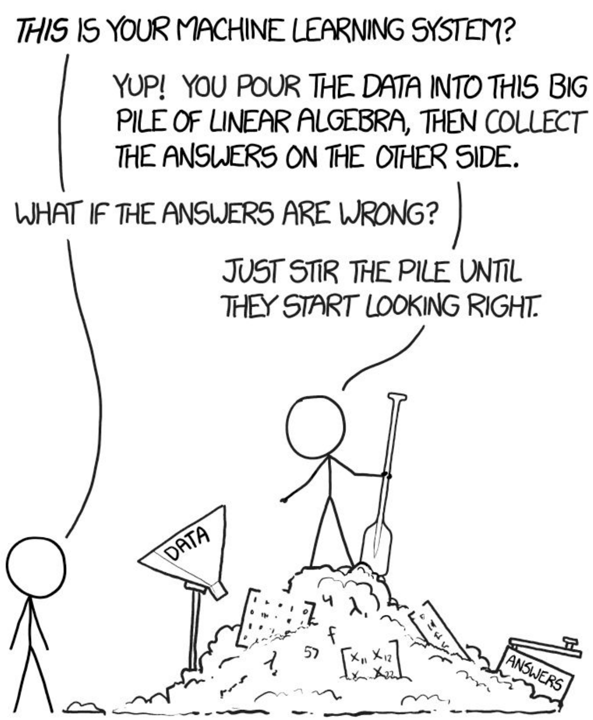
A data-driven graph framework for geometric understanding of deep learning

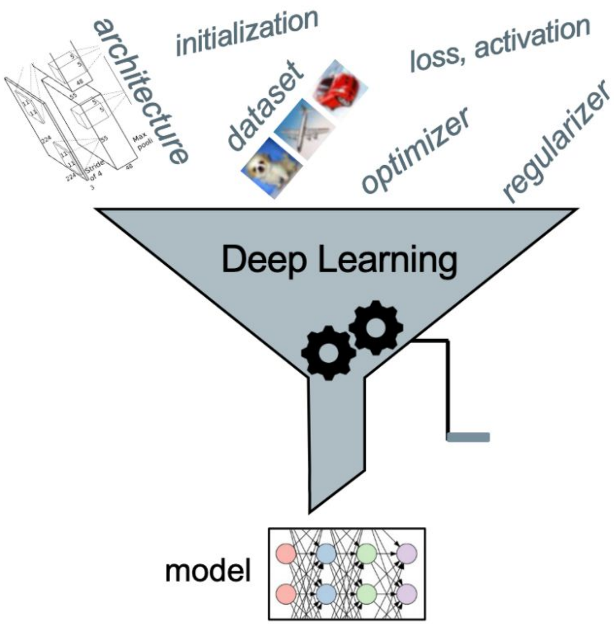
Sarath Shekkizhar, Antonio Ortega
{shekkizh, aortega}@usc.edu



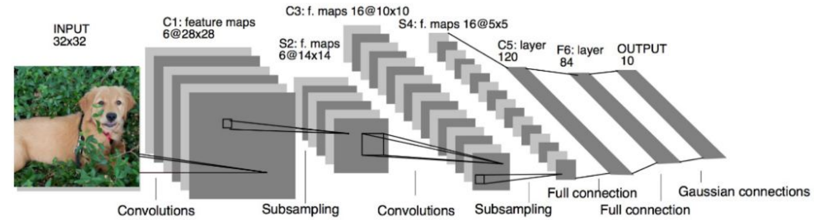
GSP Workshop
June, 2023

STAC Lab
USC Viterbi
School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

What is deep learning?

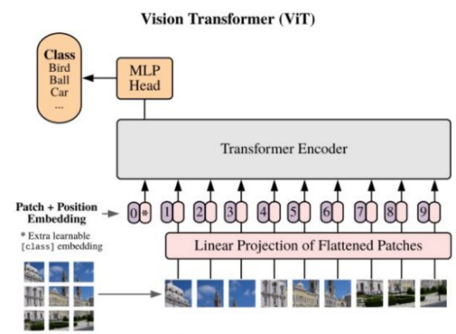


~1998-2020: ConvNets dominate vision



[LeCun et al 1998]

2020: *Transformers* (from NLP) dominate vision



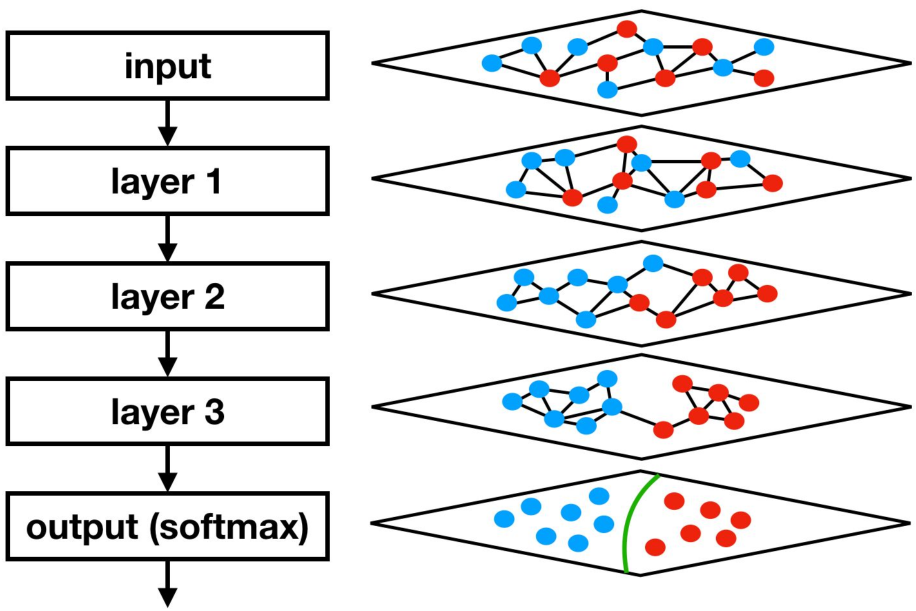
[Dosovitskiy et al 2020]

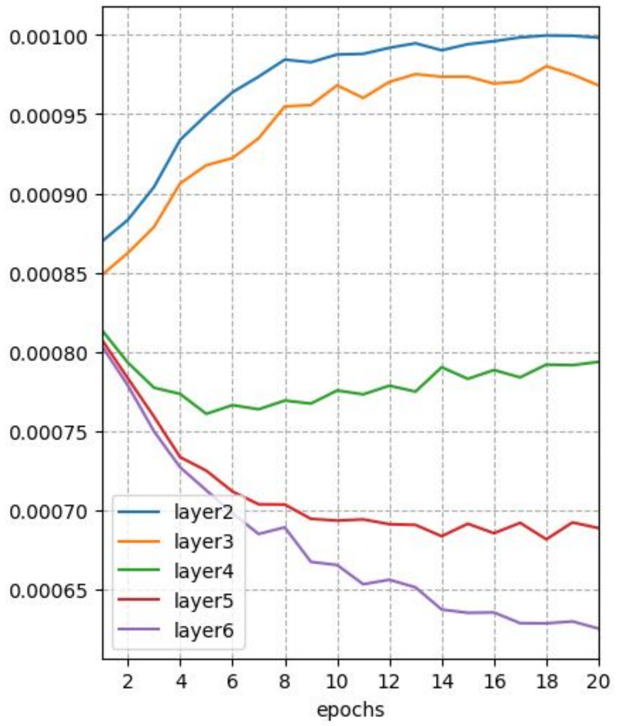
PC: Preetum Nakkiran

- ❑ Informal: Set of **modular** components combined/trained for a task
- ❑ Advances in models = new components
- ❑ Very successful in practice (often surprising which choices work)

