14, 2000, people will understand
quantitatively why big neural nets working on large databases are not so bad. (Understanding means that there will be clear conditions and bounds)

Vapnik bets (one fancy dinner) that Jackel is wrong.

But .. If Vapnik figures out the bounds and conditions, Vapnik still wins the bet.

****

2. Vapnik bets (one fancy dinner) that by March 14, 2005, no one in his right mind will use neural nets that are essentially like those used in 1995.

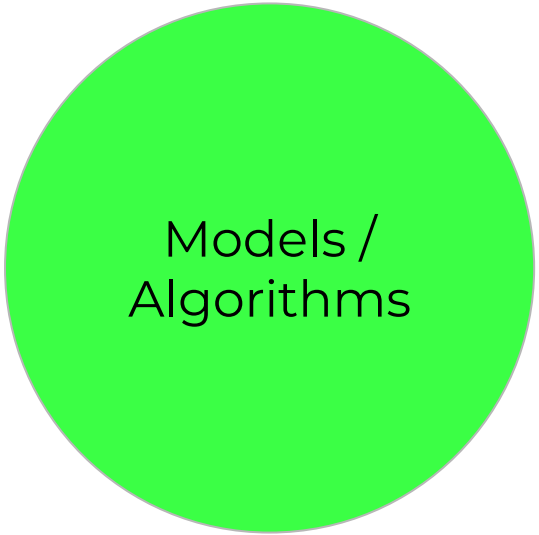
Jackel bets (one fancy dinner) that Vapnik is wrong

