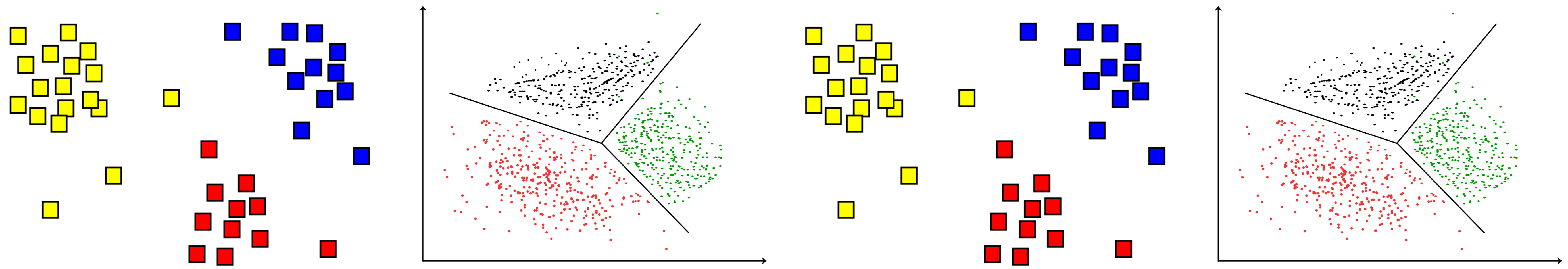
THE UNIVERSITY OF BRITISH COLUMBIA

CPSC 425: Computer Vision



Lecture 18: Visual Classification 1, Bag of Words

Menu for Today

Topics:

— Visual **Classification**

— **Bag of Words** Representations

Readings:

— **Today's** Lecture: Szeliski 11.4, 12.3-12.4, 9.3, 5.1-5.2

Reminders:

— **Quiz 4** will be available tonight (Topics: SIFT, Image Warping, Stereo)

— **Quiz 5** will be next Monday (Topics: Optical Flow, Classification)

— Issue with **Assignment 5** (see Piazza, instructions have been updated)

CVPR 2025

The **IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)** is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. With its high quality and low cost, it provides an exceptional value for students, academics and industry researchers.

Important Dates

Submitting 8 papers

Paper Submission Deadline	Nov 14 '24 11:59 PM PST *	00 weeks 00 days 10:27:52
Supplementary Materials Deadline	Nov 21 '24 11:59 PM PST *	01 weeks 00 days 10:27:52
Reviews Released	Jan 23 '25 01:59 AM CST *	
Rebuttal Period Ends	Jan 30 '25 01:59 AM CST *	
Final Decisions	Feb 26 '25 01:59 AM CST *	
All dates	Timezone: America/Vancouver	

CVPR 2025

The **IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)** is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. With its high quality and low cost, it provides an exceptional value for students, academics and industry researchers.

Important Dates

Submitting 8 papers

Paper Submission Deadline	Nov 14 '24 11:59 PM PST *	00 weeks 00 days 10:27:52
Supplementary Materials Deadline	Nov 21 '24 11:59 PM PST *	01 weeks 00 days 10:27:52
Reviews Released	Jan 23 '25 01:59 AM CST *	
Rebuttal Period Ends	Jan 30 '25 01:59 AM CST *	
Final Decisions	Feb 26 '25 01:59 AM CST *	
All dates	Timezone: America/Vancouver	

Training of Vision-Language Models


Jiayun Luo


Rayat Hossain

A big tan stuffed bear sitting in front of the store where there are many sale items on display; the door appears to be closed with no people in sight.