Graph based view of deep learning

- **Some universal ideas:** **Hierarchy, Invariances**
- Current data-driven analysis of deep learning limited to
  - End-performance (**accuracy**) or to **single model**



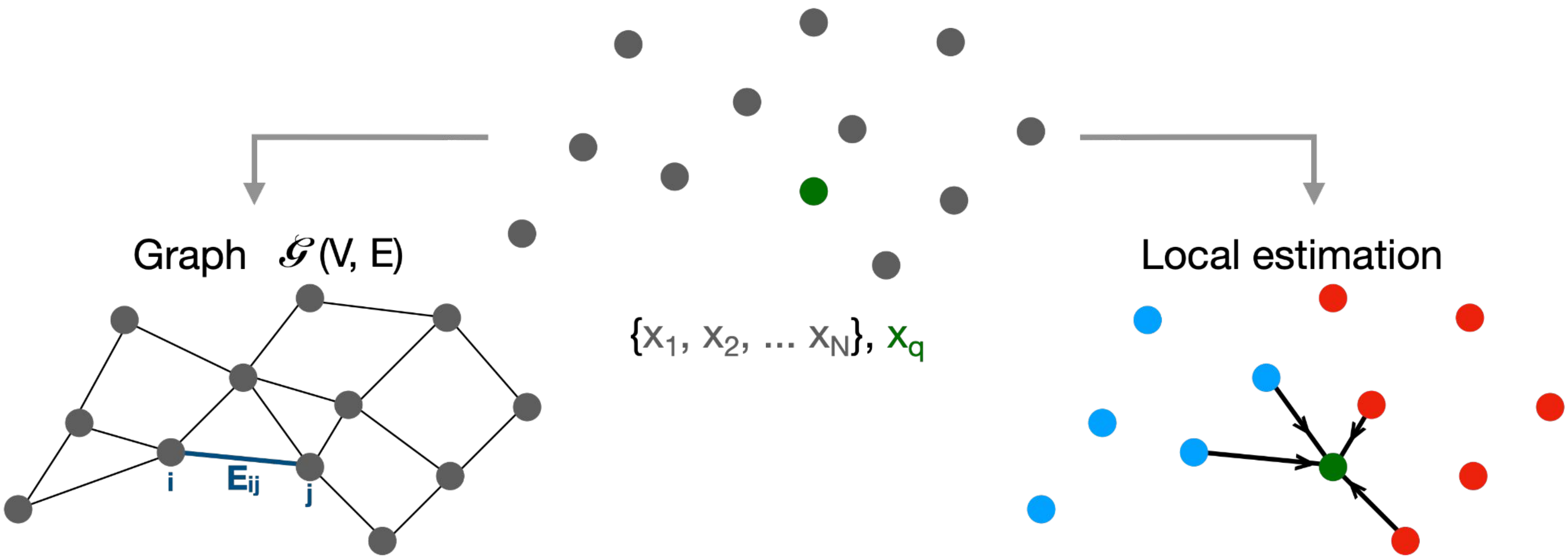


- Graphs **abstract** changes in dimensions, architecture, modality
  - **Geometric comparison** across models and layers

Overview

- ❏ Neighborhood ⇋ Non-negative sparse approximation
- ❏ Understanding deep learning models
  - ❏ Insight 1: Interpolation vs Model Size
  - ❏ Insight 2: Geometry of Self-supervised Models

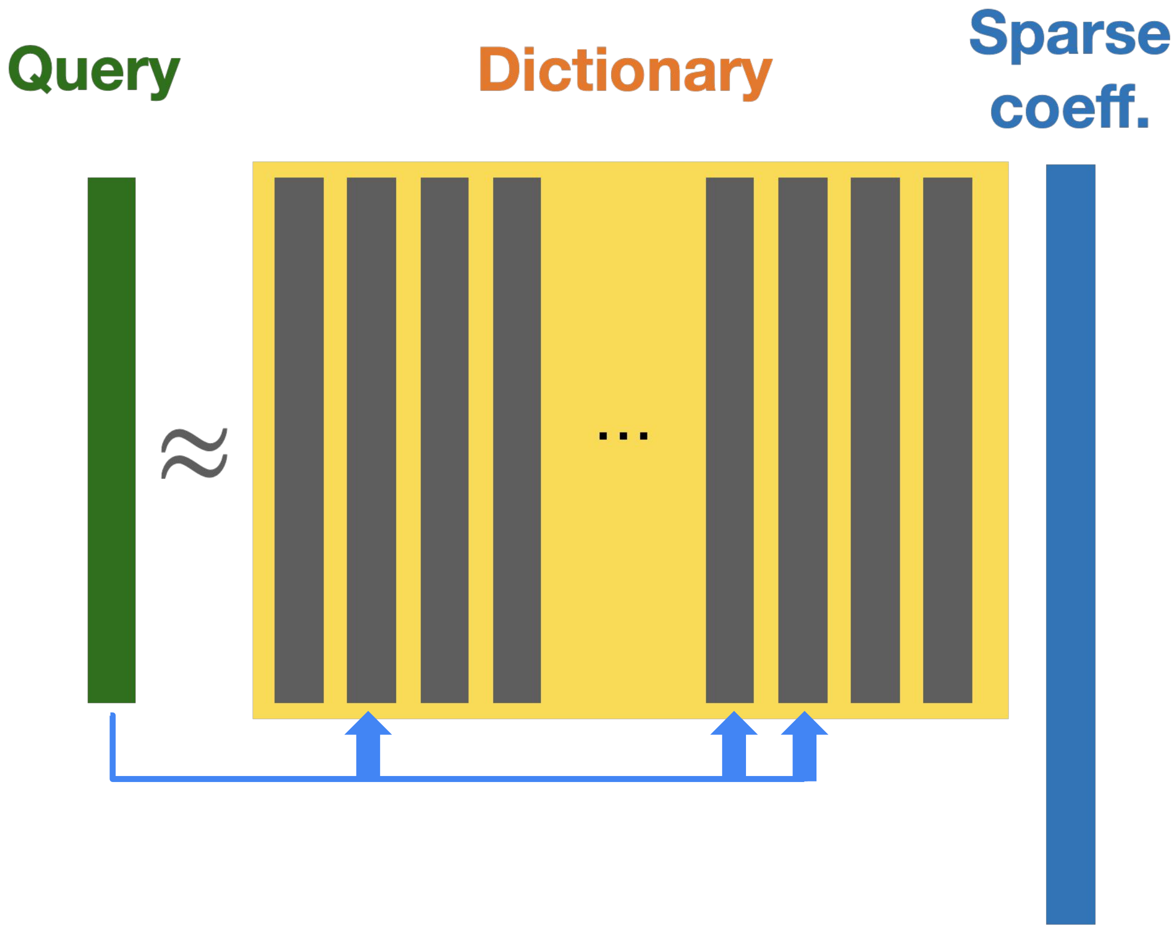
How to define a neighborhood?