3/14/95
V. Vapnik

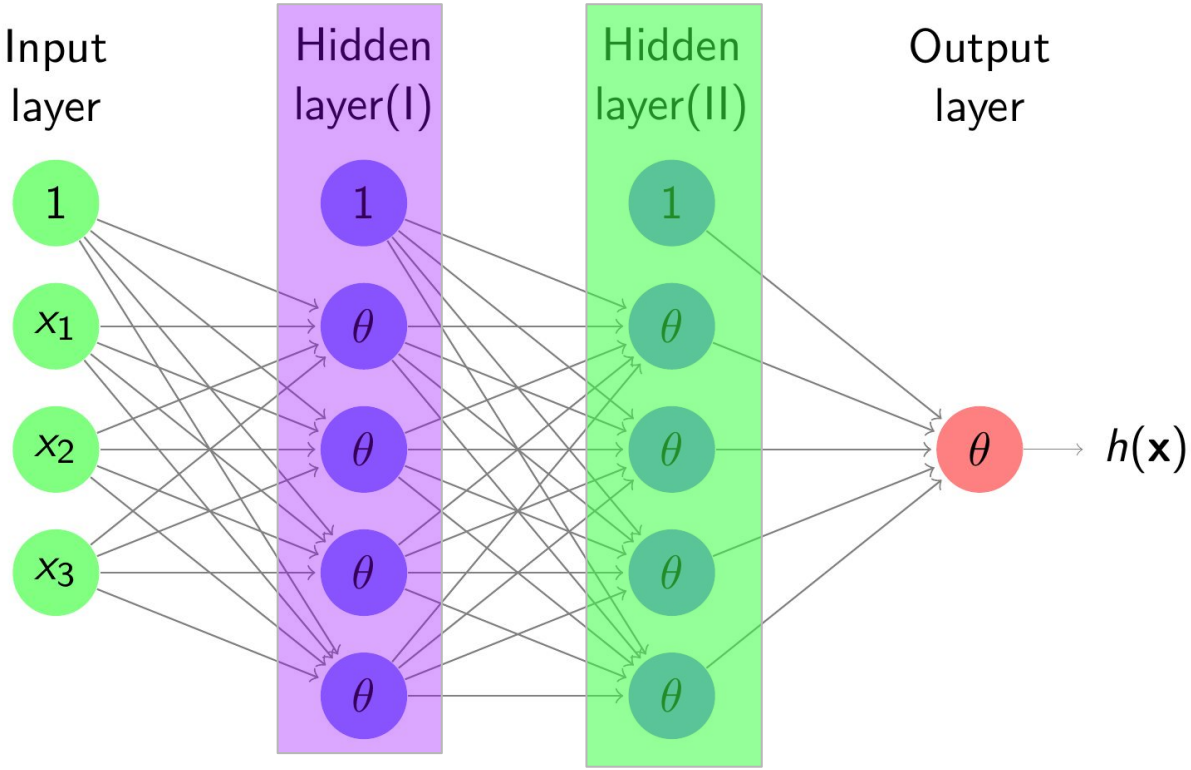
3/14//95
L. Jackel

3/14/95
Witnessed by Y. LeCun

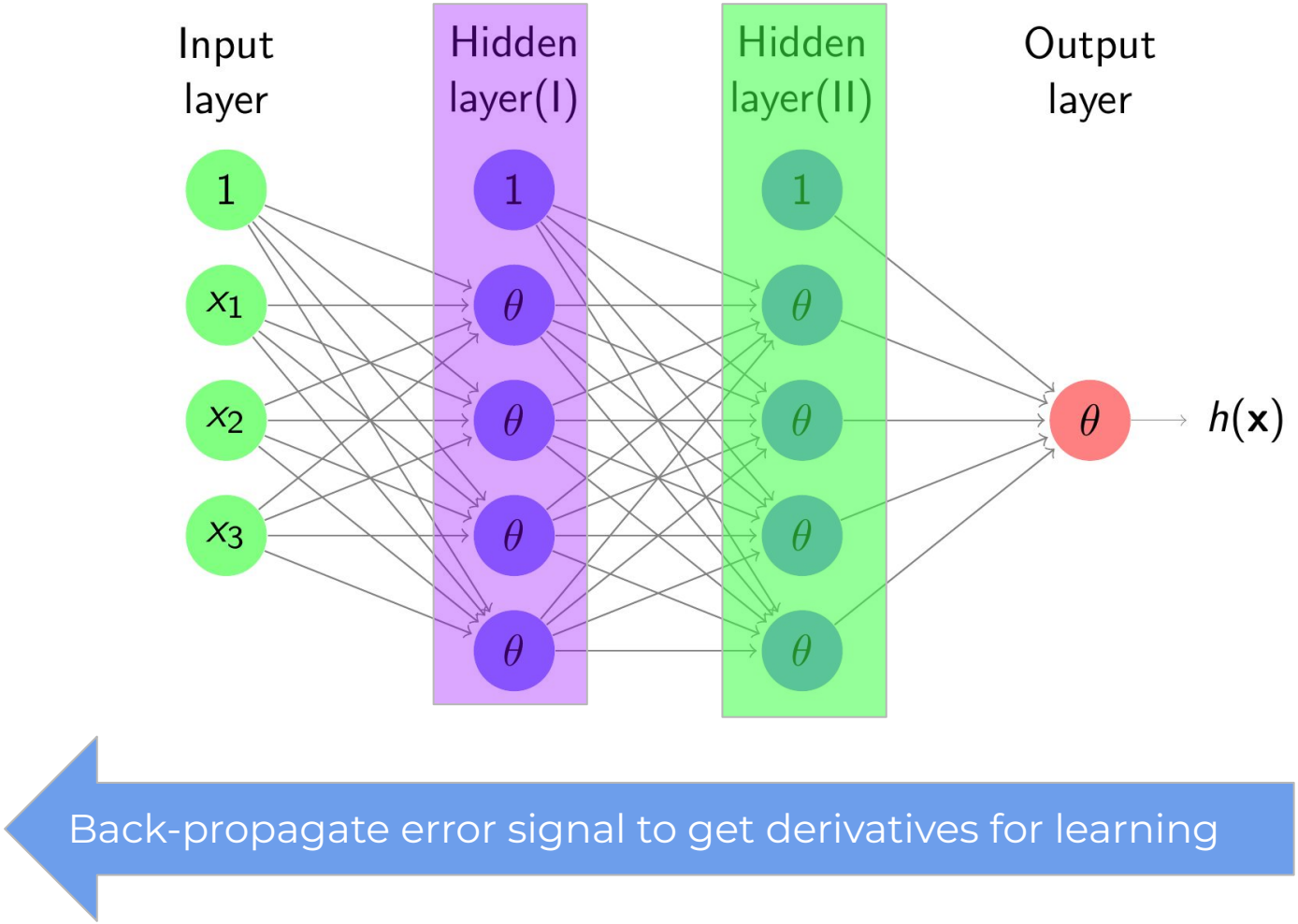
Intro



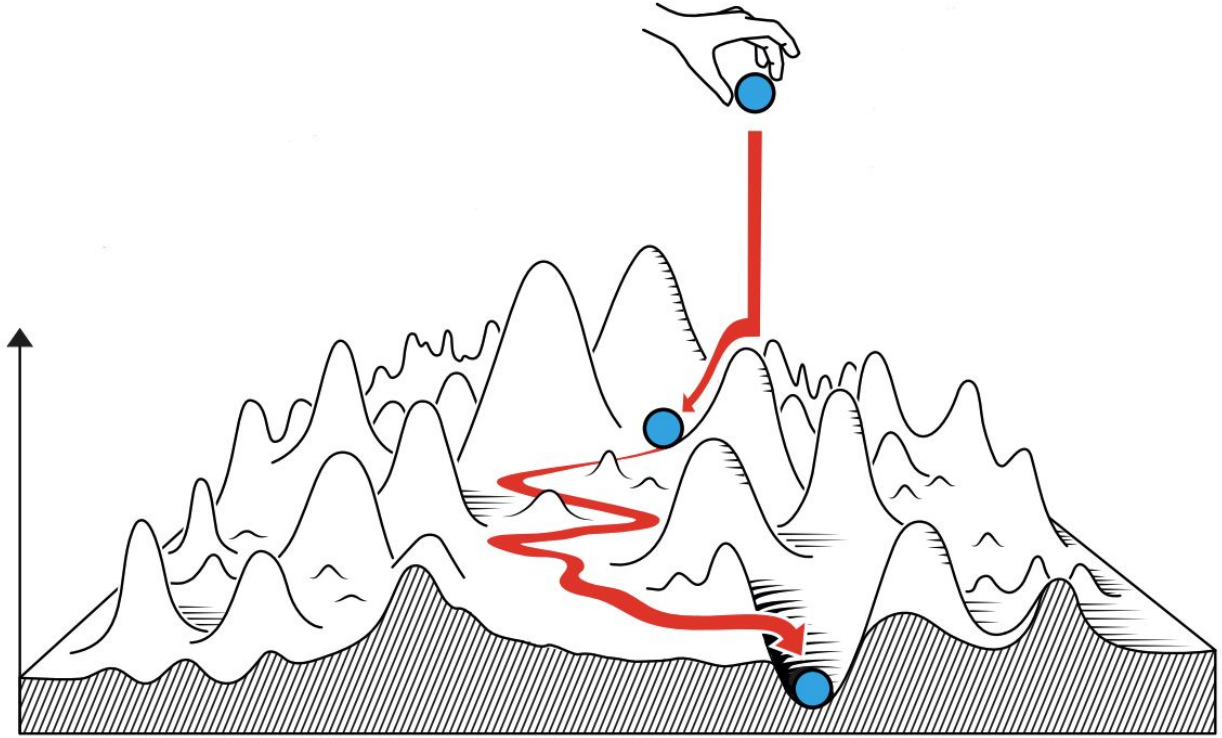
- Neural Network
- Number of layers => Deep
- Important specific components (convolutions, attention, ...)

