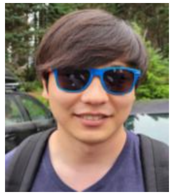
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Google Research

# LLC: Accurate, Multi-purpose Learnt Low-dimensional Binary Codes

Aditya Kusupati

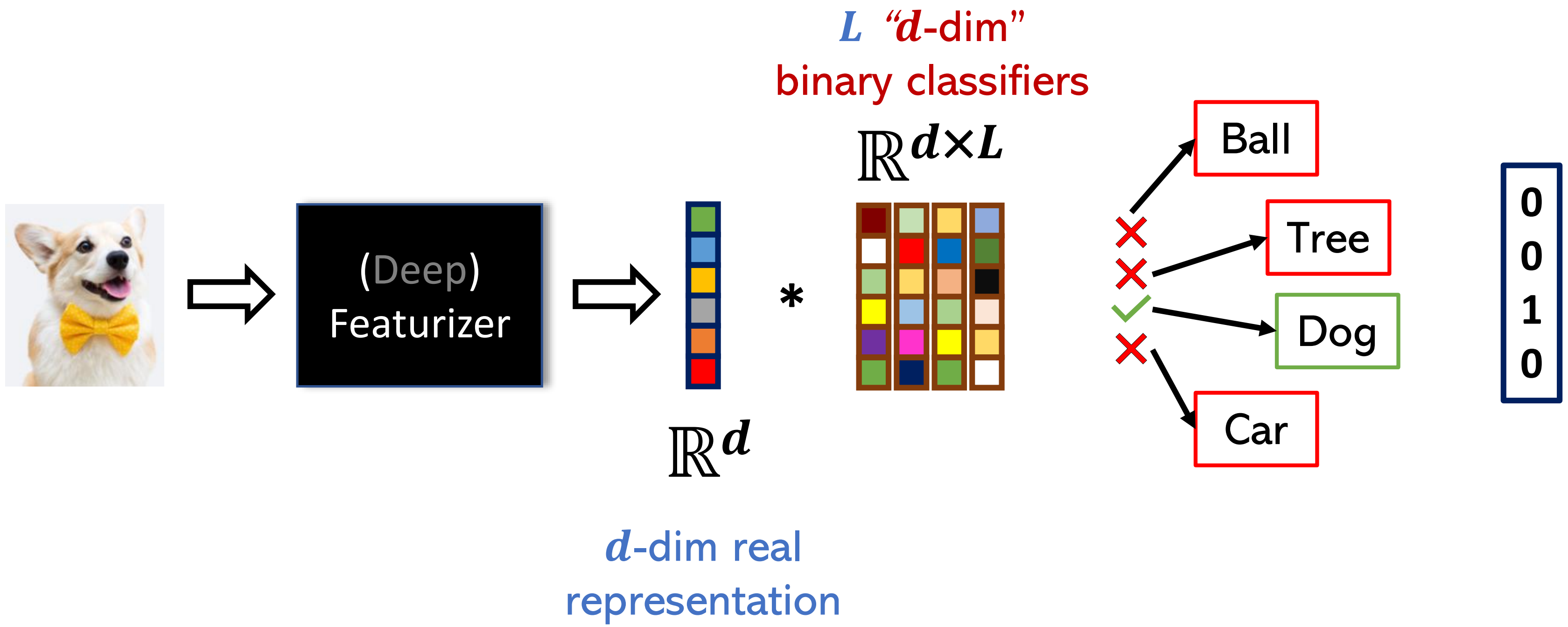




# Multiclass Classification

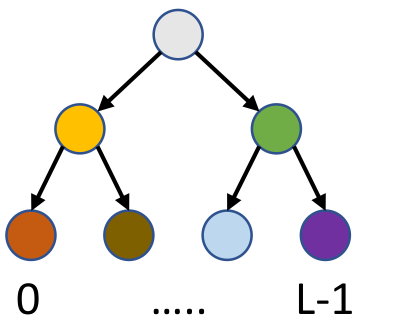
- Answers an MCQ among $L$ options (classes / labels)
- (Typically) solved using $L$ binary classifiers w/ only one correct answer
- Total probability of all classes sums to $\mathbf{1}$



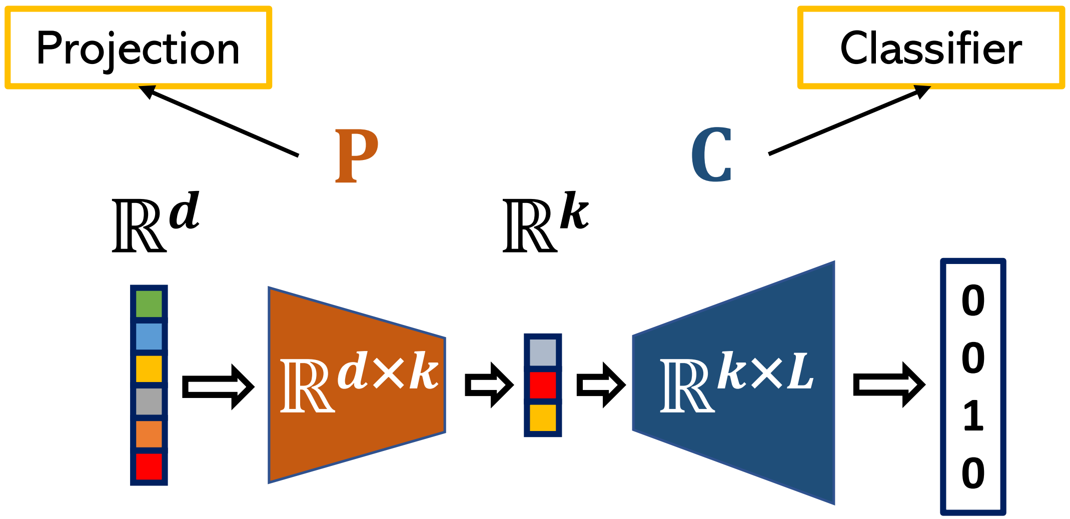
# Multiclass Classification: Costs & Trade-offs

- Constant featurization costs
- Linear Classifier: $\mathbb{R}^{d \times L}$ – compute & memory scale as $O(d * L)$