Training of Vision-Language Models

Jiayun Luo

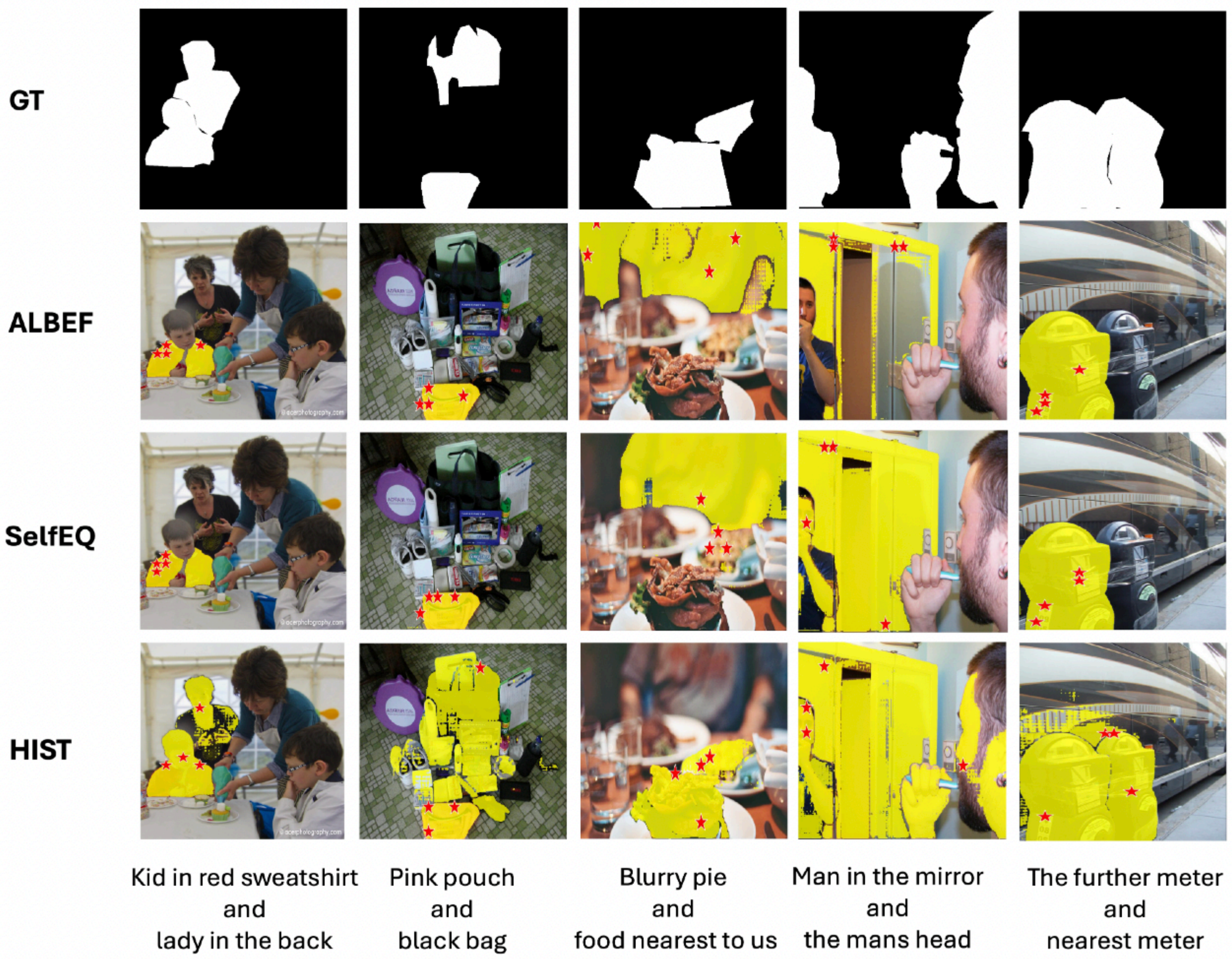
Rayat Hossain

Subject Level — items — bear

Phrase Level — items on display for sale in store window — a tan big stuffed and sitting bear

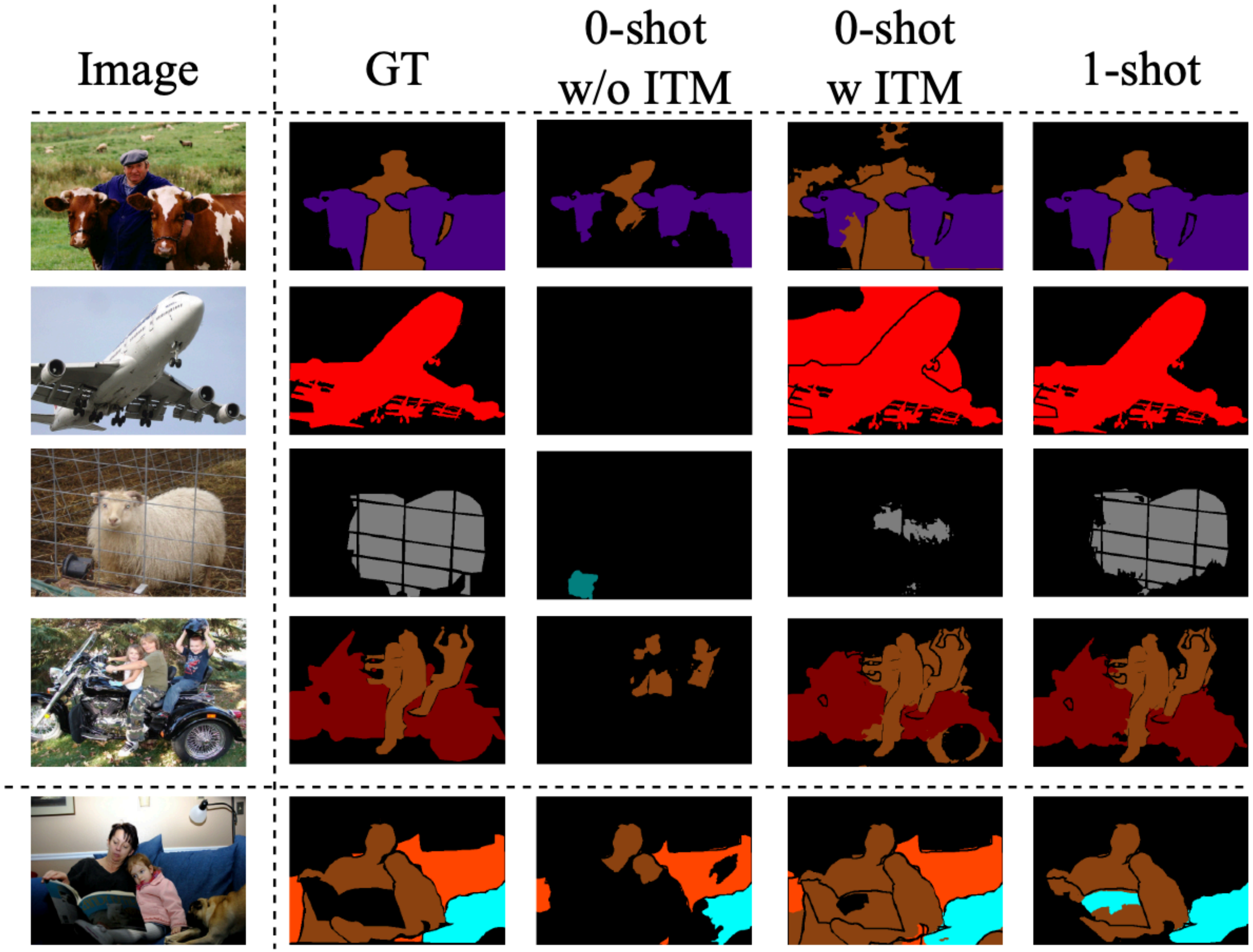
Combined Phrases Level — items on display for sale in store window **and** a tan big stuffed and sitting bear