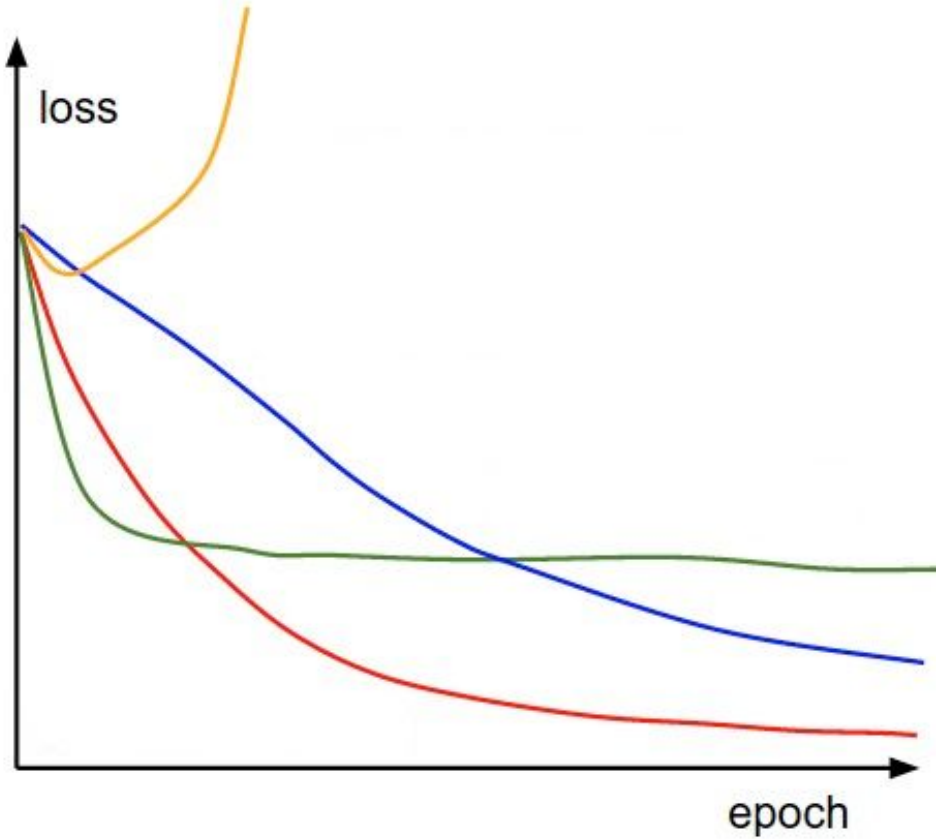


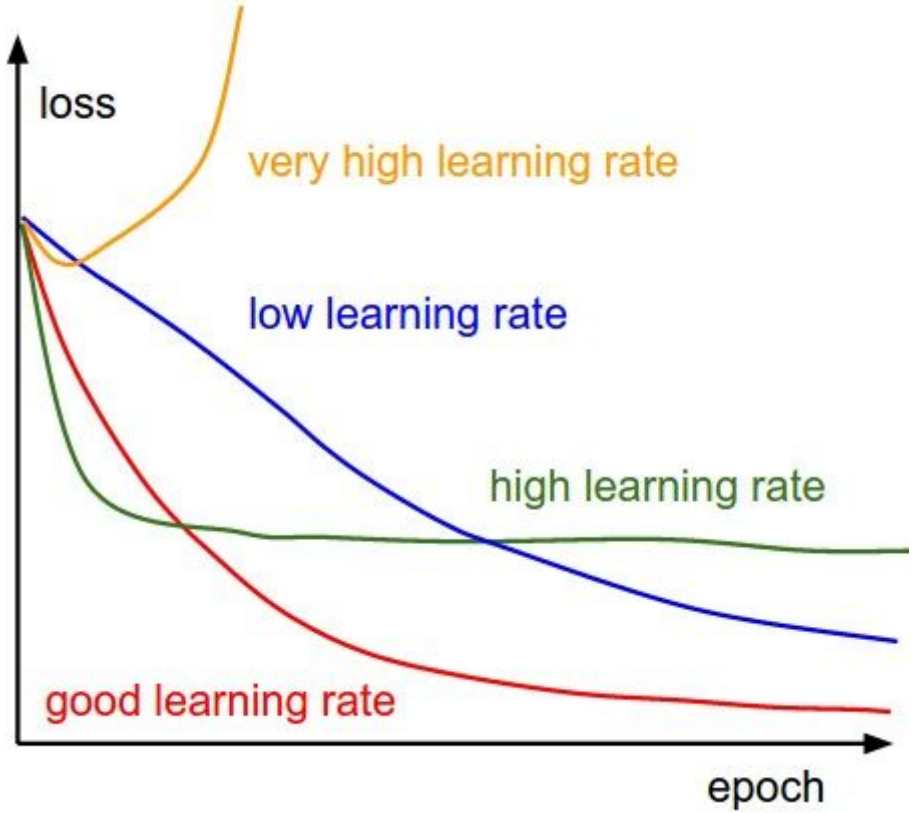
$$\left\{ f(\boldsymbol{x};\boldsymbol{\theta}) = \mathbf{W}_L \boldsymbol{\sigma}_L(\mathbf{W}_{L-1} \cdots \boldsymbol{\sigma}_2(\mathbf{W}_2 \boldsymbol{\sigma}_1(\mathbf{W}_1 x))) \;\middle|\; \boldsymbol{\theta} = \{\mathbf{W}_1, \ldots, \mathbf{W}_L\} \right\}$$


Input layer
Hidden layer(I)
Hidden layer(II)
Output layer
1
x_1
x_2
x_3
1
1
θ
θ
θ
θ
θ
θ
θ
θ
θ
$h(\mathbf{x})$
Back-propagate error signal to get derivatives for learning

Optimization

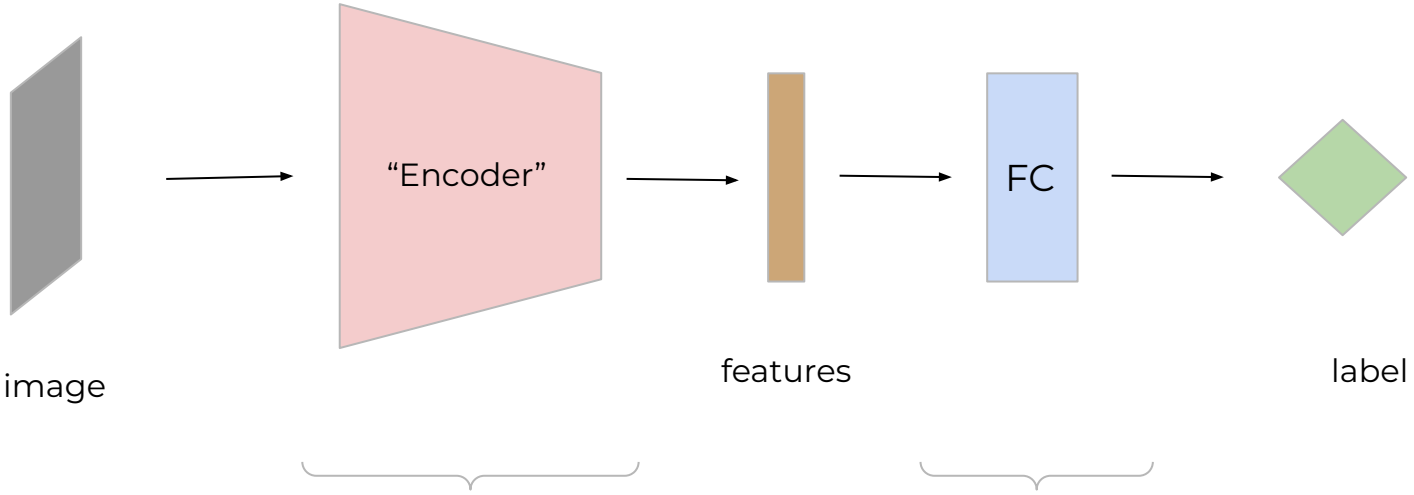



loss
epoch

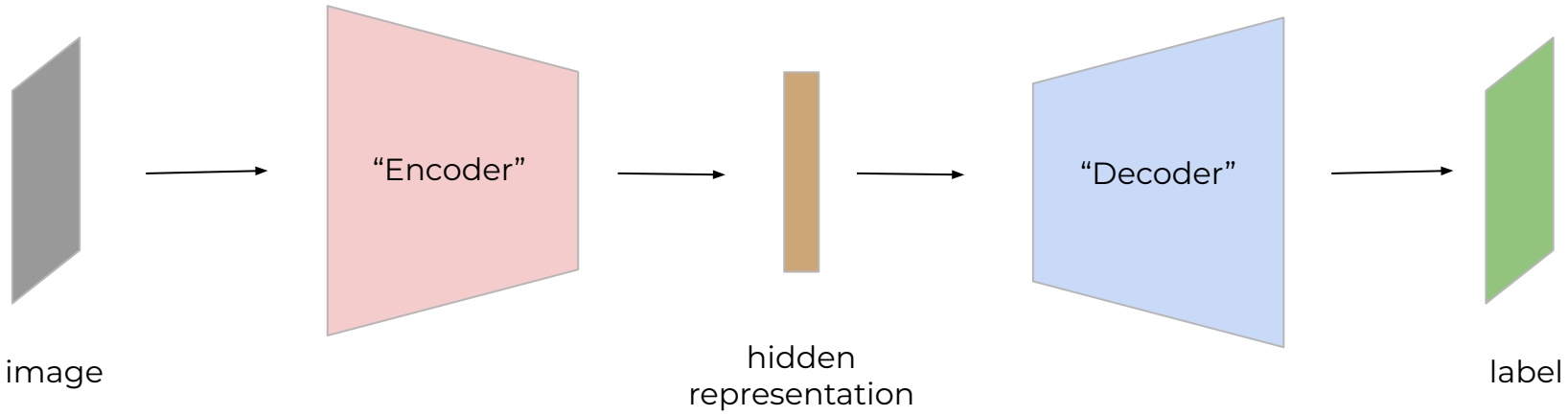

loss
very high learning rate
low learning rate
high learning rate
good learning rate
epoch

Typical CV architecture

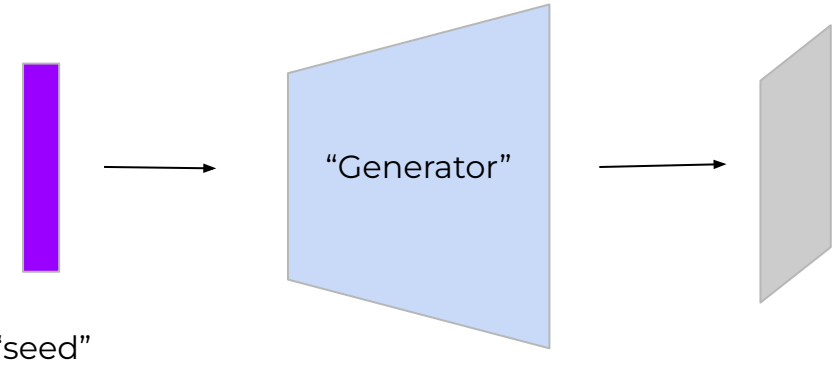
Typical CV architecture [classification]