## Alternatives:




Trees / ANNS

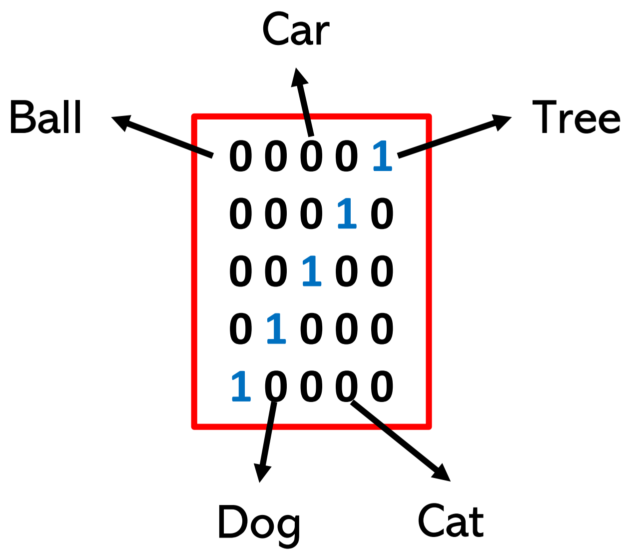



Embeddings

# Multiclass Classification: Output Codes

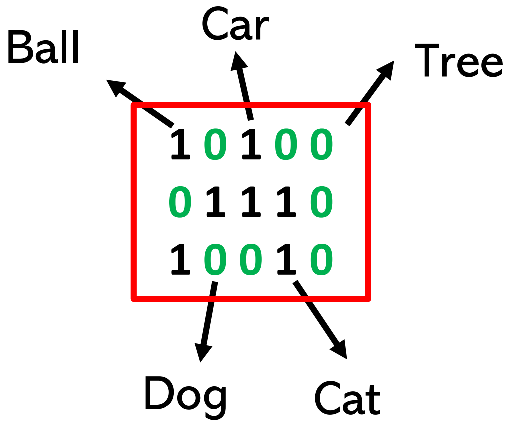
- Typically, one-hot vector per class

- Very sparse; can we do better?



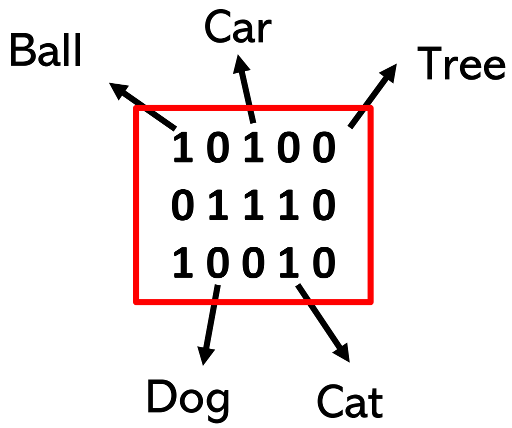
# Multiclass Classification: Output Codes

- Error Correcting Output Codes
- Codebook for classes (binary codes)
- Learn instance codes using codebook as a multi-label problem



# Output Codes: Shortcomings

- Construction of output codes
  - Attributes / Hierarchy / Random



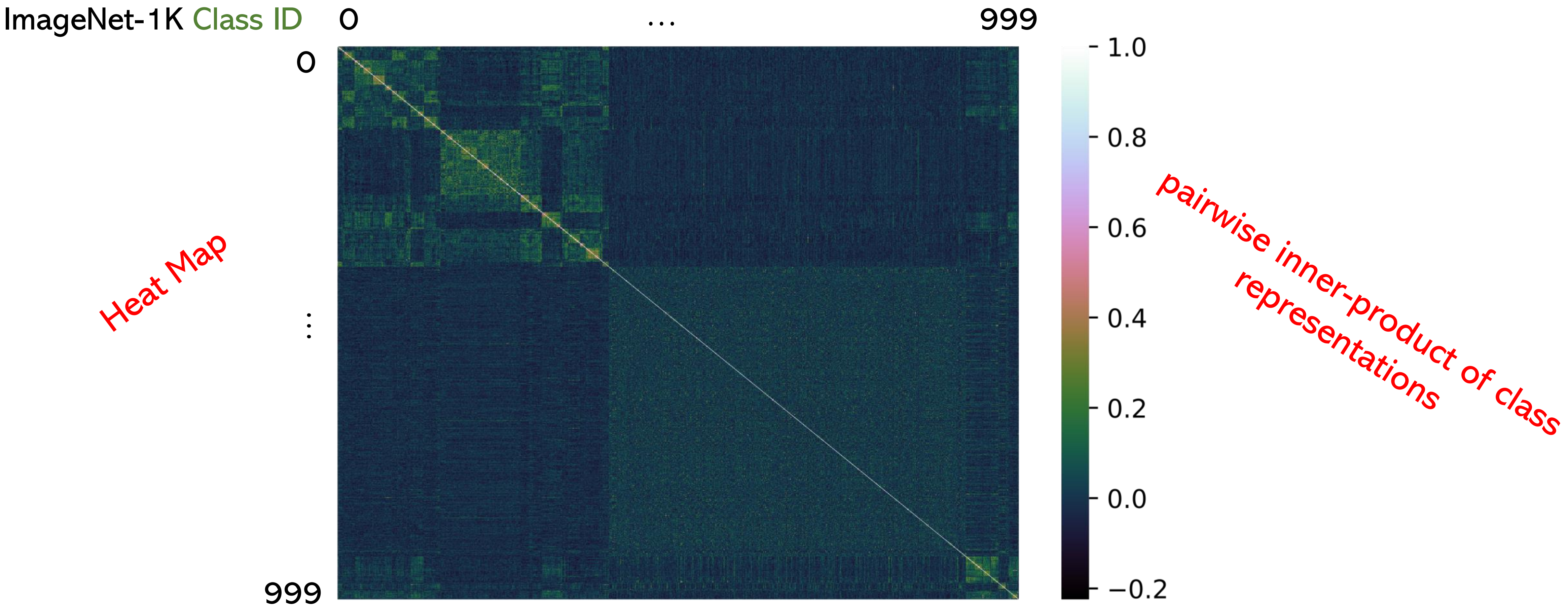

- Extremely hard to learn
  - Suffer from codebook collapse with most optimization techniques