- **First step** in graph based analysis, non-parametric estimation
- Definition impacts characterization (think “k” in kNN)
- Need: A principled formulation that is adaptive to data

Sparse Signal Approximation

Idea: Represent **input** using **few** elements (**atoms**) from a **dictionary**

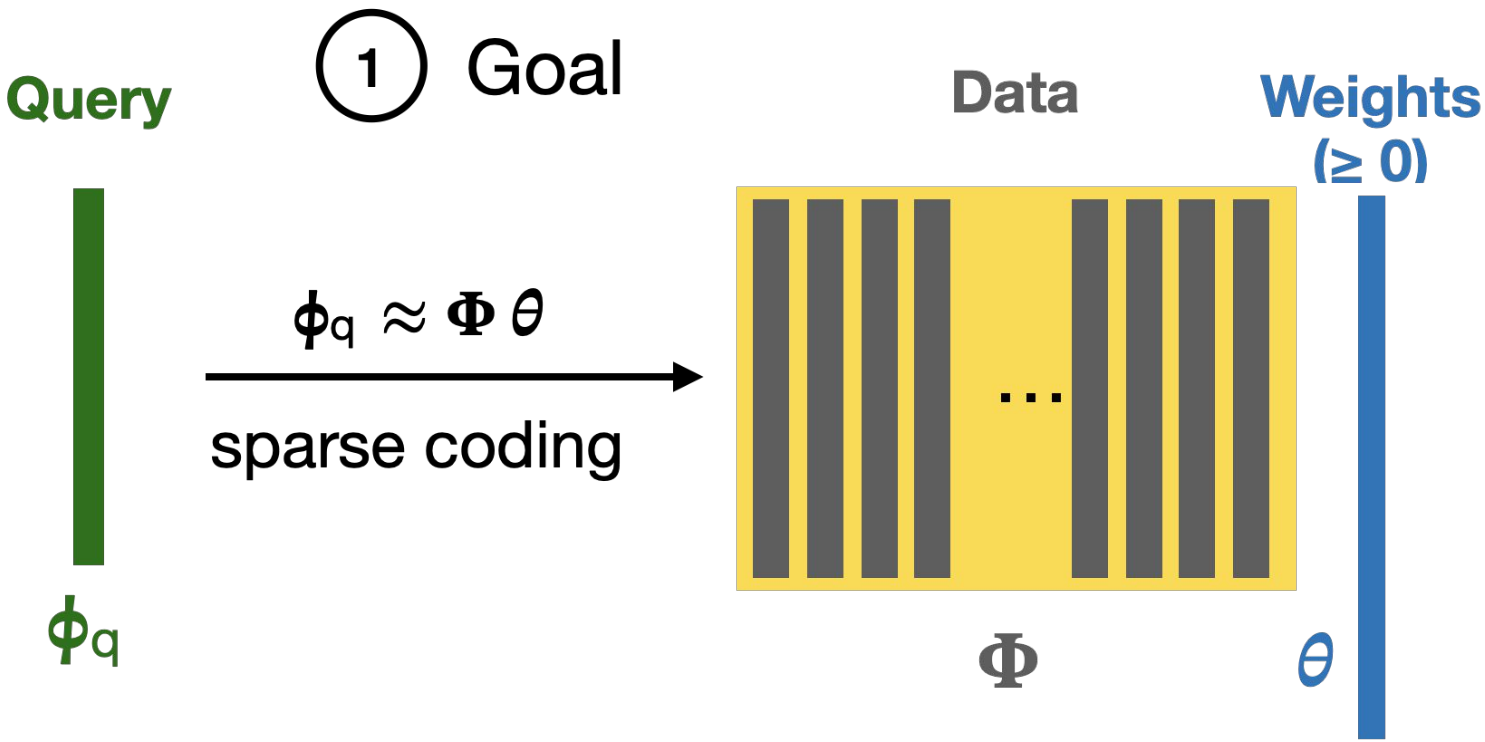


Want: Each selected element to **represent "new" information**

Neighborhood ⇋ Sparse signal approximation *(non-negative)

Setup:

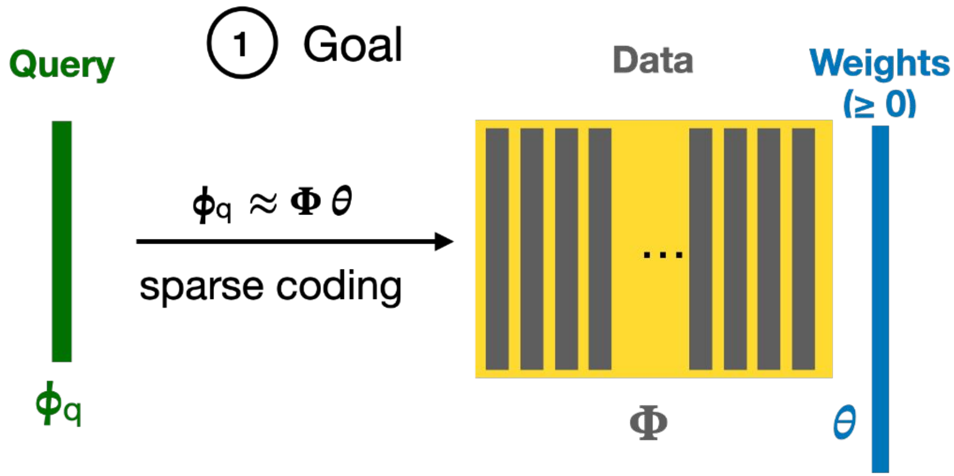
- ❏ **Given: Kernel similarity** ($\in$ [0, 1] - Normalized kernels)
- ❏ **Form a dictionary** based on kernel representation of data.



Neighborhood ≒ Sparse signal approximation
*(non-negative)

Setup:

- ❏ **Given: Kernel similarity** ($\in [0, 1]$ - Normalized kernels)
- ❏ **Form a dictionary** based on kernel representation of data.