Segmentation …



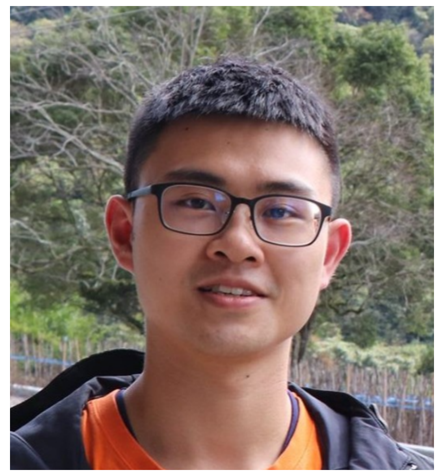
Method	Training	Flickr30K	Referit
InfoGround [10]	VG-boxes	76.7	-
VMRM [8]	VG-boxes	81.1	-
AMC [41]	VG-boxes	86.6	73.2
ALBEF [21]	Zero-shot	79.4	59.7
BLIP [22]	Zero-shot	80.1	51.6
g [33]	VG	75.6	66.0
g++ [32]	VG	80.0	**70.3**
ALBEF + SelfEQ† [12]	VG	81.7	67.0
ALBEF + SelfEQ [12]	VG	<u>81.9</u>	67.4
(Ours) ALBEF + HIST	VG	**83.6**	<u>69.5</u>
(Ours) BLIP + HIST	VG	81.5	58.1
g [33]	COCO	75.4	61.0
g++ [32]	COCO	78.1	61.5
ALBEF + SelfEQ† [12]	COCO	84.1	62.8
ALBEF + SelfEQ [12]	COCO	84.1	<u>62.8</u>
(Ours) ALBEF + HIST	COCO	**85.3**	**63.4**
(Ours) BLIP + HIST	COCO	<u>84.7</u>	57.6

Few-shot Segmentation



Rayat Hossain

Visual Program Synthesis

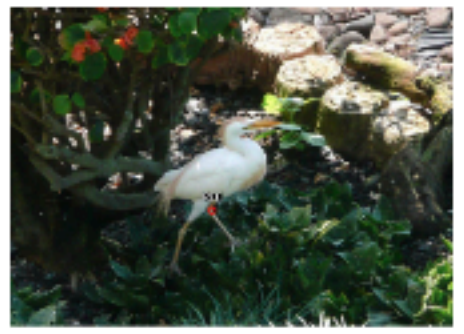


Wan-Cyuan (Chris) Fan

TASK DEFINITION: In this task, you are given a prompt and two images. In the first image, there is only one point labeled with a red circle and REF tag. In the second image, there are four points labeled with red circle and a letter tag of A, B, C, and D. You have to … the second image corresponds to the point in the first image. You may have to know where these points are to answer the question. Here are three examples of the user task.

EXAMPLES from the task:

EXAMPLE 0



TASK REQUEST PROMPT #:
<img src='...'> <img src='...'> … Which point on … (A) Point A (B) Point B (C) Point C…

The correct answer is: (C)

EXAMPLE 1

TASK REQUEST PROMPT #:
<Image> <Image> … Which point … (D) Point.

The correct answer is: (D)

⋮

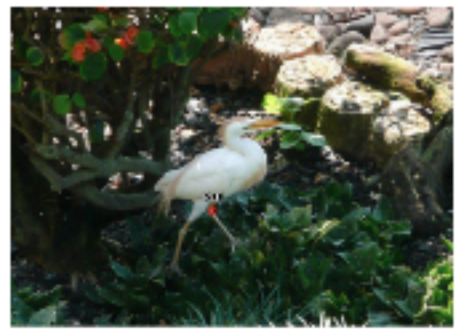
(OPTIONAL) USER CONSTRAINTS: For example, execution time need to be less than 5 sec per sample, or models with fewer than 3B parameters…

Visual Program Synthesis

TASK DEFINITION: In this task, you are given a prompt and two images. In the first image, there is only one point labeled with a red circle and REF tag. In the second image, there are four points labeled with red circle and a letter tag of A, B, C, and D. You have to … the second image corresponds to the point in the first image. You may have to know where these points are to answer the question. Here are three examples of the user task.

EXAMPLES from the task:

EXAMPLE 0




TASK REQUEST PROMPT #:
<img src='...'> <img src='...'> … Which point on … (A) Point A (B) Point B (C) Point C…

The correct answer is: (C)