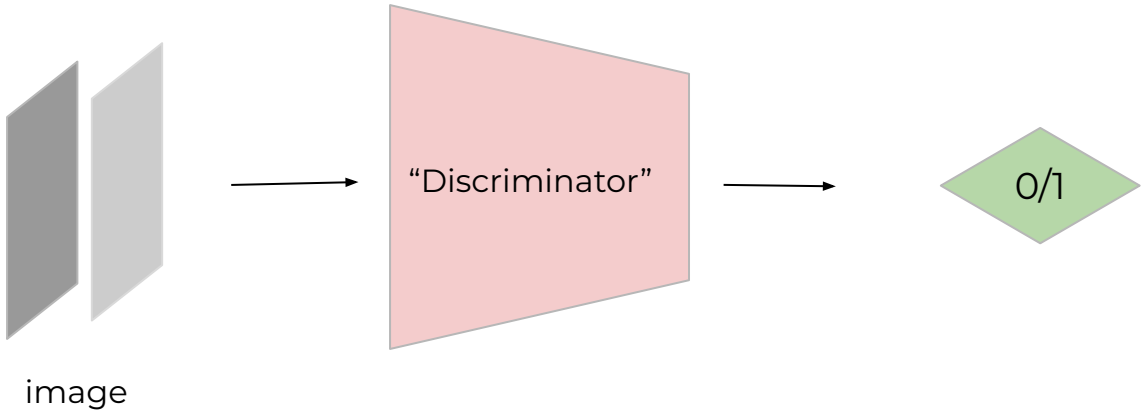


Typical CV architecture [segmentation]

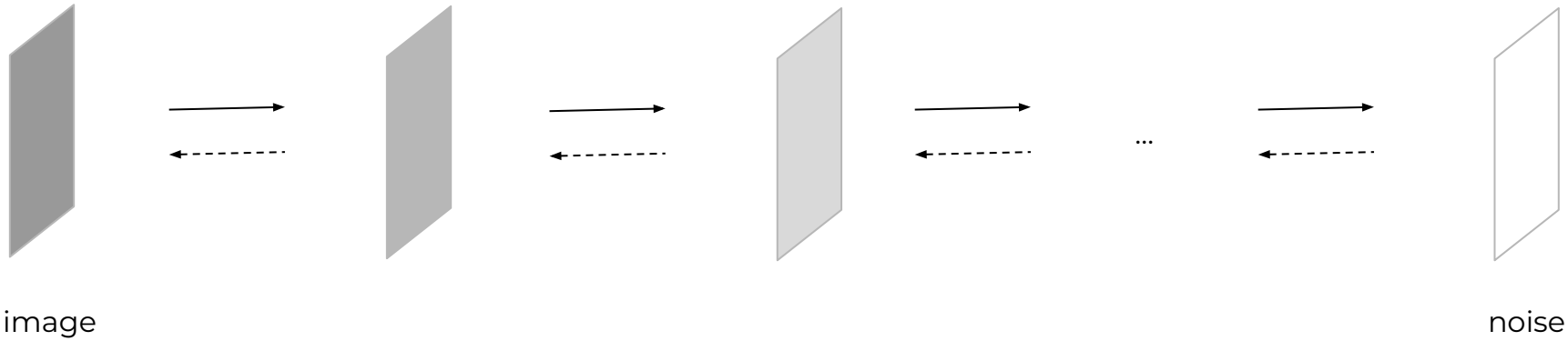


Typical CV architecture [Generation, GAN]

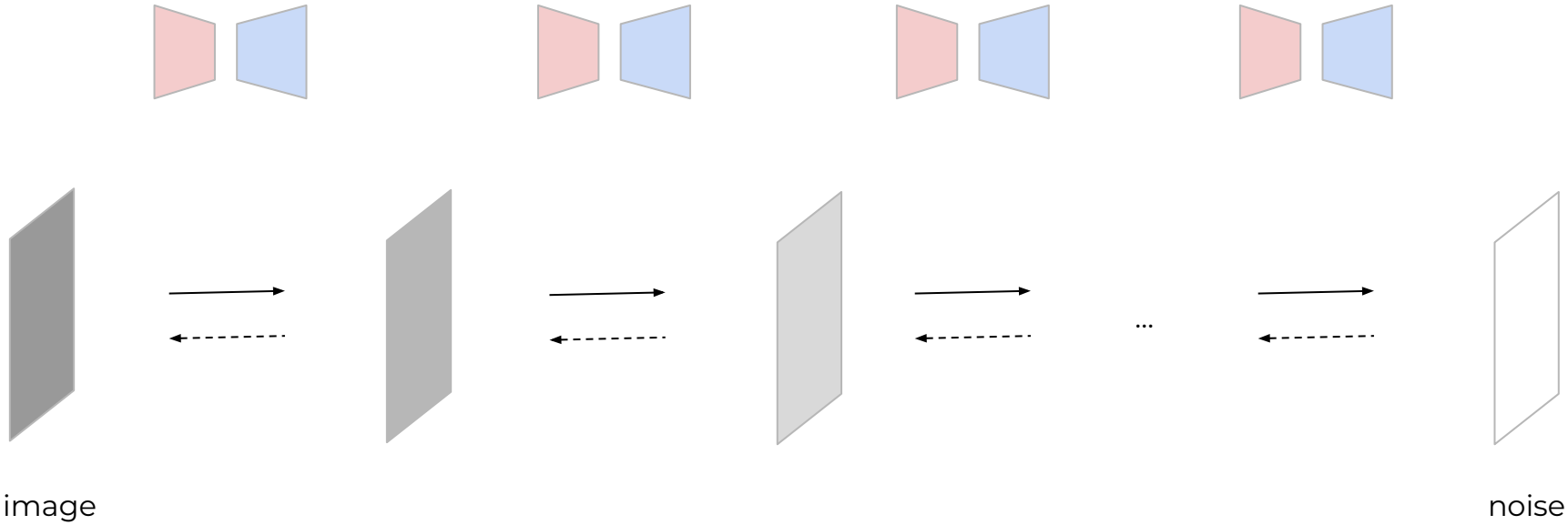


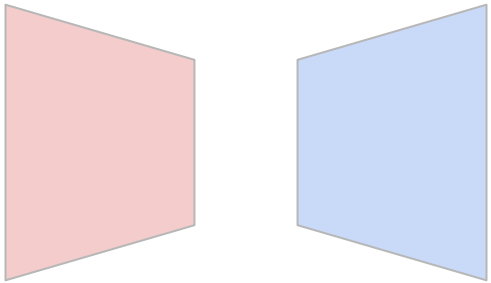


Typical CV architecture [Generation, Diffusion]



Typical CV architecture

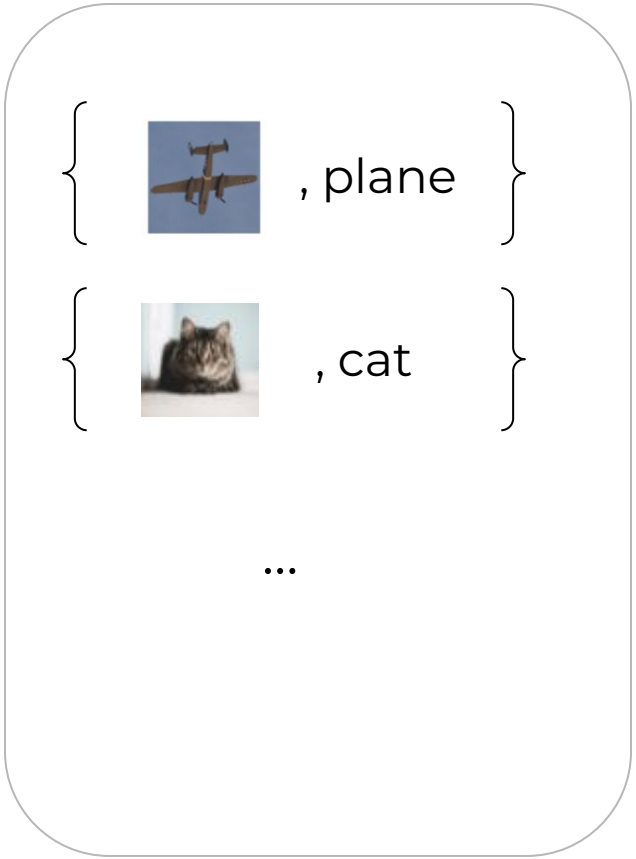




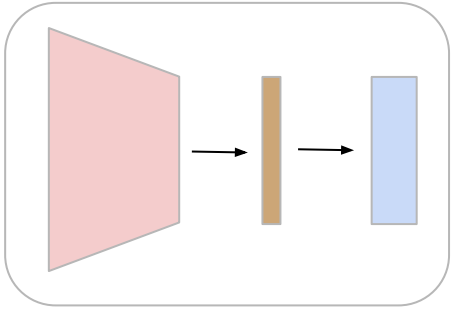
- convolutions
  - ...
- attention
  - ...
- ...