2 k-NN, ε-Neighborhood: Thresholding

- Select "k" atoms with max $\phi_q^T \phi_a$
- Let S = selected atoms
- Optimal only if Φ_S is orthogonal

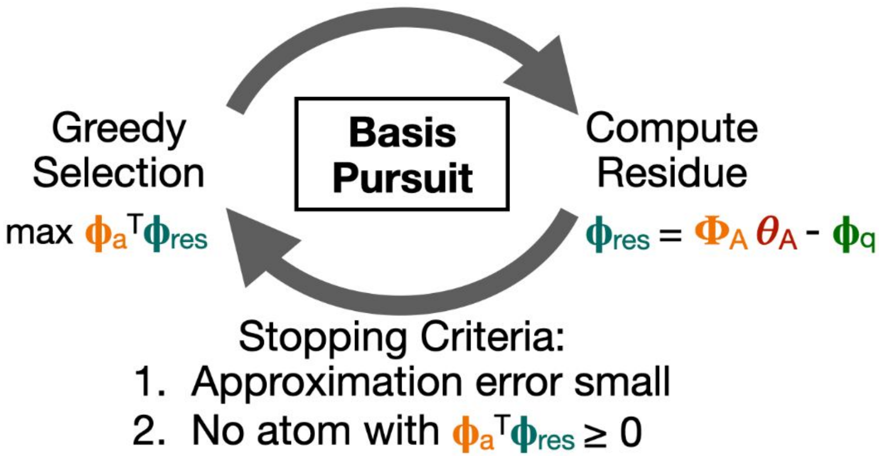
i.e., $\phi_j^T \phi_k = 0$

- Not the case in general

Is **thresholding** the best we can achieve?

Non-negative basis pursuit

- Select atoms such that **residue is represented** at each step
- Optimize selected atoms so that **residue is orthogonal**



Matching Pursuit (MP)

$\theta_a = \phi_a^T \phi_{res}$

Orthogonal Matching Pursuit (OMP)

$\theta_A = \mathrm{argmin}_{\theta \geq 0} \| \phi_q - \Phi_A \theta \|^2$

Stagewise Orthogonal Matching Pursuit

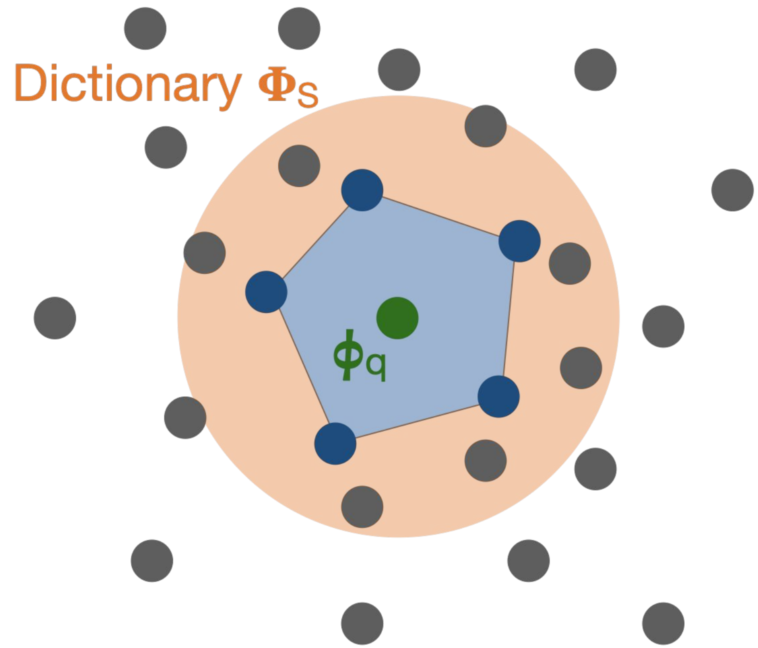
Greedy selection: Select all atoms above threshold

Leads to **adaptive, optimal** (OMP), **sparse** neighborhood definition

Cons: **Expensive iterative search**. Does not leverage problem setup

Non-Negative Kernel regression (NNK)

- Adapt dictionary for a query by using only **a relevant subset S**
  - E.g. kNN, Approximate neighbors
- Constrained optimization for **residue orthogonality**
- Equivalent to **1-stage OMP**



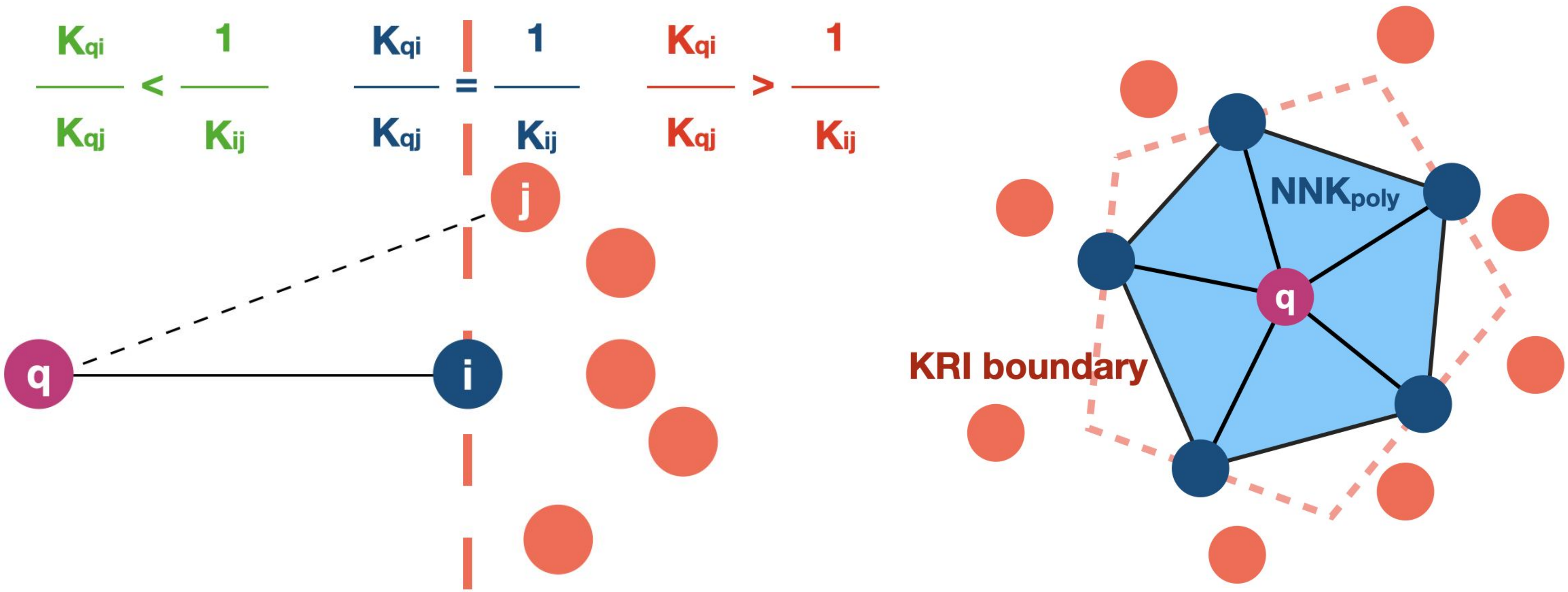
NNK Objective: $\min_{\theta \geq 0} \ \frac{1}{2}\theta^\top \mathbf{K}_{S,S}\theta - \mathbf{K}_{S,q}^\top \theta$