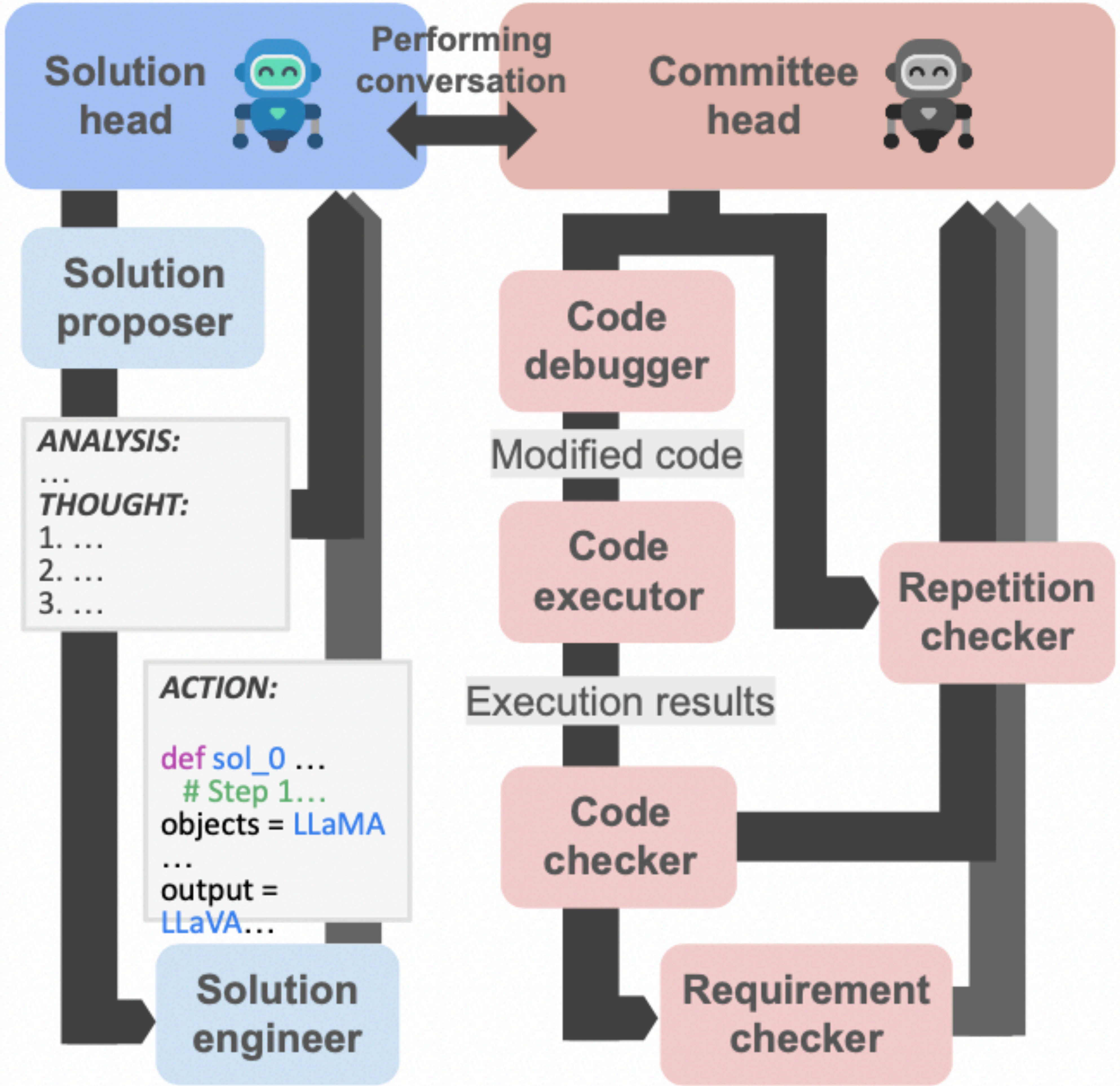
EXAMPLE 1


TASK REQUEST PROMPT #:
<Image> <Image> … Which point … (D) Point.

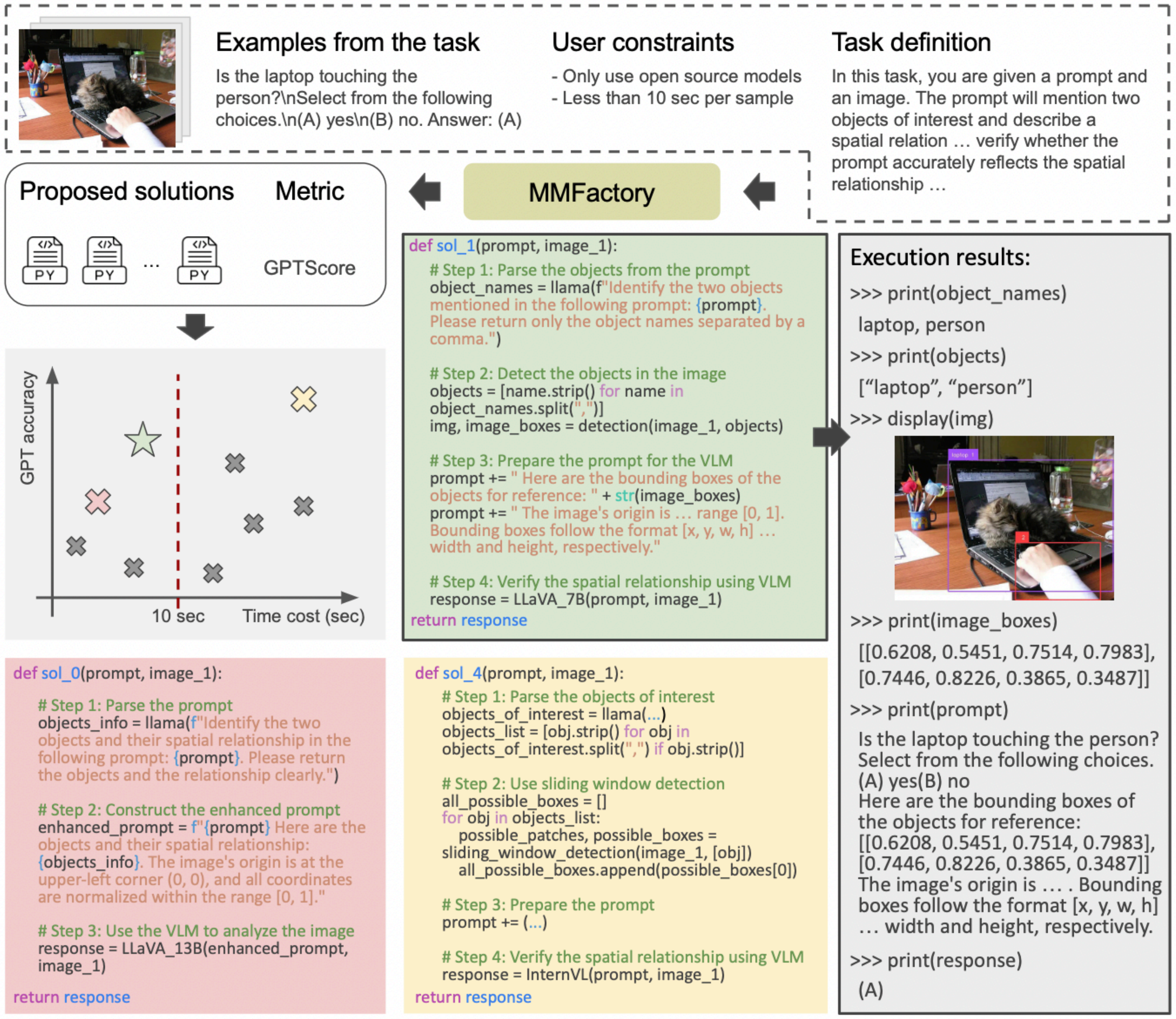
The correct answer is: (D)