- Inaccurate & often not tight - $O(\log L)$ with large constants

*Can we learn accurate & tight output codes without using any side information?*

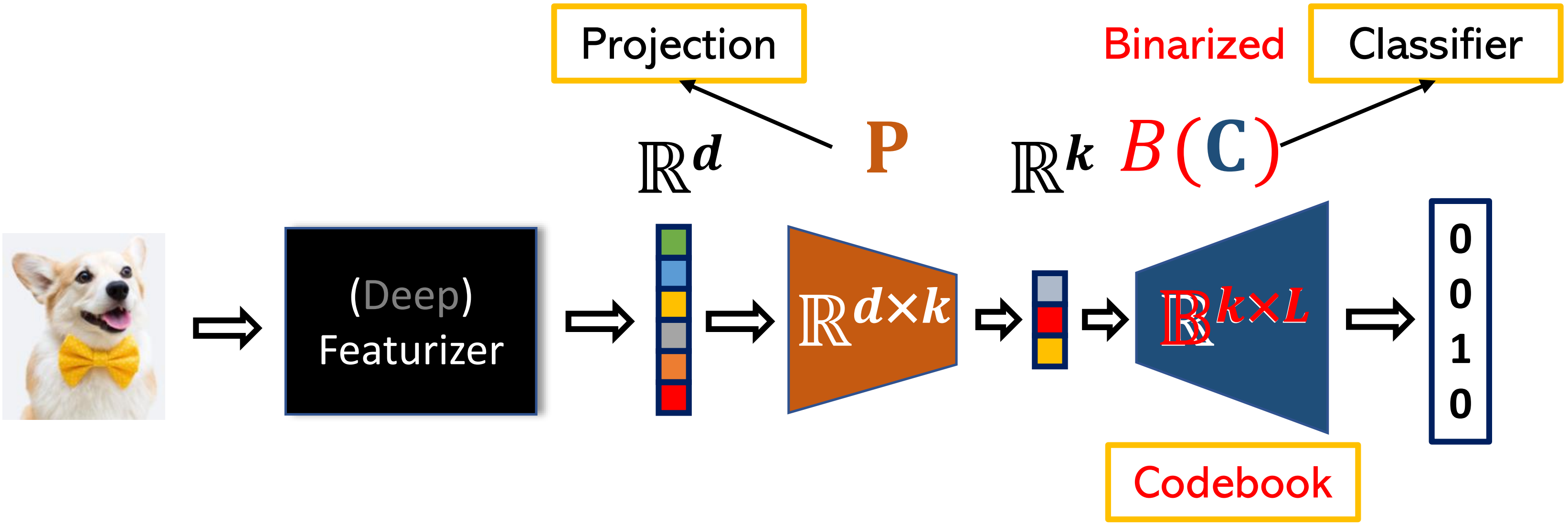
# Key Obs.: Accurate Class Representations

- Classifier works great; ImageNet - $\mathbf{1000}$ $(L)$ 2048-dim real vectors $(\mathbb{R}^d)$



# Low-dimensional Class Codes

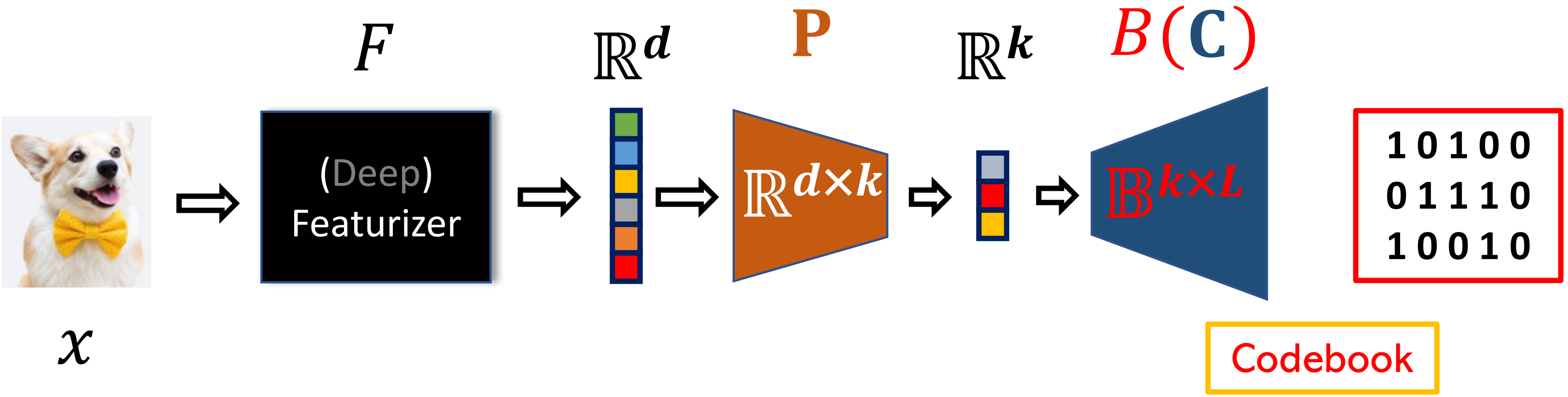
- Linear classifier on low-dimensional space ($\mathbb{R}^k$; $k \approx \log(\boldsymbol{L}) \ll d$)



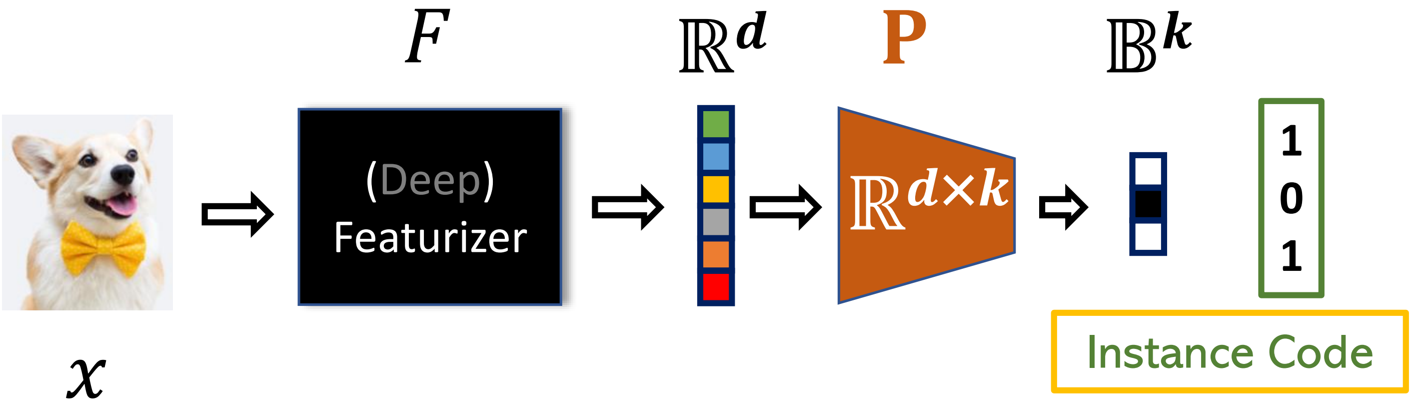

- Learn binarized ($\mathbb{B}^{k\times L}$) {0/-1, 1} version of the linear classifier
- $B(\mathbf{C})$ is the collection of class codes aka Codebook