Runtime: $O(Nd\,|S|) + O(|S|^3)$

Geometry: Kernel Ratio Interval (KRI)

NNK construction depends on the relative position* of data

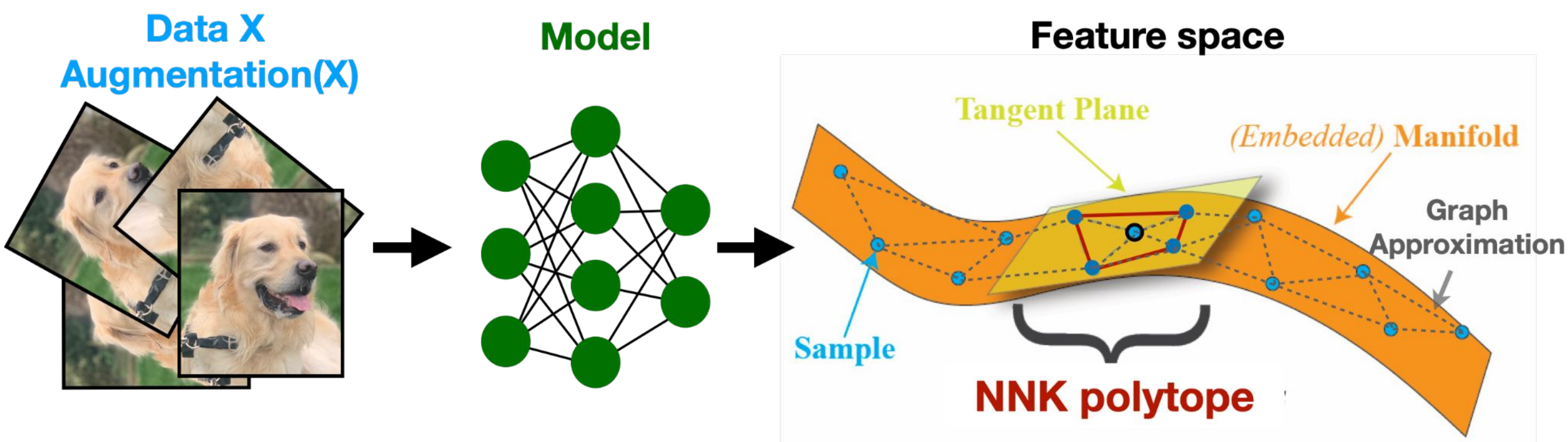
***metric on (i, j)**



NNK determines neighbors based on **hyperplanes and polytopes**

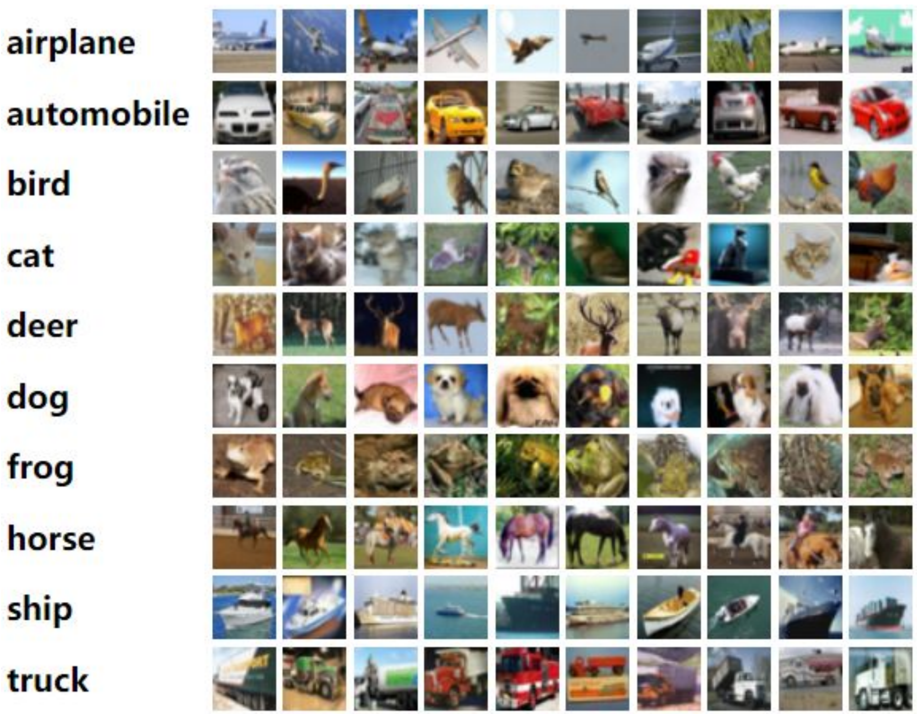
NNK graphs in deep embedding spaces

- ❑ Manifold metrics comparable across models and architectures



- ❑ NNK for extracting manifold properties
  - ❑ No. of NNK neighbors ⇔ **Intrinsic dimension**
  - ❑ NNK polytope diameter ⇔ **Invariance**
  - ❑ Polytope complexity⇔ **Embedding space complexity**
- ❑ Stability, Invariance via augmentations (perturbation of data)