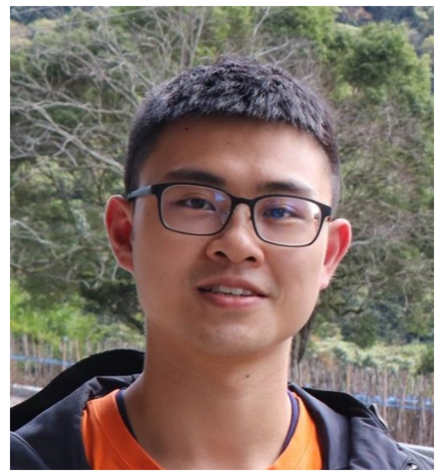
⋮

(OPTIONAL) USER CONSTRAINTS: For example, execution time need to be less than 5 sec per sample, or models with fewer than 3B parameters…



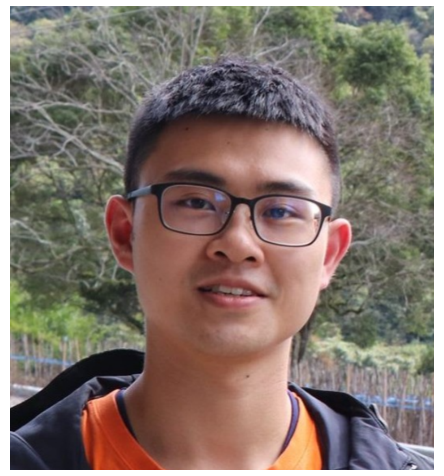
Visual Program Synthesis





Wan-Cyuan (Chris) Fan

Visual Program Synthesis



Wan-Cyuan (Chris) Fan

Method	Depth	Spatial	Jigsaw	Vis corr.	Sem. Corr.	Art	Count	Fun. Corr.	Local.	Multi-view	Refl.	Fore.	IQ	Sim.
Open-source multimodal LLMs														
OpenFlamingo-v2	54.03	43.36	47.33	25.58	30.22	52.14	21.67	**36.15**	52.00	41.35	43.28	15.91	23.33	55.15
InstructBLIP-7B	51.61	56.64	52.67	30.81	30.94	47.86	29.17	23.85	44.80	**58.65**	29.85	**29.55**	23.33	46.32
InstructBLIP-13B	51.61	65.73	52.67	29.65	32.37	50.43	30.83	22.31	52.00	54.14	**46.27**	13.64	26.00	46.32
CogVLM	50.81	67.13	52.67	20.93	23.57	49.57	46.32	23.85	43.20	57.14	26.87	24.24	26.67	46.32
LLaVA-v1.5-7B	52.42	61.54	11.33	25.58	23.02	47.86	43.33	21.54	48.80	49.62	36.57	28.03	24.00	46.32
LLaVA-v1.5-13B	53.23	67.83	58.00	29.07	32.37	47.86	**50.00**	20.77	47.20	41.35	45.52	27.27	28.00	46.32
Ours (LLaVA-7B)	51.61	**78.82**	56.67	33.14	32.37	54.70	41.23	21.54	**56.56**	55.64	37.04	26.52	23.33	**58.52**
Ours (LLaVA-13B)	**58.06**	69.93	**64.00**	**34.30**	**34.53**	**58.12**	47.24	23.85	51.64	51.13	45.06	26.52	**28.00**	45.93
API-based models														
Qwen-VL-Max	58.87	77.62	3.33	22.67	29.29	37.61	55.83	28.46	49.60	53.38	**49.25**	47.73	22.00	51.47
Gemini Pro	50.00	67.13	54.00	37.21	22.14	49.57	**65.00**	32.31	46.40	41.35	46.27	45.45	27.33	55.88
Claude 3 OPUS	57.26	57.34	32.67	31.40	20.71	60.68	49.17	22.31	46.40	57.89	27.61	62.12	21.33	**70.59**
GPT-4o	74.19	69.23	55.33	75.00	53.96	**82.91**	51.67	39.23	56.00	**60.15**	38.81	**85.61**	**30.00**	65.44
GPT-4o (+ SoM + orig.)	75.0	82.5	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4o (+ Visprog)	46.8	37.8	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4o (+ Sketchpad)	**83.9**	81.1	70.7	80.8	**58.3**	77.19	66.67	42.11	65.44	45.61	33.13	78.95	22.81	84.21
Ours (GPT-4o)	80.25	**81.82**	**75.33**	**85.47**	58.27	83.01	61.67	**55.38**	**59.02**	60.20	35.07	84.82	28.67	75.32

CVPR 2025

The **IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)** is the premier annual computer vision event comprising the main conference and several co-located workshops and short courses. With its high quality and low cost, it provides an exceptional value for students, academics and industry researchers.