Training strategies

Supervised way



$F(\mathbf{x_n})$



$e(F(\mathbf{x_n}), y_n)$

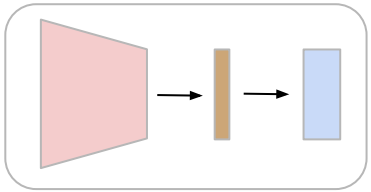
$e(F(\mathbf{x_n}), y_n)$

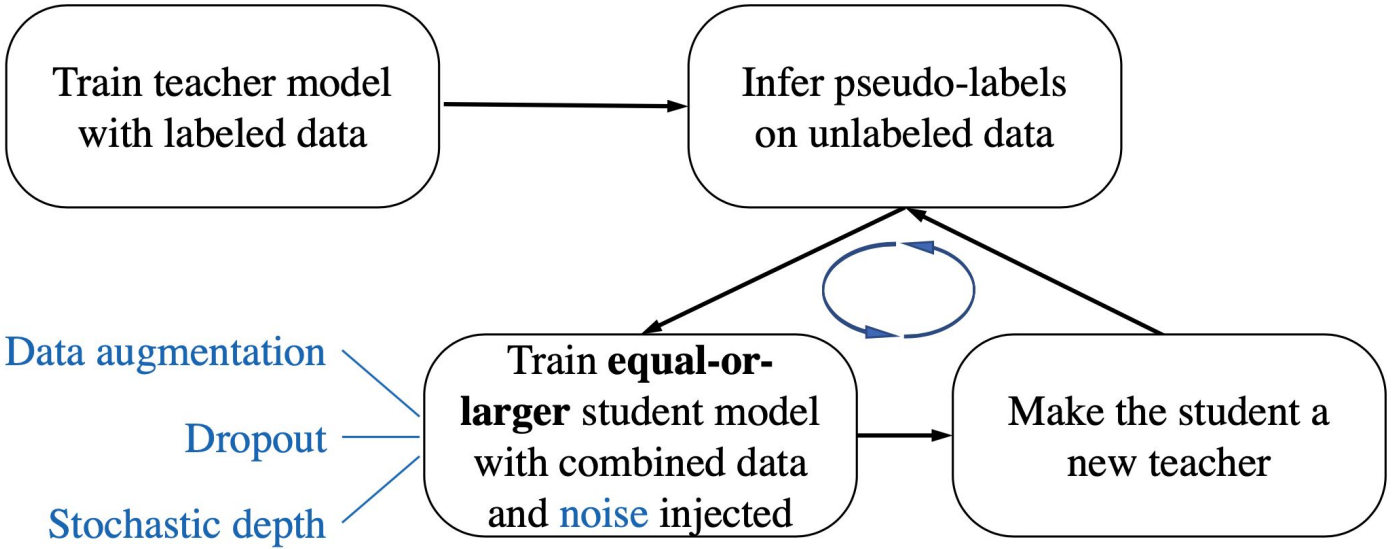
...