# The **LLC** Method

- Phase 1 – Codebook Learning

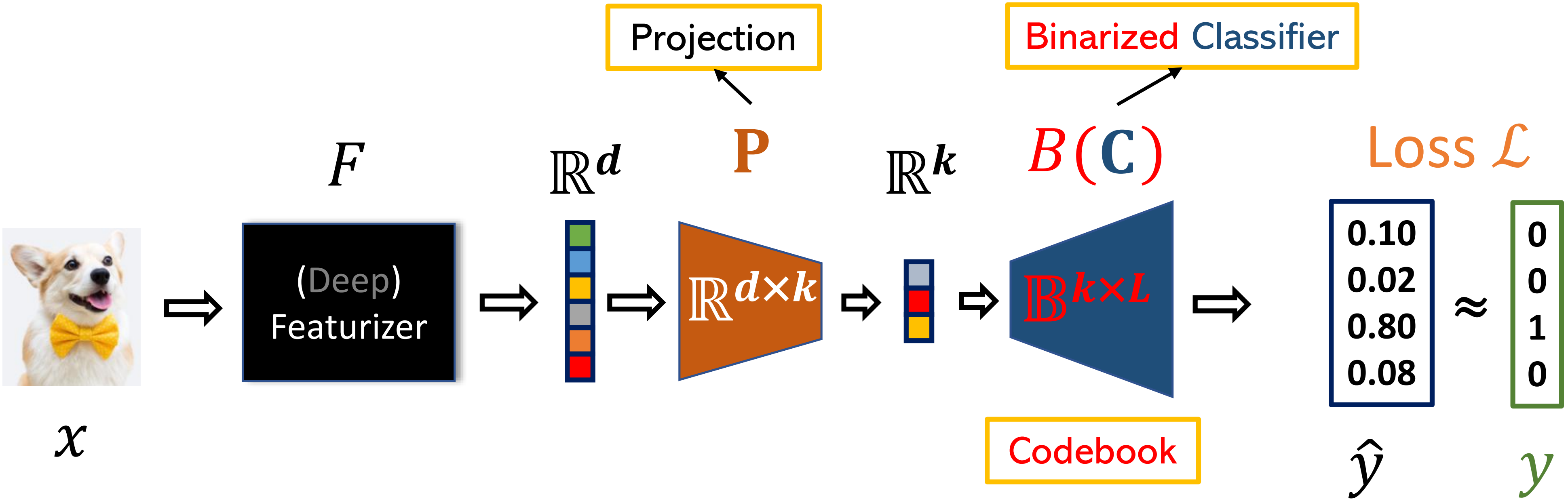



- Phase 2 – Instance Code Learning



# **LLC**: Phase 1 – Codebook Learning

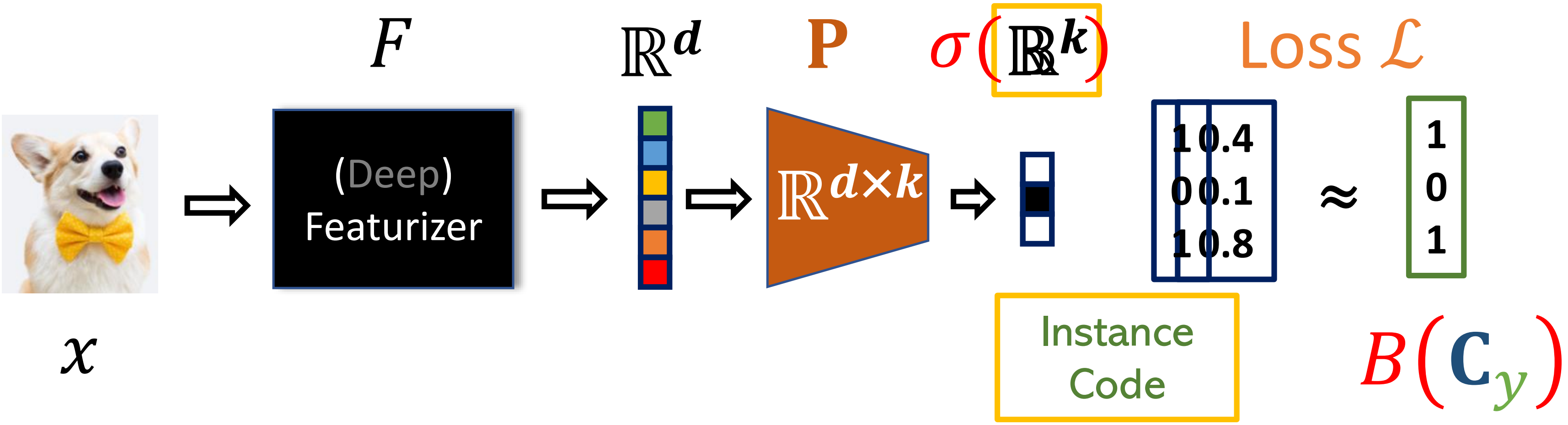
- Learnt end-to-end - $F$, $\mathbf{P}$ & $B(\mathbf{C})$ - w/ Stochastic Gradient Descent (SGD)




- Binarization is learnt through Straight Through Estimator (STE)