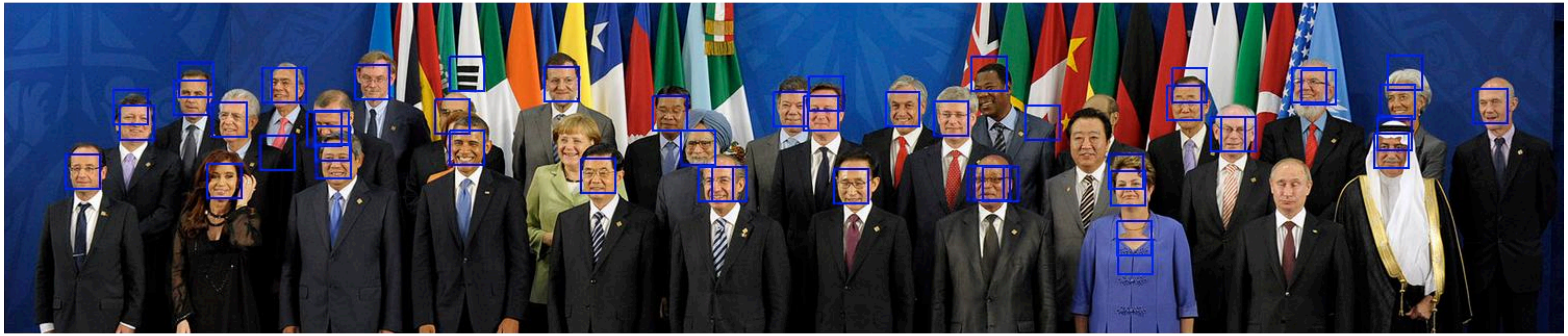
Important Dates

Paper Submission Deadline	Nov 14 '24 11:59 PM PST *	00 weeks 00 days 10:27:52
Supplementary Materials Deadline	Nov 21 '24 11:59 PM PST *	01 weeks 00 days 10:27:52
Reviews Released	Jan 23 '25 01:59 AM CST *	
Rebuttal Period Ends	Jan 30 '25 01:59 AM CST *	
Final Decisions	Feb 26 '25 01:59 AM CST *	
All dates	Timezone: America/Vancouver	

Object **Recognition / Detection**

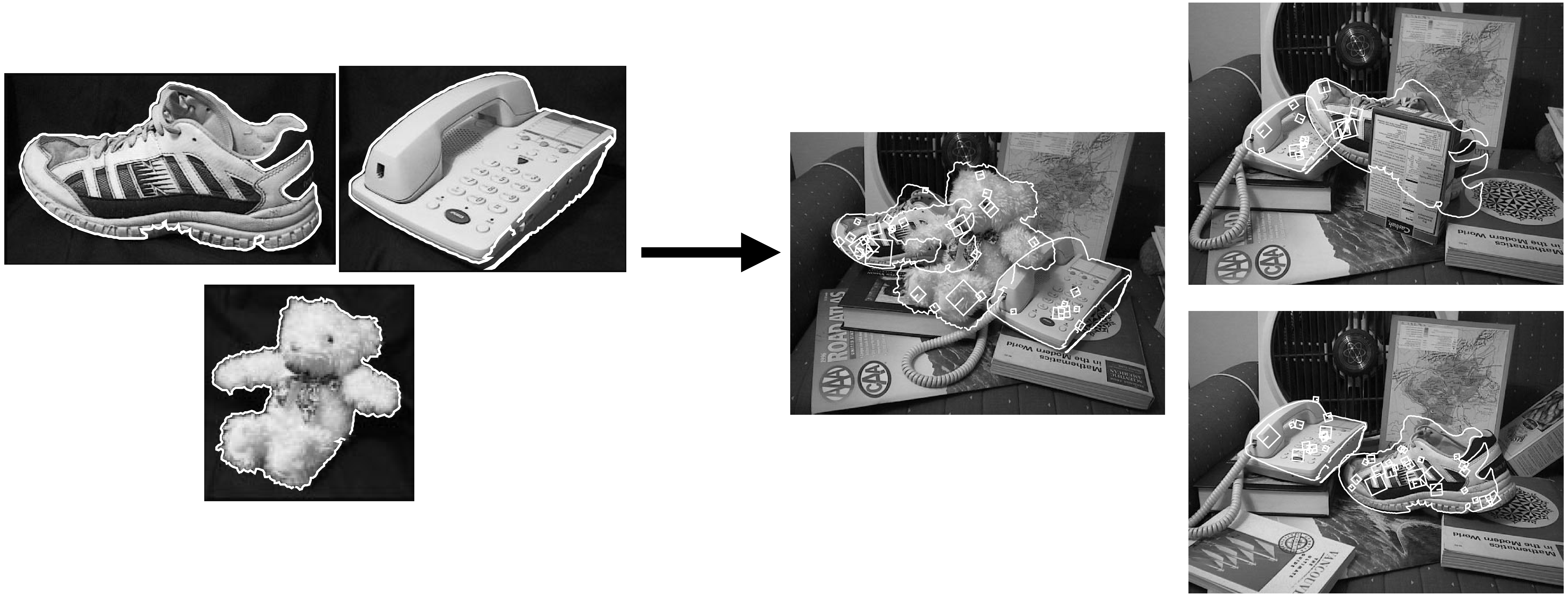
Template matching …

*  → Non-max suppress + threshold →



Object **Recognition / Detection**

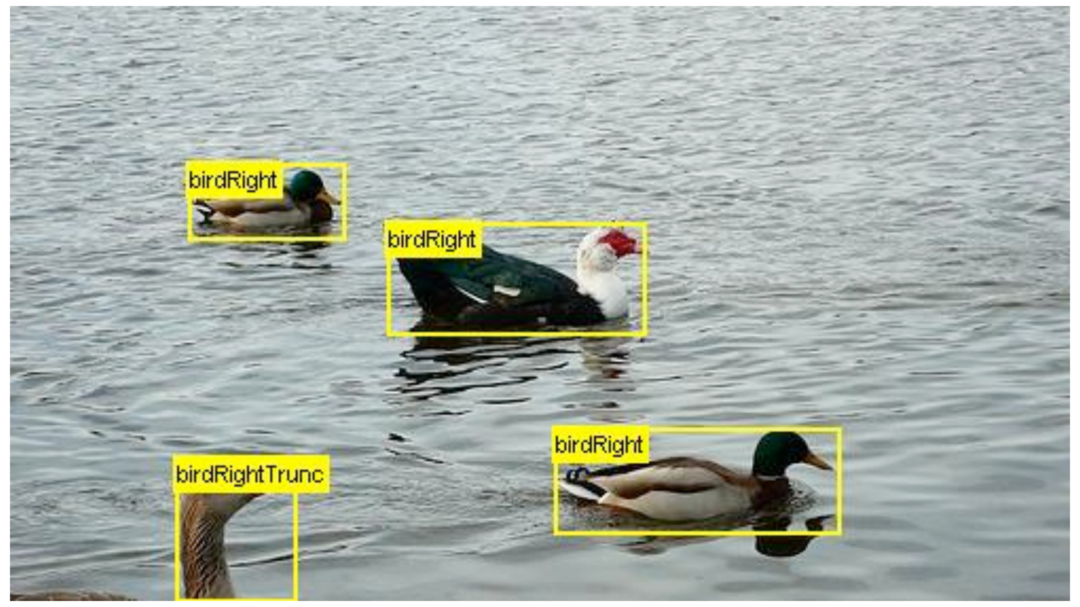
Object recognition with SIFT features and RANSAC [Lowe 1999]