$$L_{train}(\omega) = \frac{1}{N}\sum_{n=1}^{N} e(F(\mathbf{x_n}), y_n)$$

require images paired with high-quality metadata

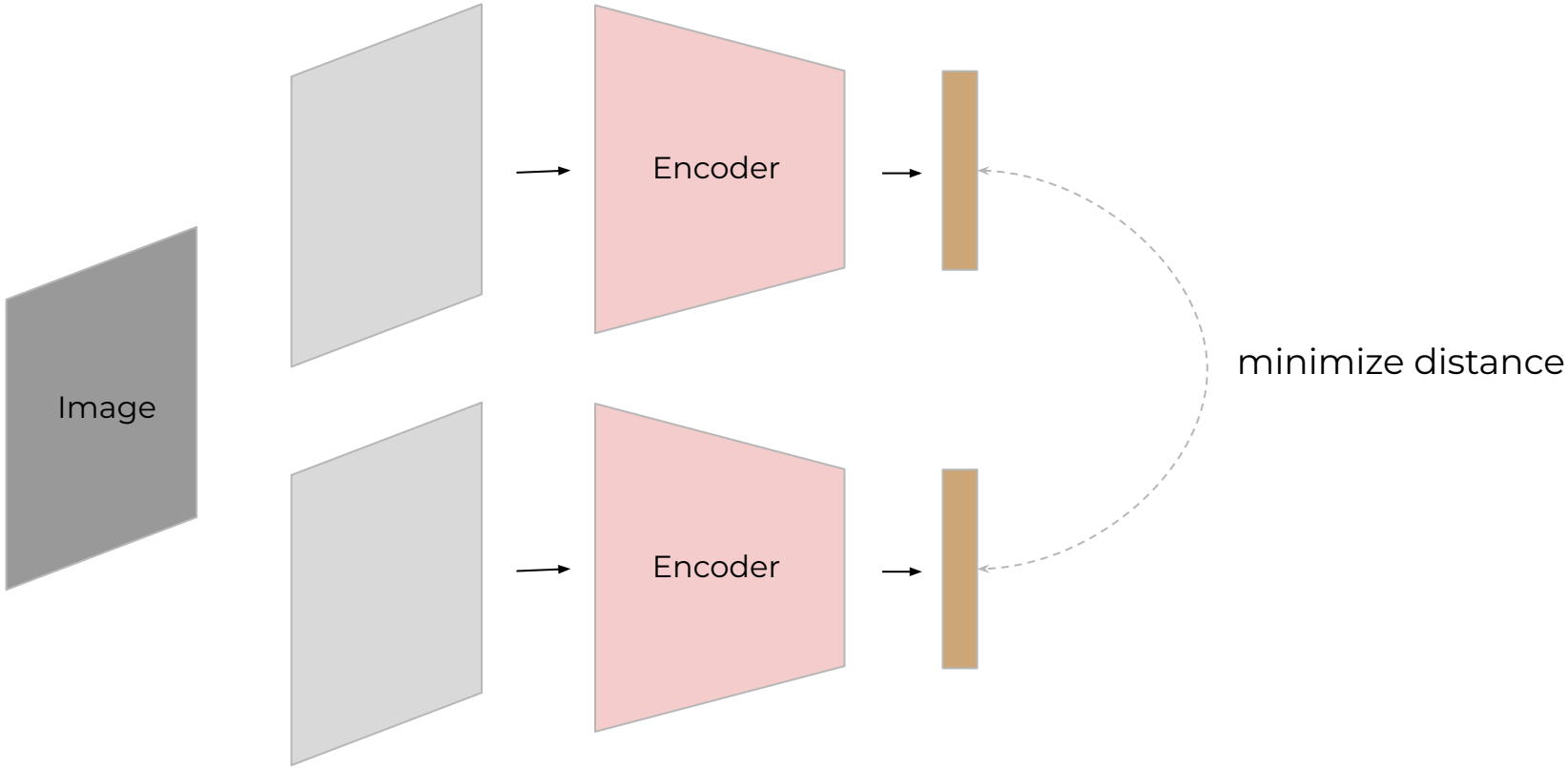
Weakly-supervised way





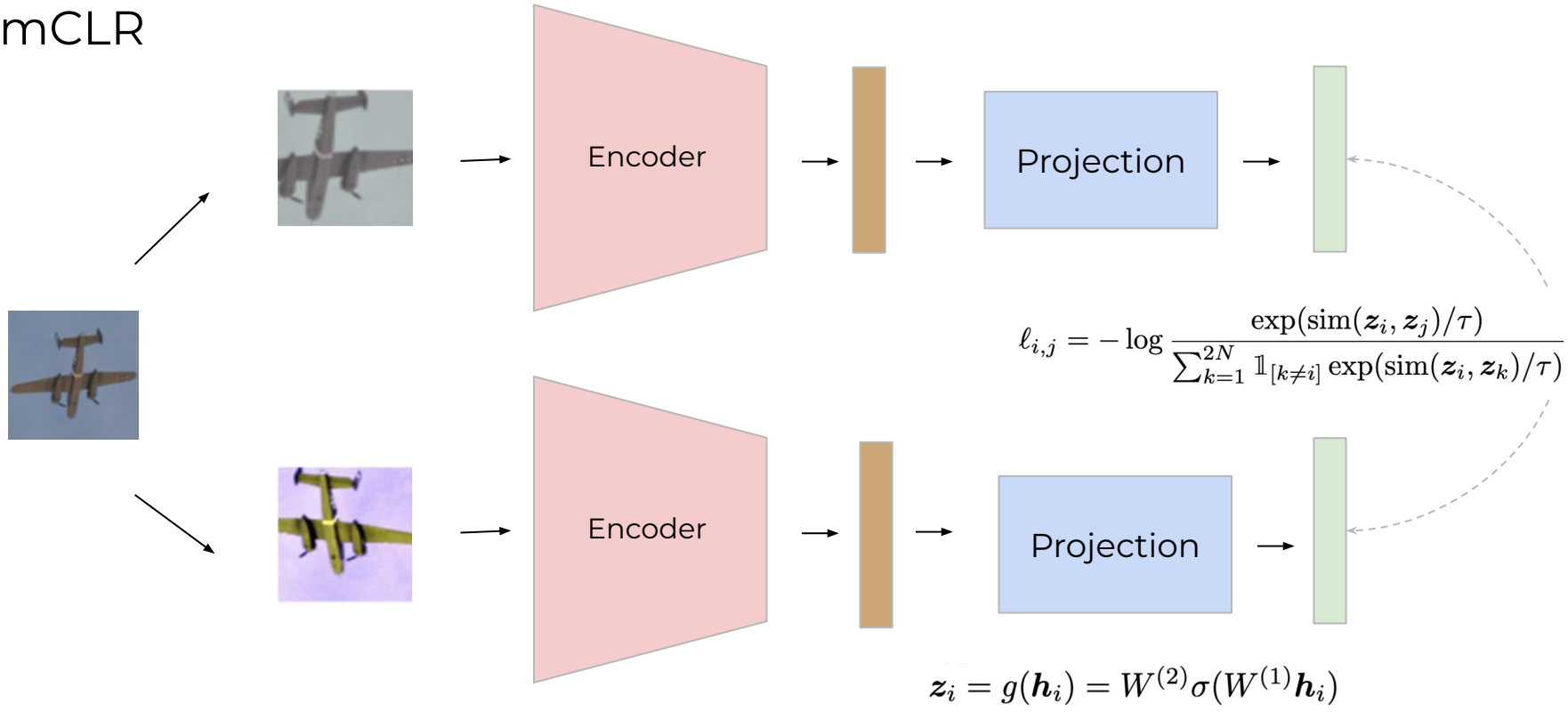
1911.04252

Self-supervised way



Self-supervised way

SimCLR



$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

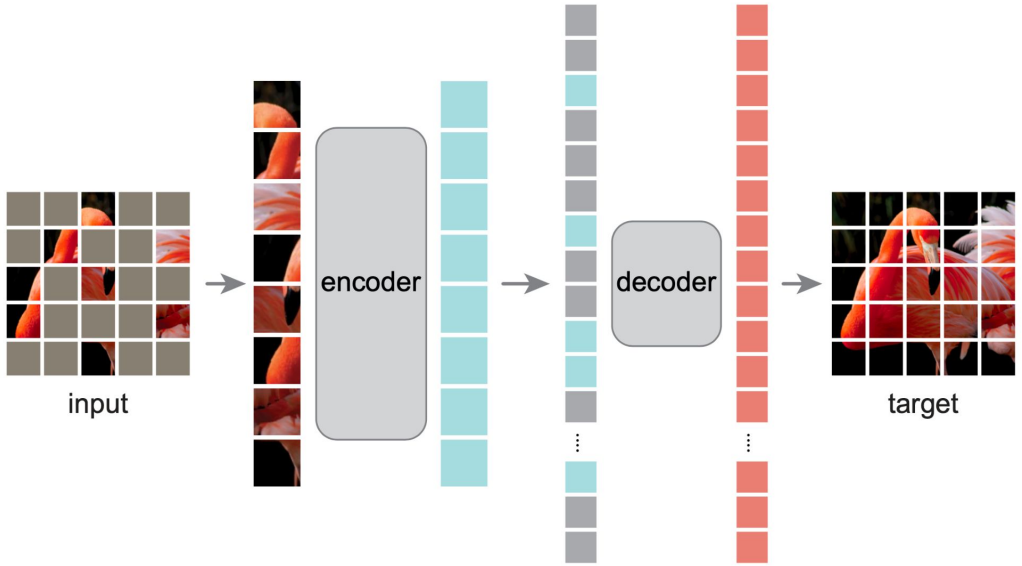
$$\boldsymbol{z}_i = g(\boldsymbol{h}_i) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h}_i)$$

$$\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i) = \text{ResNet}(\tilde{\boldsymbol{x}}_i)$$

2002.05709

Self-supervised way

MAE



2111.06377

Self-supervised way

MAE

2111.06377

Self-supervised way

- do not require annotated data
- robust to input distribution shifts, provide strong global and local features, and generate rich embeddings that facilitate physical scene understanding.
- produce versatile and robust generalist features since SSL models are not trained for any specific downstream task
- requiring no human intervention, is well-suited for lifelong learning amid the growing volume of web data