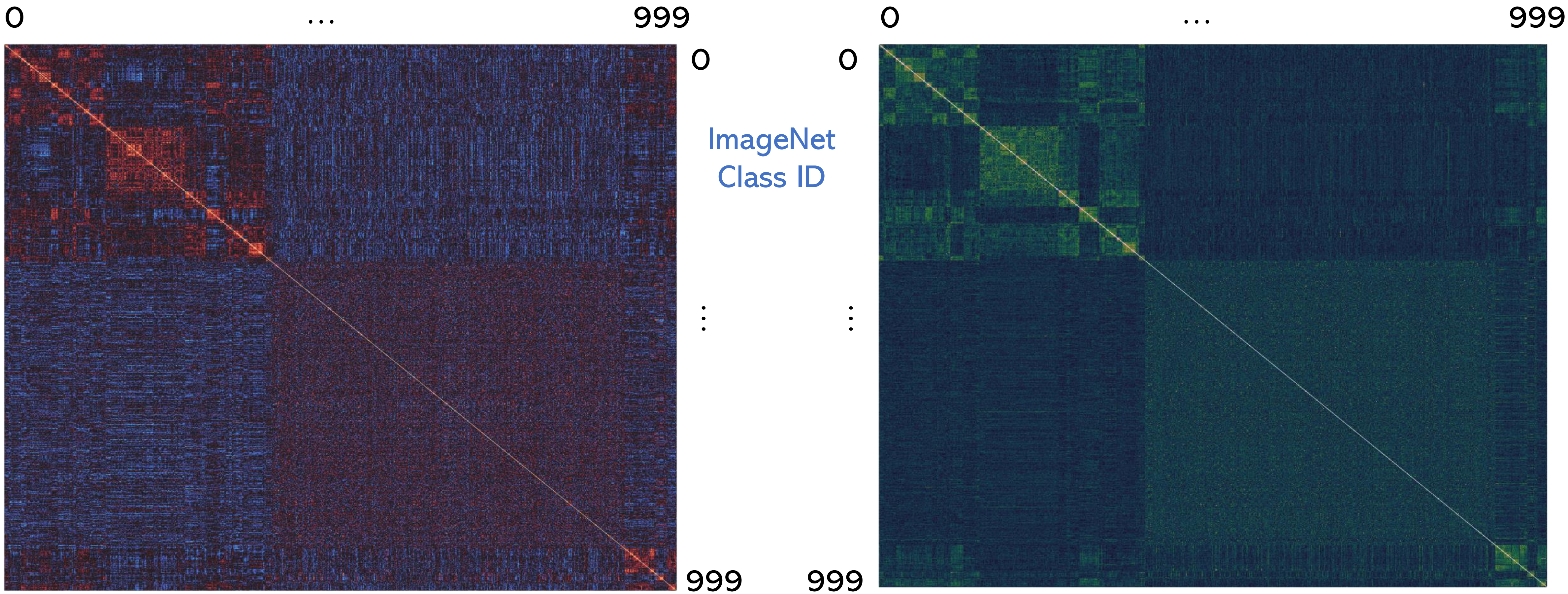
# **LLC**: Phase 2 – Instance Code Learning

- Warm start w/ $F$, $\mathbf{P}$: generate $k \approx \log(\boldsymbol{L})$ dim instance representation
- Columns of $B(\mathbf{C}) \in \mathbb{B}^{\boldsymbol{k \times L}}$ - target output labels per class $B(\mathbf{C}_y)$
- Solve the multi-label problem as $k$ binary classification problems
- Binarize the multi-label predictions to obtain instance codes $(\mathbb{B}^{\boldsymbol{k}})$