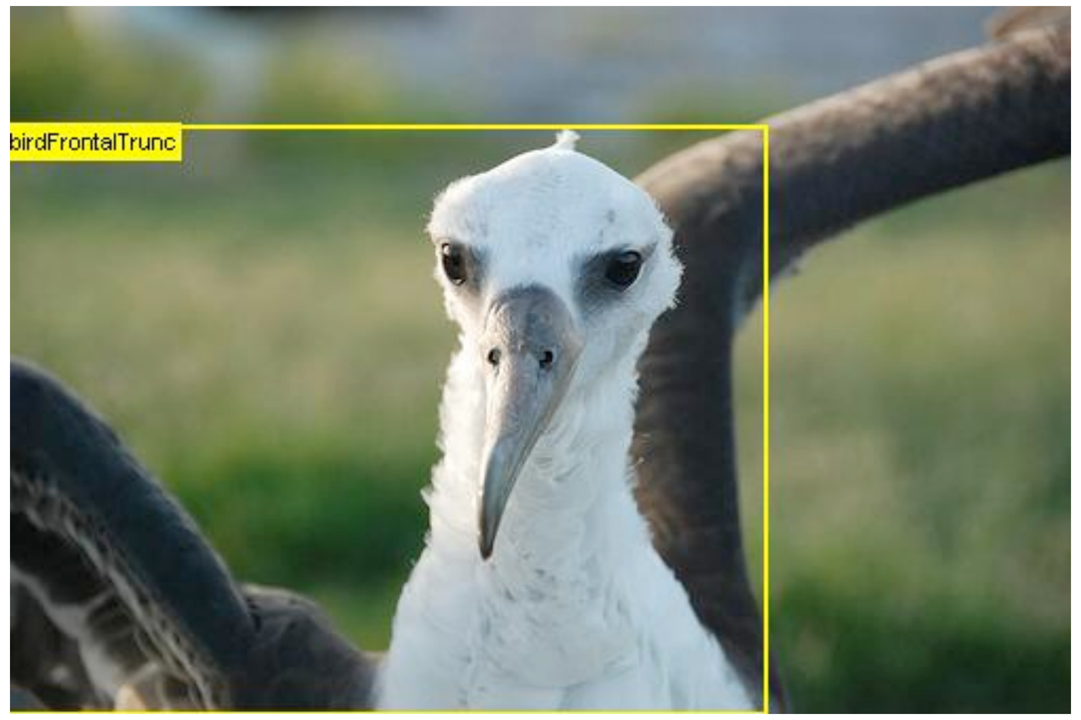


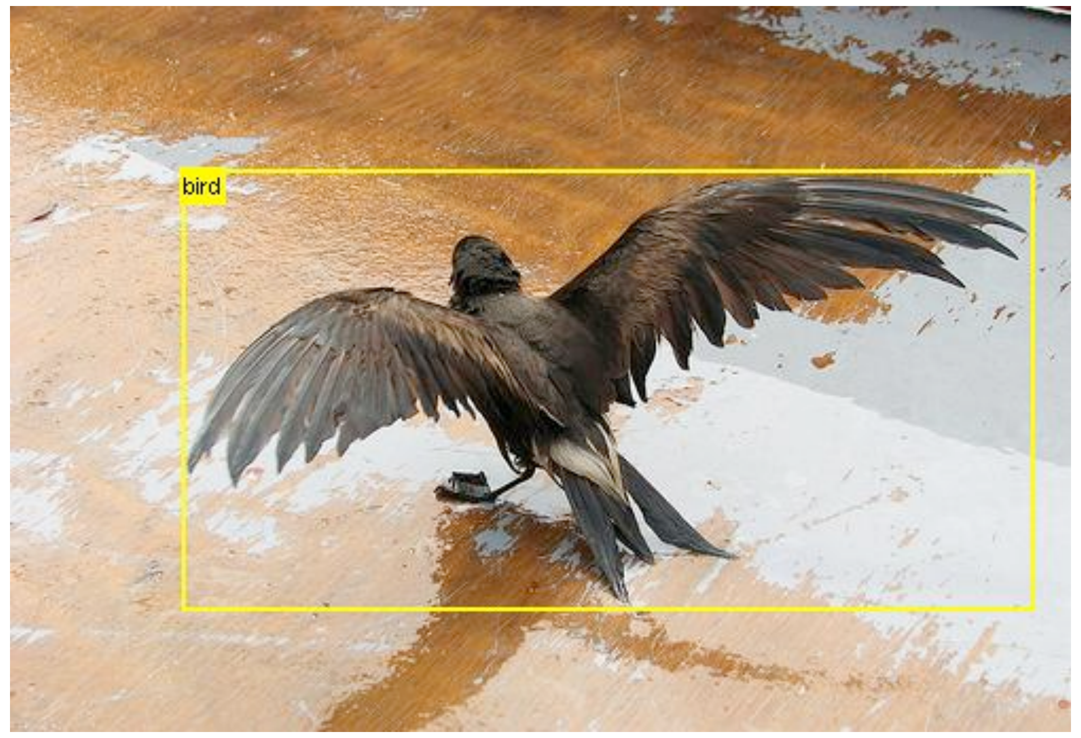
What is present? Where? What orientation?