Self-supervised way

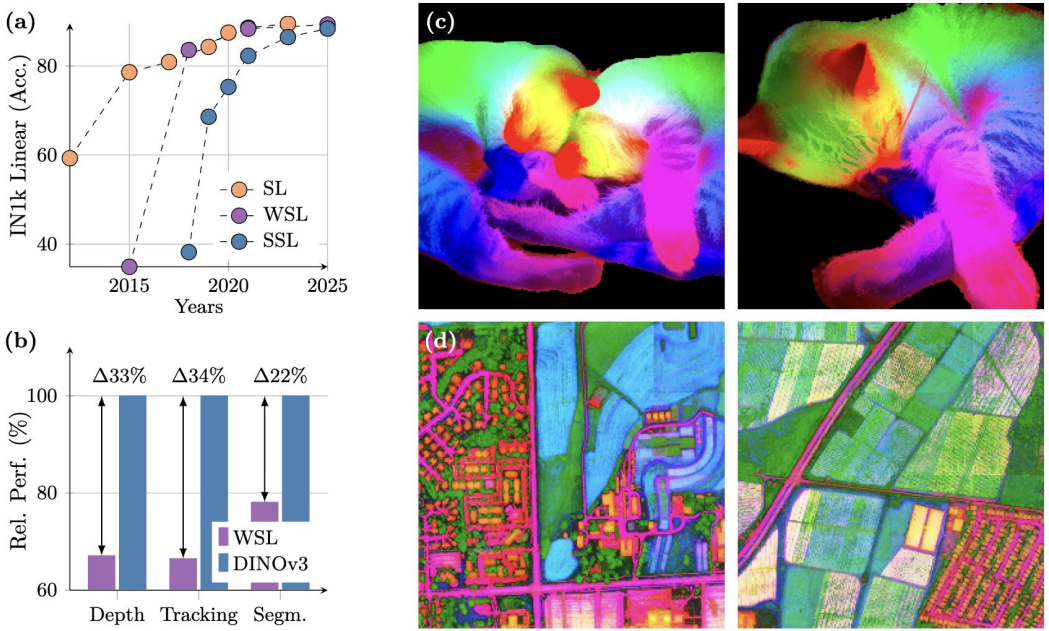


Figure 1: (a) Evolution of linear probing results on ImageNet1k (IN1k) over the years, comparing fully- (SL), weakly- (WSL) and self-supervised learning (SSL) methods. Despite coming into the picture later, SSL has quickly progressed and now reached the Imagenet accuracy plateau of recent years. On the other hand, we demonstrate that SSL offers the unique promise of high-quality dense features. With DINOv3, we markedly improve over weakly-supervised models on dense tasks, as shown by the relative performance of the best-in-class WSL models to DINOv3 (b). We also produce PCA maps of features obtained from high resolution images with DINOv3 trained on natural (c) and aerial images (d).

2508.10104

APIs multimodal LLMs

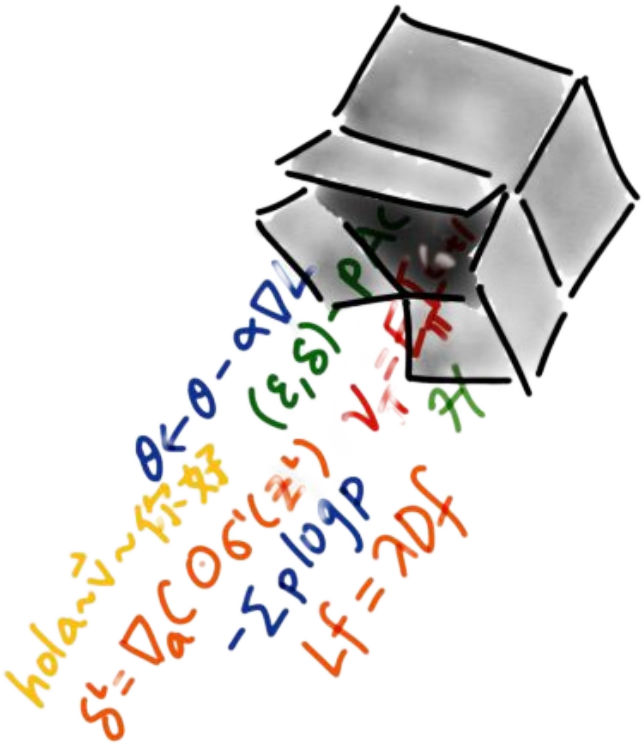
Local multimodal LLMs

CV DL foundational models

CV DL fundamentals

“classical” CV

math

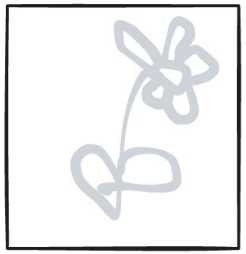


Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong
"Mathematics for Machine Learning"

https://mml-book.github.io/

w

state of the world



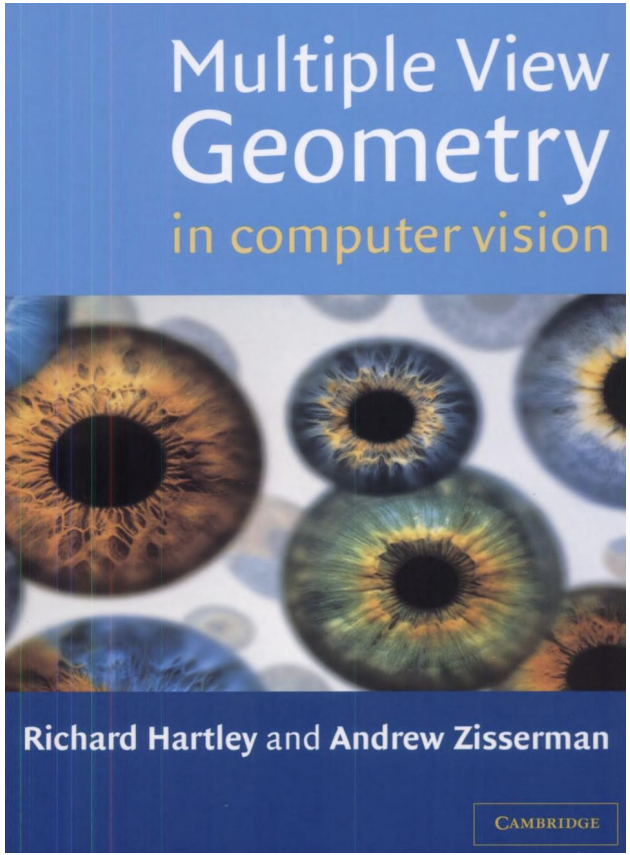
x

measurements

- **discriminative** model Pr(**w**|**x**)
- **generative** model Pr(**x**|**w**)
- Pr(**w**|**x**) = Pr(**x**|**w**) * Pr(**w**) / ∫[Pr(x | w) * Pr(w)] dw

Simon J.D. Prince
"Computer Vision: Models, Learning, and Inference"

https://udlbook.github.io/cvbook/
https://udlbook.github.io/udlbook/



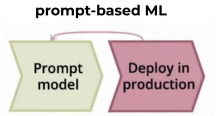
Richard Hartley and Andrew Zisserman
"Multiple View Geometry in Computer Vision"

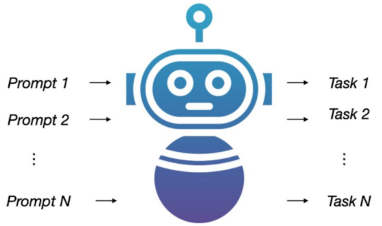
CS231A: Computer Vision, From 3D Perception to 3D Reconstruction and beyond

https://www.google.com.ua/books/edition/Multiple_View_Geometry_in_Computer_Visio/

Summary

APIs multimodal LLMs



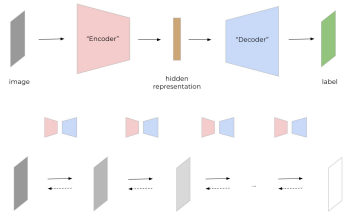


Local multimodal LLMs

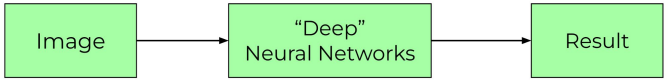
GEMMA, LLAMA, ...

CV foundational models

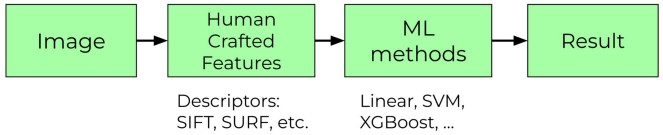
DINO, SAM, ...