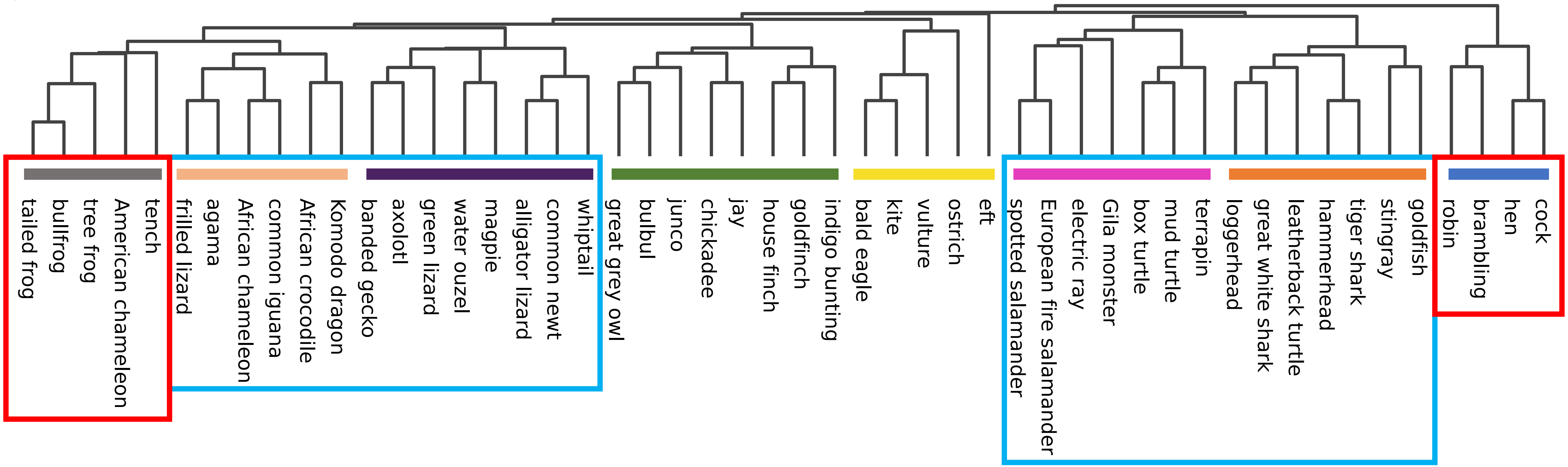
# Heat Maps: Comparison for ImageNet-1K

- 20-bits produce a visually similar heat map as 2048-dim real numbers



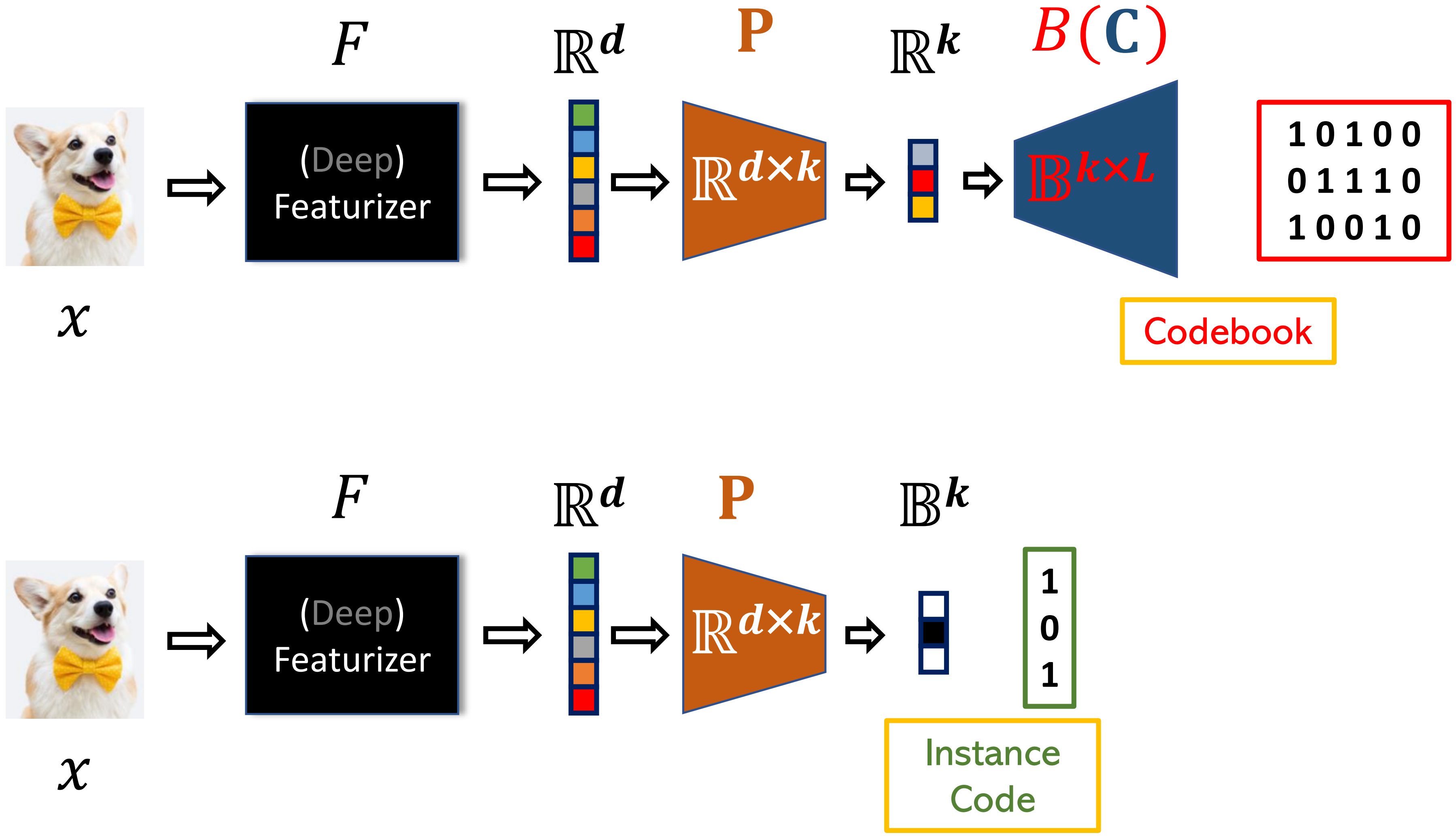
# Discovered Taxonomy: 20-bit Codebook

- Agglomerative clustering builds an intuitive hierarchy
- 20-bits capture important high-level semantic information




Discovered Taxonomy for 50 classes of ImageNet-1K

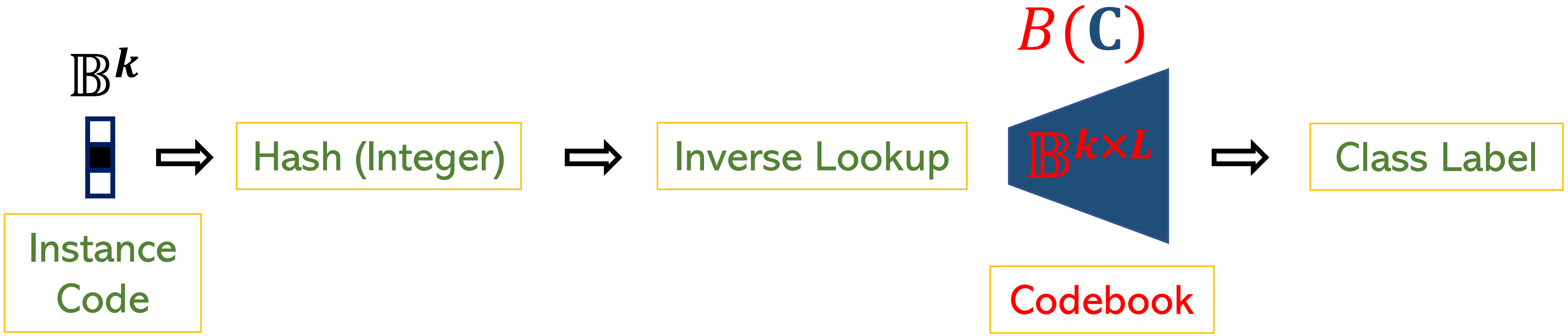
# LLC: Class & Instance Codes



# **LLC**: Decoding Schemes for Classification

Exact Decoding (ED):