Datasets

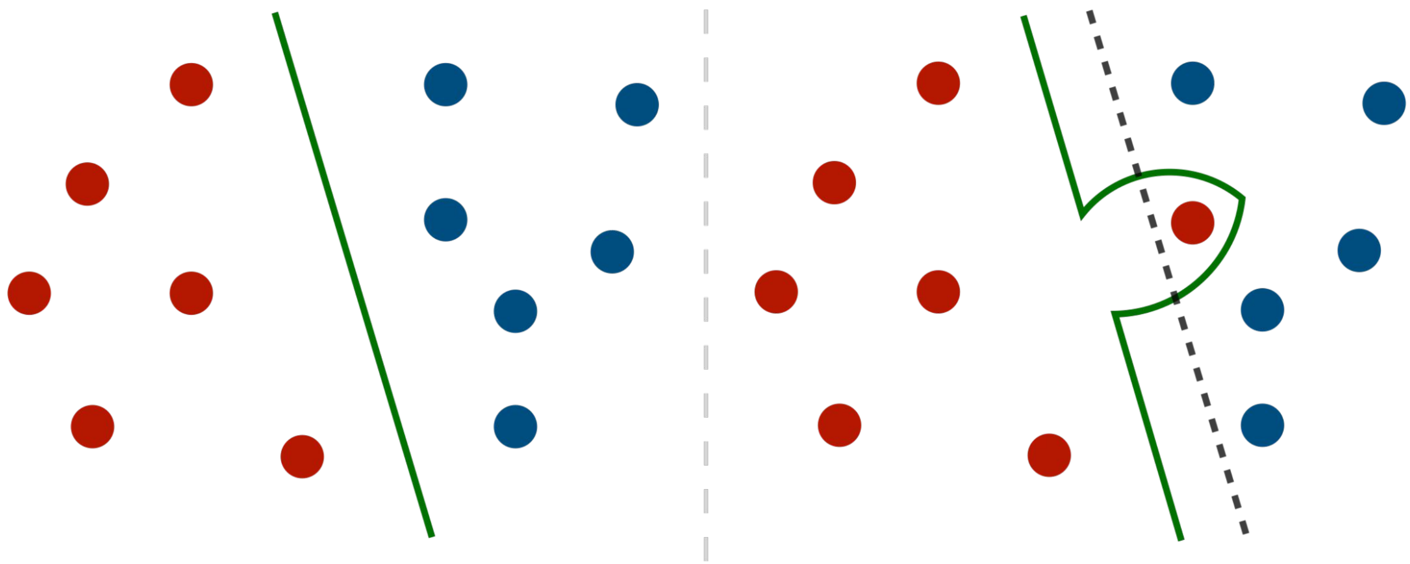


CIFAR10: Train (50k), Test (10k)
10 class

ImageNet: Train (1.2M), Val (50k)
1000 class

Revisiting interpolative estimators

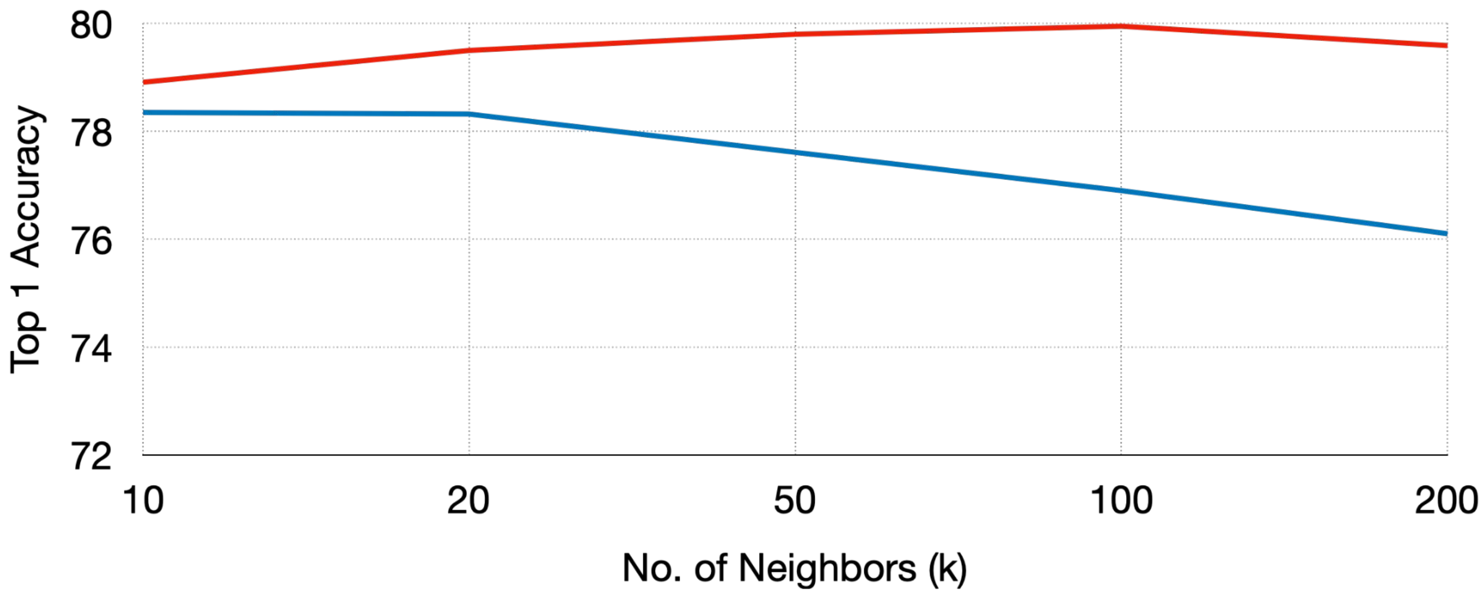
- Neural networks are trained to **zero loss** (Interpolative)
  - Can fit any data given time and capacity (Zhang '17)
  - Can generalize* even when data is noisy
- Involves complex **parametric / classification boundary**



- Graph: Empirical, **local characterization** of classification space

Classification performance: kNN vs NNK

- Setup: Classification of imagnet using embeddings from encoder
  - Self-supervised learning model: DINO ‘21
- Plot: kNN vs NNK performance for different choices of “k”