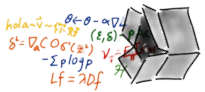
CV DL fundamentals



“classical” CV



math



Overview of the module

Lecture I Intro, big picture

Lecture II Essential architectures (CNN)

Lecture III Essential architectures (Transformers)

Lecture IV CV Foundational models in depth (DINO, SAM)

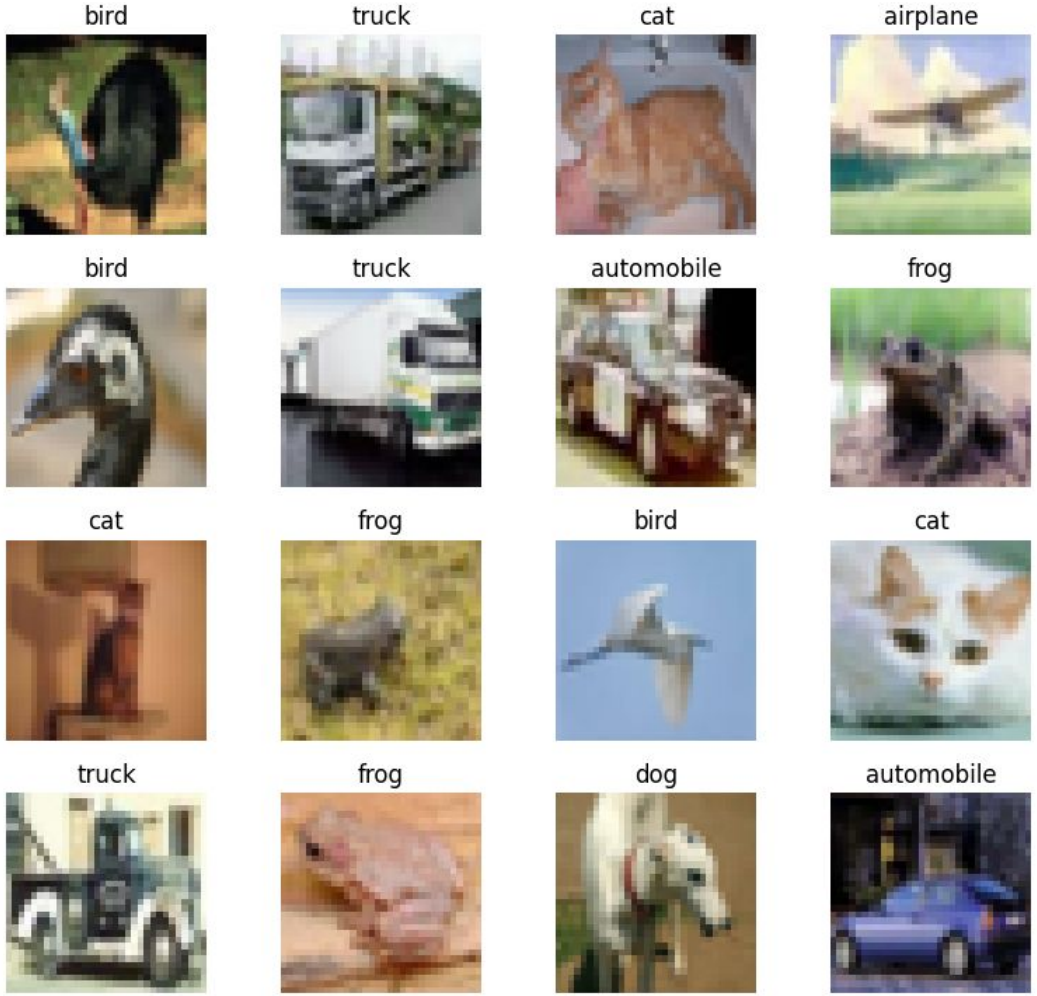
Lecture V Multimodal Foundational models (GEMMA, APIs)

Lecture VI Object Detection

Time to practice!

notebook

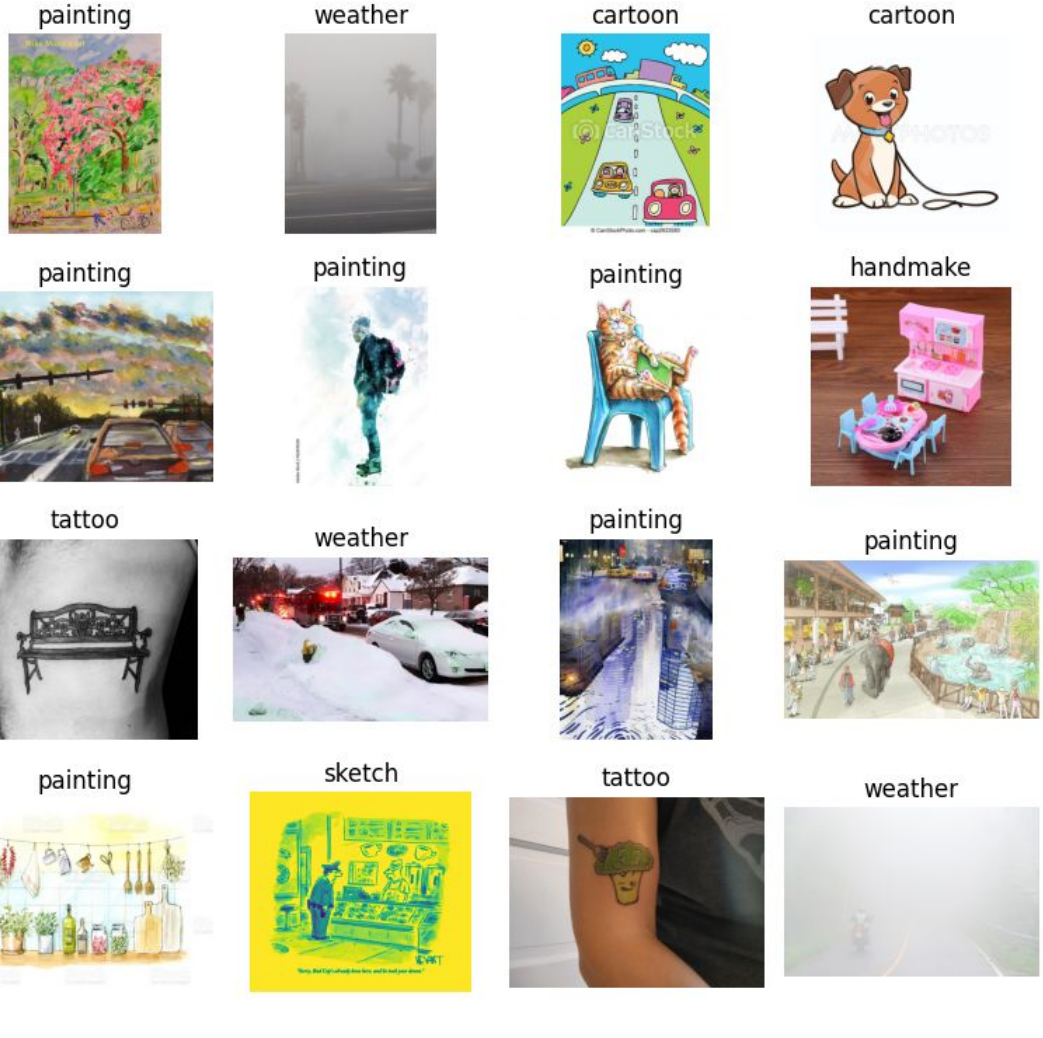
CIFAR10


bird
truck
cat
airplane
bird
truck
automobile
frog
cat
frog
bird
cat
truck
frog
dog
automobile

DTD


freckled
banded
bubbly
swirly
bubbly
striped
marbled
interlaced
meshed
swirly
frilly
marbled
smeared
stained
grid
meshed

COCO-O


painting
weather
cartoon
cartoon
painting
painting
painting
handmake
tattoo
weather
painting
painting
painting
sketch
tattoo
weather