- Instance code needs to exactly match the ground truth class code

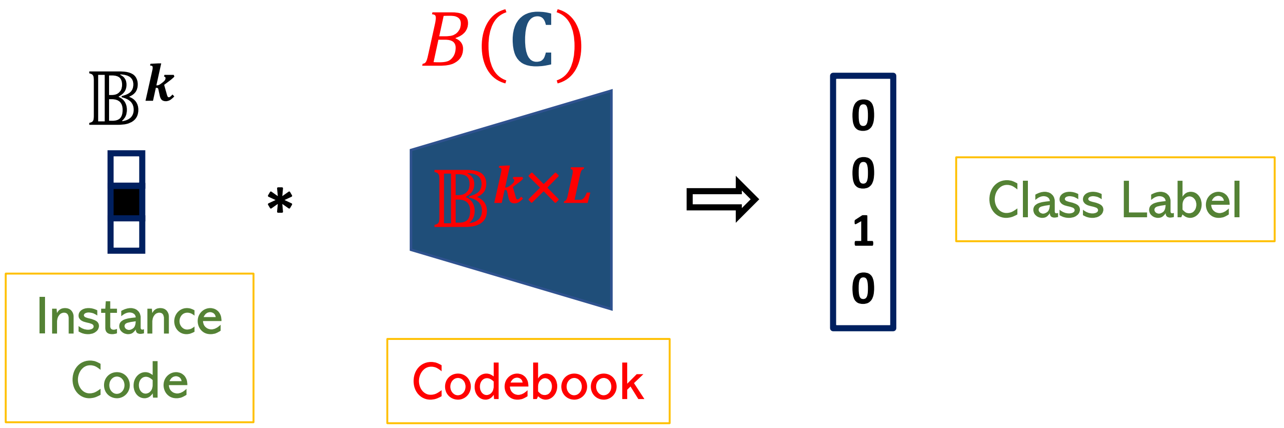



- Extremely efficient – Scales as $O(k \approx \log L) + O(1)$ hash lookup

# **LLC**: Decoding Schemes for Classification

Minimum Hamming Decoding (MHD):

- Retrieves the closest class code for an instance code using Hamming distance




- Efficient - Scales as $O(k * L)$ instead of the typical $O(d * L * r)$

  $r$ is the relative cost of real-valued compute over bits

# **LLC:** Applications

- Efficient Multiclass Classification
  - ImageNet-1K with 20 bits
  - Multiple decoding schemes for compute and accuracy trade-offs
- Efficient Retrieval
  - ImageNet-100 with 10 bits
  - Potential for low-latency high recall retrieval in search systems
- Out-of-Distribution (OOD) Detection
  - Out-of-the-box without tuning for threshold

# **LLC**: Image Classification for ImageNet-1K

- Comparison across 20-bit codebooks using ResNet50 backbone

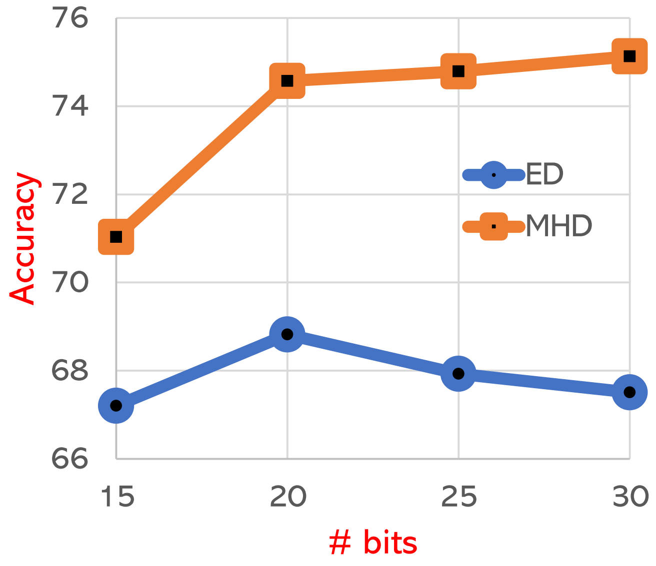
| **Codebook** | **# Unique Codes** | **ED Accuracy (%)** | **MHD Accuracy (%)** |
|---|---|---|---|
| Random | 1000 | 64.07 | 66.91 |
| CCA | 813 | 55.17 | 57.03 |
| SVD | 969 | 65.12 | 69.18 |
| **LLC (Ours)** | **1000** | **68.82** | **74.57** |

- LLC learns more accurate, tight & reliable codebooks

- ResNet50 with 2048-dim real representation + linear classifier: **77%**

# **LLC**: Image Classification for ImageNet-1K

- Classification accuracy *vs.* # bit codes for LLC on ResNet50 backbone




- MHD accuracy ↑s gradually; but ED accuracy ↑s & ↓s
  - LLC gets only 19.2 & 28.5 bits right for 20 & 30 bit models respectively

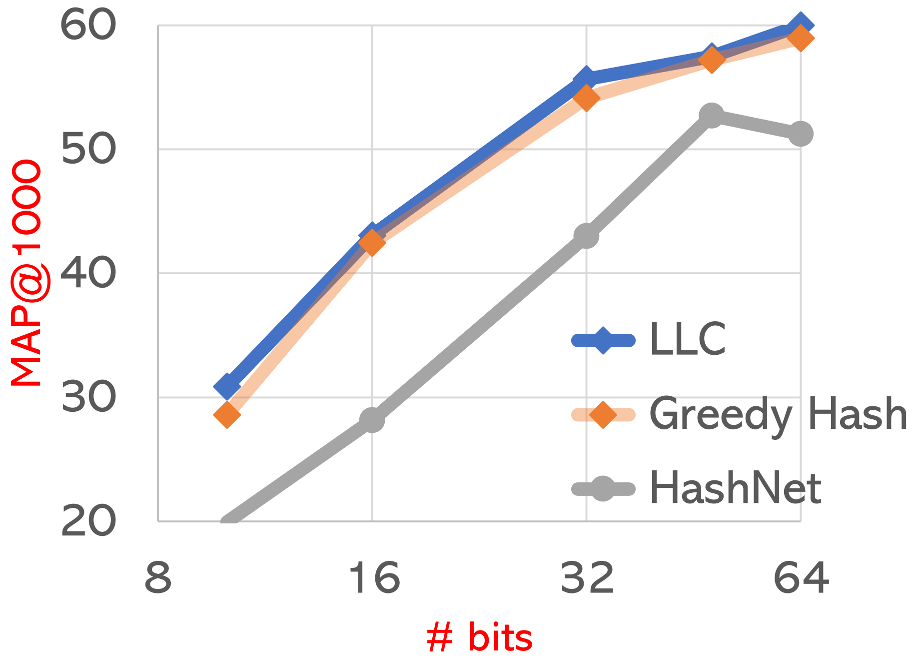
# LLC: Image Retrieval for ImageNet-100

- Encode the image and database using instance binary codes
- For a query image, retrieve images of the same class
- MAP@1000 performance for retrieval using ResNet50 backbone

| Representation | 8 dims | 10 dims | 64 dims |
|---|---|---|---|
| LLC (1-bit) | - | 64.07 | 67.73 |
| Real (16-bits) | 50.41 | 66.57 | 77.94 |