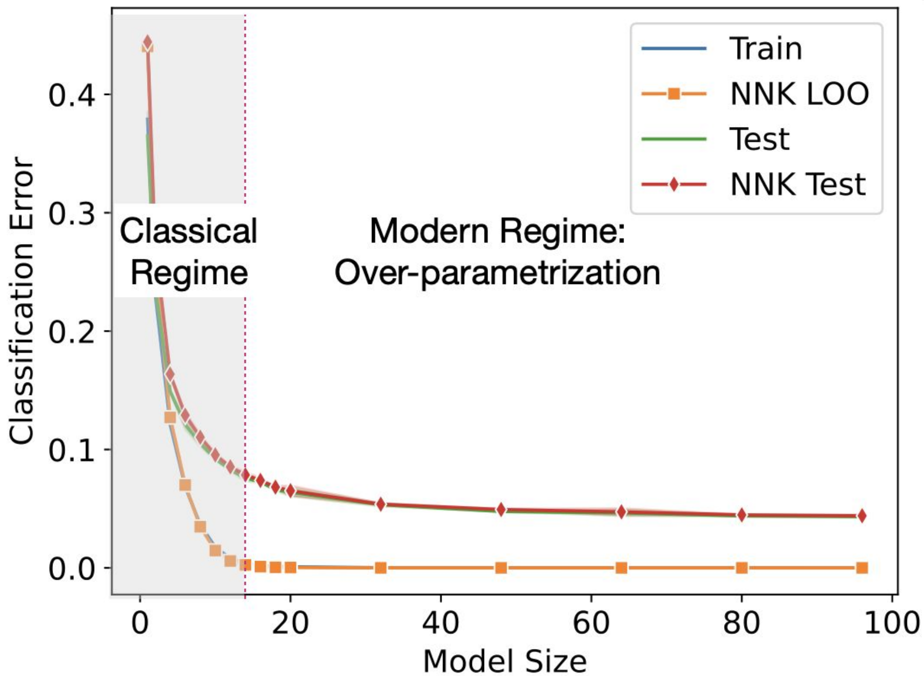


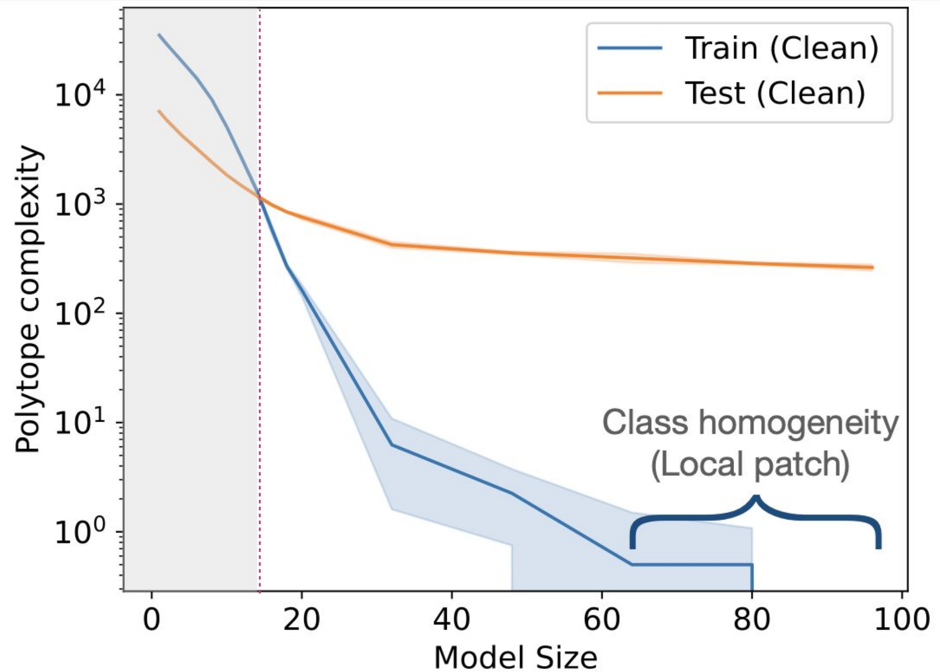
NNK interpolation on ImageNet achieves **79.9**% accuracy

Fine-tuned softmax classifier on same model 79.7%

Interpolation vs Model size

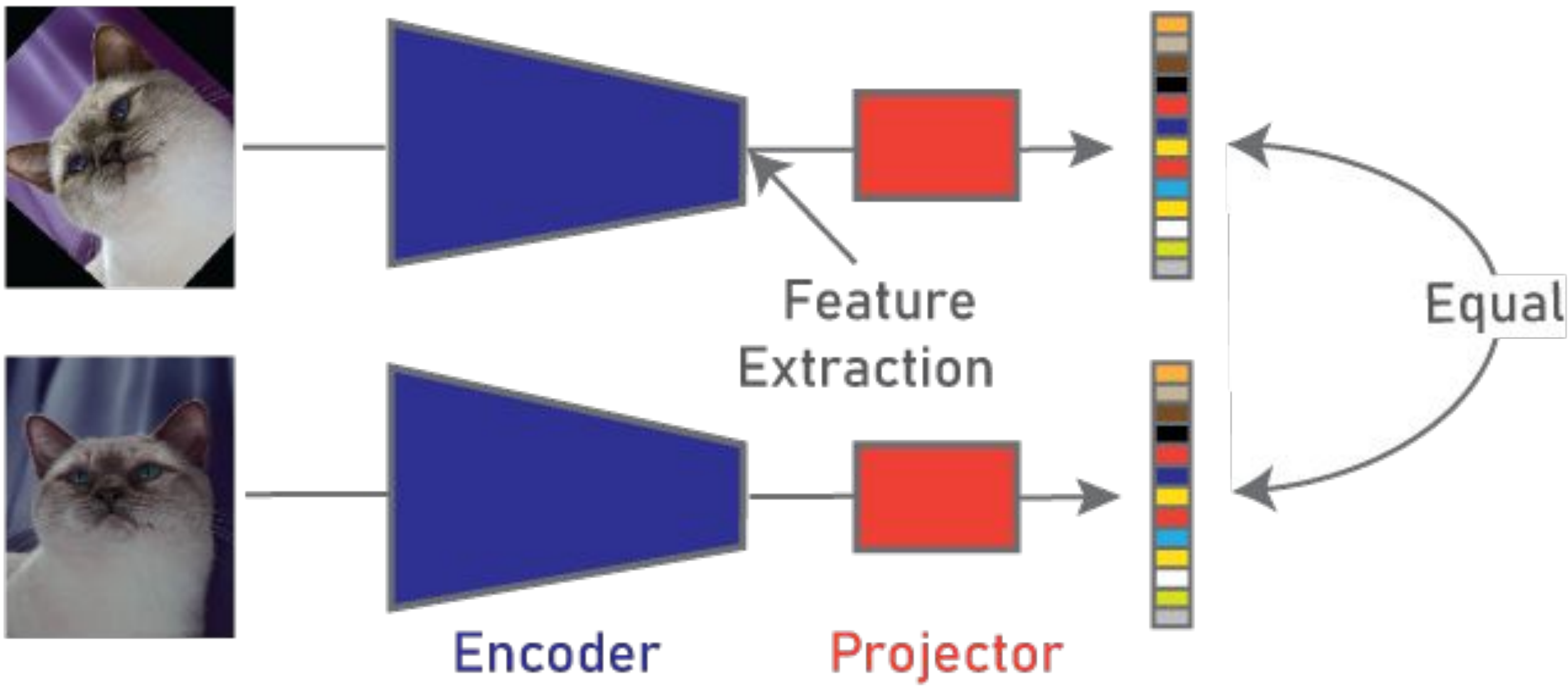
- ❑ Setup: ResNet 18 with variable block size (Model size)
- ❑ Observe complexity of local neighborhood
- ❑ # NNK polytopes with at least one neighbor with different label





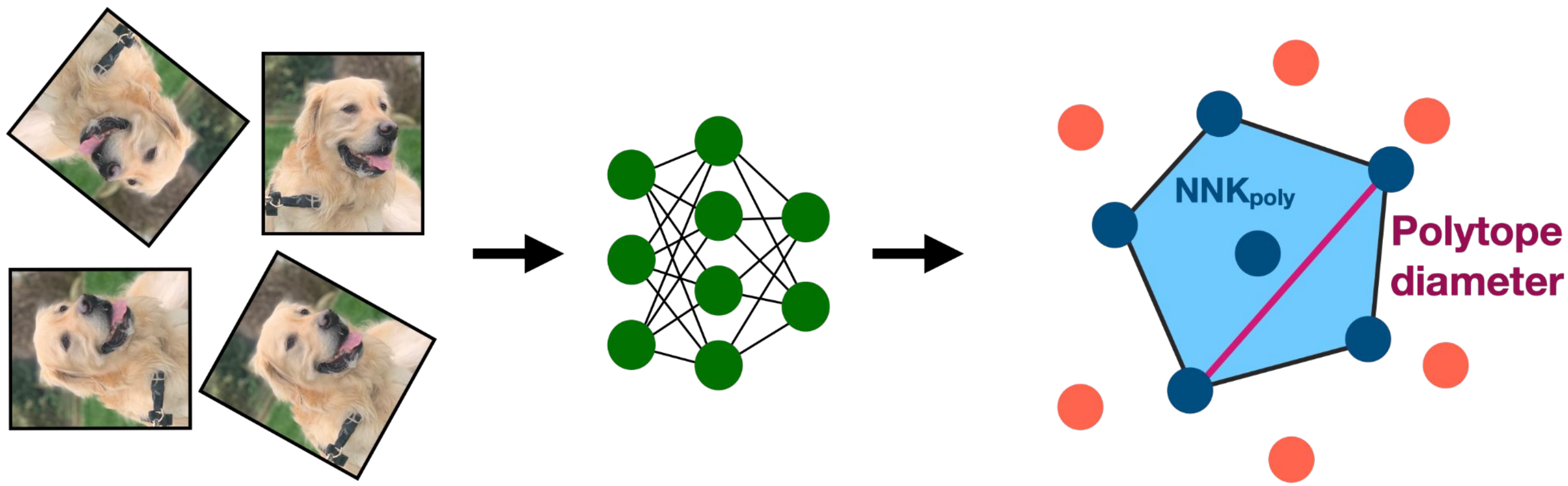
- ❑ NNK classification closely approximates the model performance
- ❑ Model size: From weighted interpolation to class homogeneity