# LLC: Image Retrieval for ImageNet-100

- MAP@1000 vs # bits comparison for retrieval using AlexNet backbone



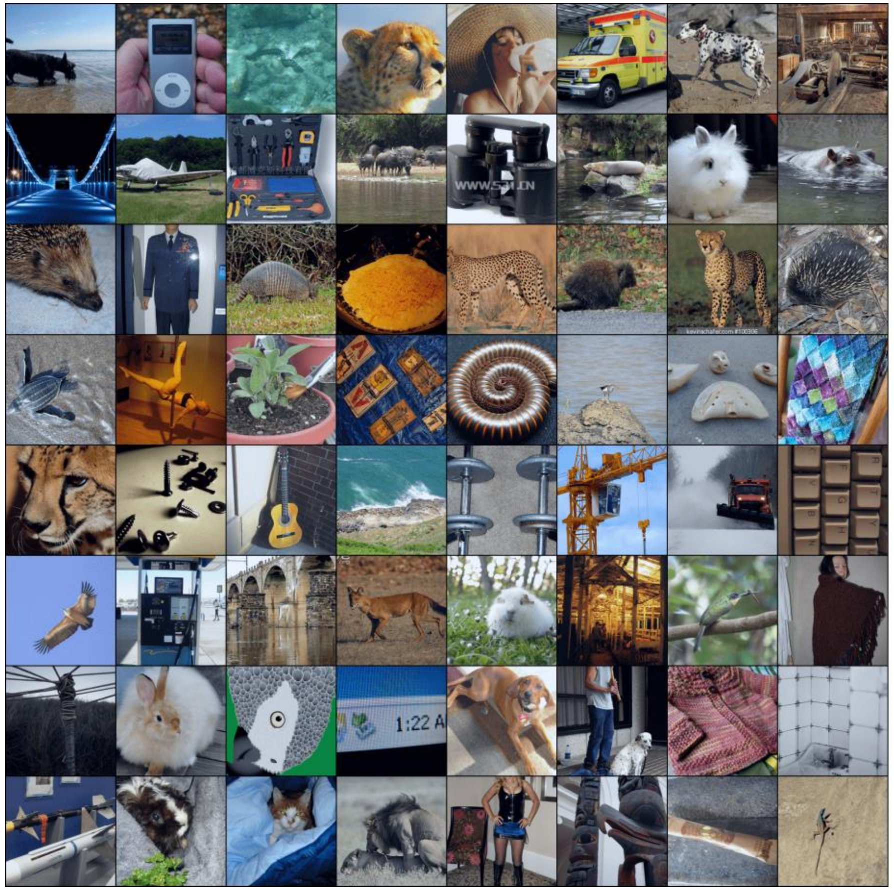

- Baselines are designed to learn only instance codes

# **LLC**: Ablations

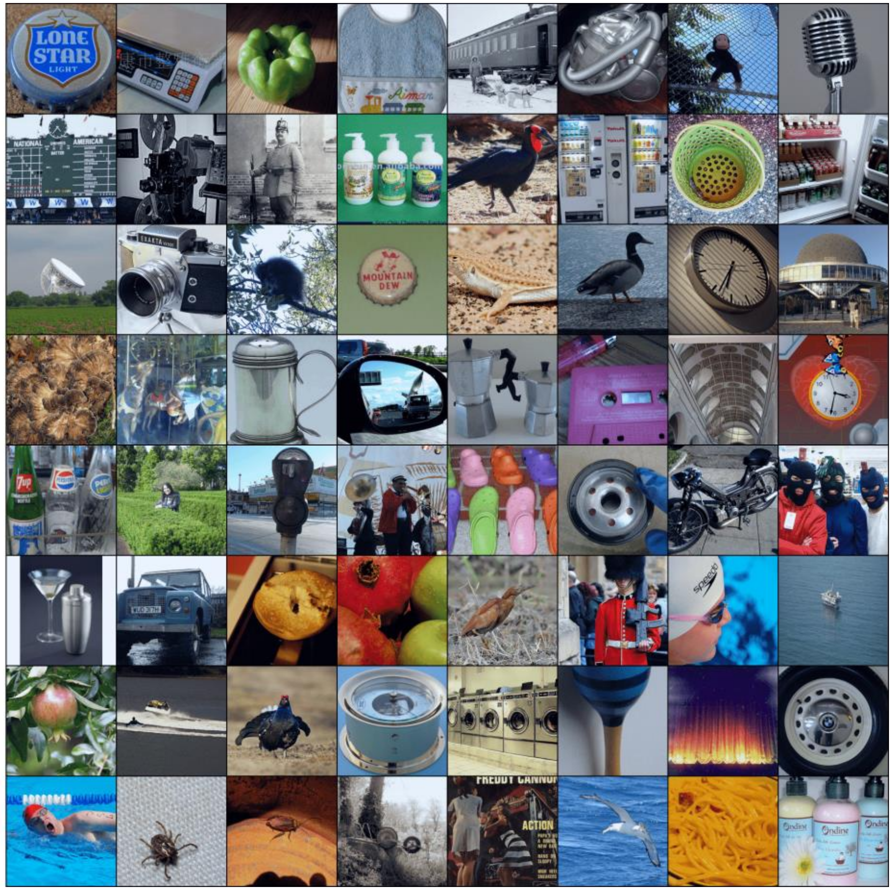
- Faster codebook learning
  - Cheaper backbones; lesser data
  - Warm start with collapsed SVD/CCA codebooks
- Linear vs. Hamming separability
  - LLC makes representation Hamming separable (hypercube)
  - Hamming separability (more efficient) → Linear separability (more accurate)
- Nested codebook learning
  - Train a 30-bit codebook, get 20-bit and 25-bit codebooks for free
  - Extremely useful in deployment across various budgets without retraining

# **LLC**: Are Learnt Codes Interpretable? NO!

Bit 4



1



0

# **LLC**: Conclusions

- First method to learn both class & instance codes simultaneously
- Learns semantically rich, highly accurate & tight binary codes reliably
- Enables applications like *classification* & *retrieval* in sub-linear costs
- Future Work:
  - Million-Billion scale classification & retrieval for search / instance classification
  - Extremely efficient edge classifiers for tiny devices
  - Cross-modal representation learning for interpretability in sub-linear costs
  - Binary representations for various domains like videos, controls etc.,