Case study: Geometry of Self-Supervised learning



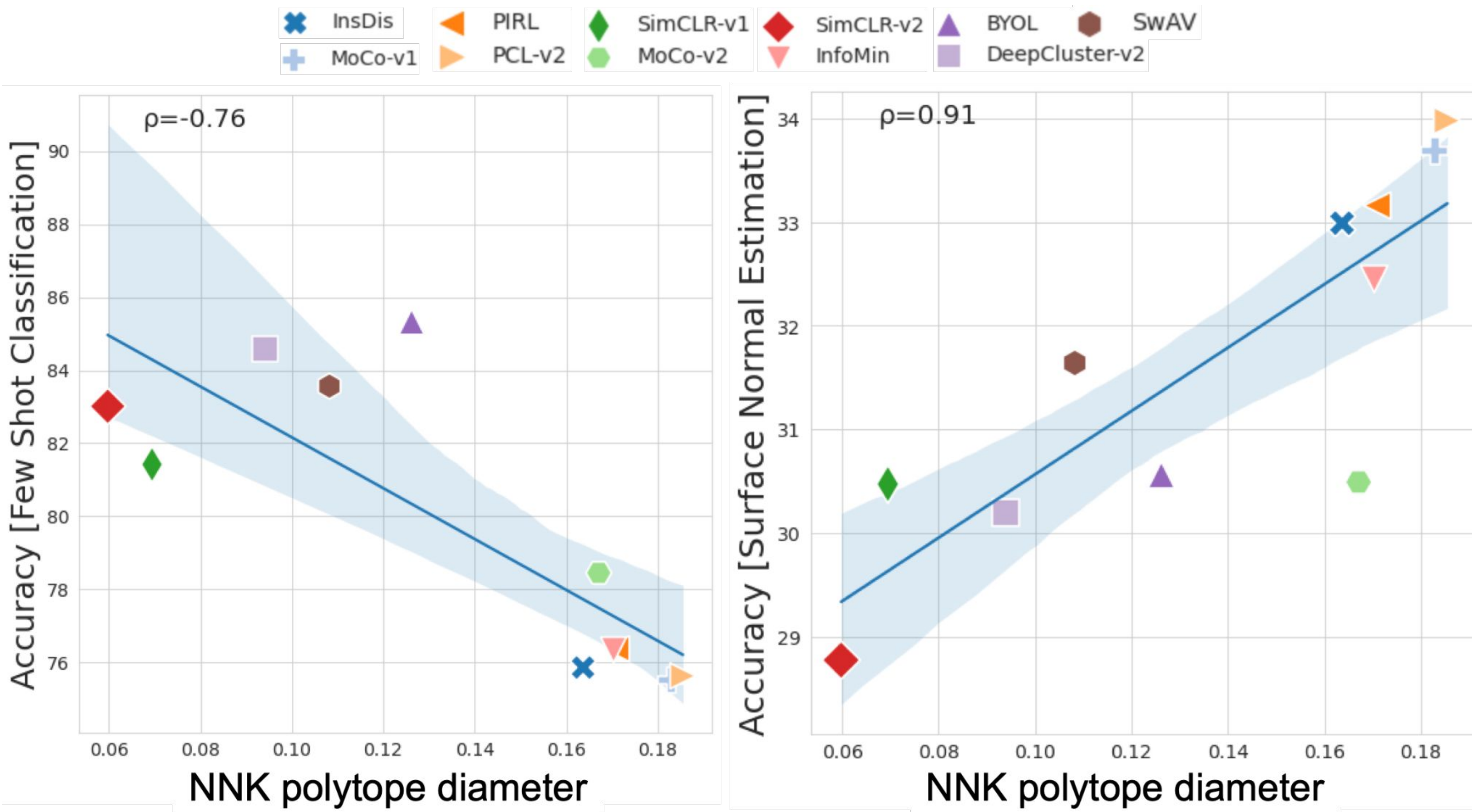
- ❑ Informal: Learning to be invariant to known prior (e.g. rotation)
- ❑ Several SSL models - which model to use for downstream task?
- ❑ How invariant is the encoder to an augmentation (perturbation)?

Setup: Rotation invariance of SSL models



- Feed an **ImageNet** image and its rotations as inputs to a model
- Obtain the NNK neighbors of inputs in encoder space
- Measure NNK polytope diameter
  - Max. distance between NNK neighbors (Range: [0, 2])
- Invariant ⇔ small diameter, Not invariant ⇔ large diameter

Results: Rotation invariance of SSL



- Rotation independent task: **Classification**
  - More invariant ⇔ Better performance
- Rotation dependent task: **Surface normal estimation**
  - Less invariant ⇔ Better estimation

Measured invariance to rotation correlates with downstream task

Summary

- Graph tools: Geometric understanding of deep learning
  - Properties of model beyond test accuracy
  - Applicable to other modalities & architectures
  - Explainability, Stability analysis, Model transfer

Resources (papers, code): shekkizh.github.io

References:

- **S.S**, A.Ortega, “Neighborhood and Graph constructions using Non-negative kernel regression”, IEEE TPAMI (under review), arXiv 2023
- C.Hurtado, **S.S**, J.Hidalgo, A.Ortega, “Study of Manifold Geometry using Multiscale Non-Negative Kernel Graphs”, ICASSP 2023
- R.Cosentino*, **S.S***, M. Soltanolkotabi, S. Avestimehr, A.Ortega, “The geometry of self-supervised learning models and its impact on Transfer learning”, arXiv 2022
- D.Bonet, J.Hidalgo, A.Ortega, **S.S**, “Channel redundancy and overlap in convolutional neural networks with Channel-wise NNK graphs”, ICASSP 2022
- **S.S**, A.Ortega “Revisiting local neighborhood methods in machine learning”, DSLW 2021
- **S.S**, A.Ortega, “Model selection and explainability in neural networks using a polytope interpolation framework”, ASILOMAR 2021