Object **Recognition / Detection**

PASCAL Visual Object Classes Challenges [2005-2012]



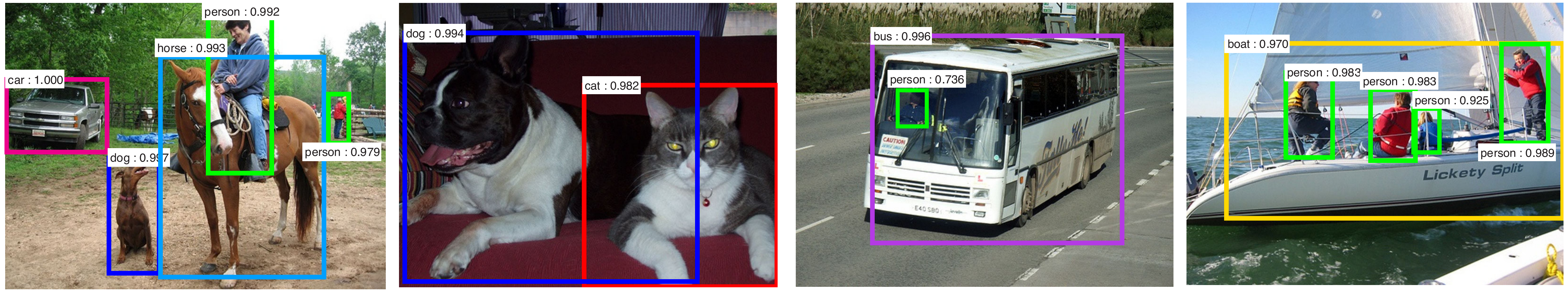




What is present? Where? What orientation?

Object **Classification** and **Detection**

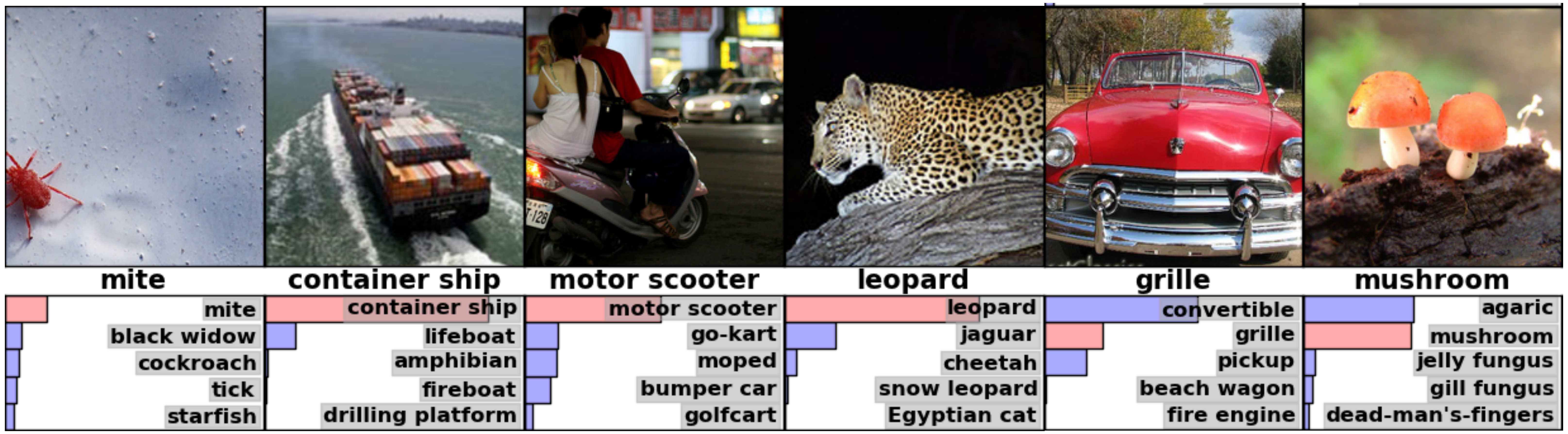
Detection: Label per region, e.g., PASCAL VOC



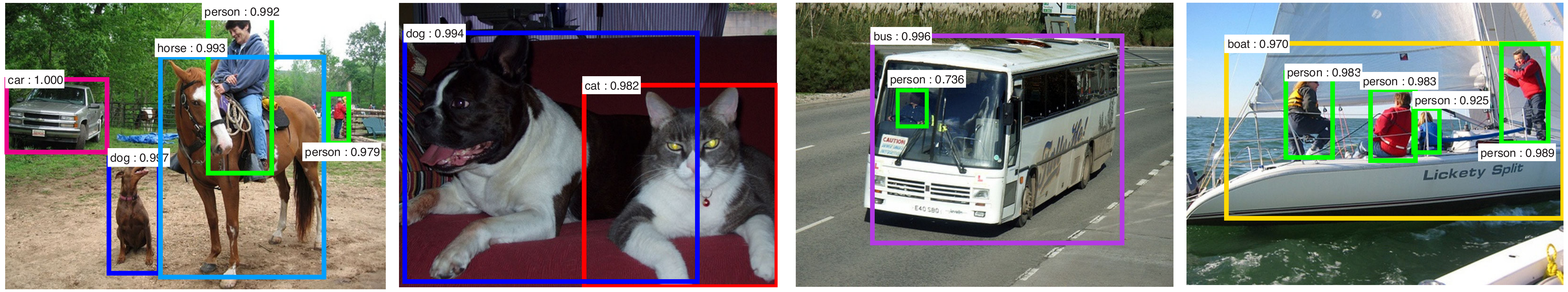
[Krizhevsky et al 2011][Ren et al 2016]

Object **Classification** and **Detection**

Classification: Label per image, e.g., ImageNet



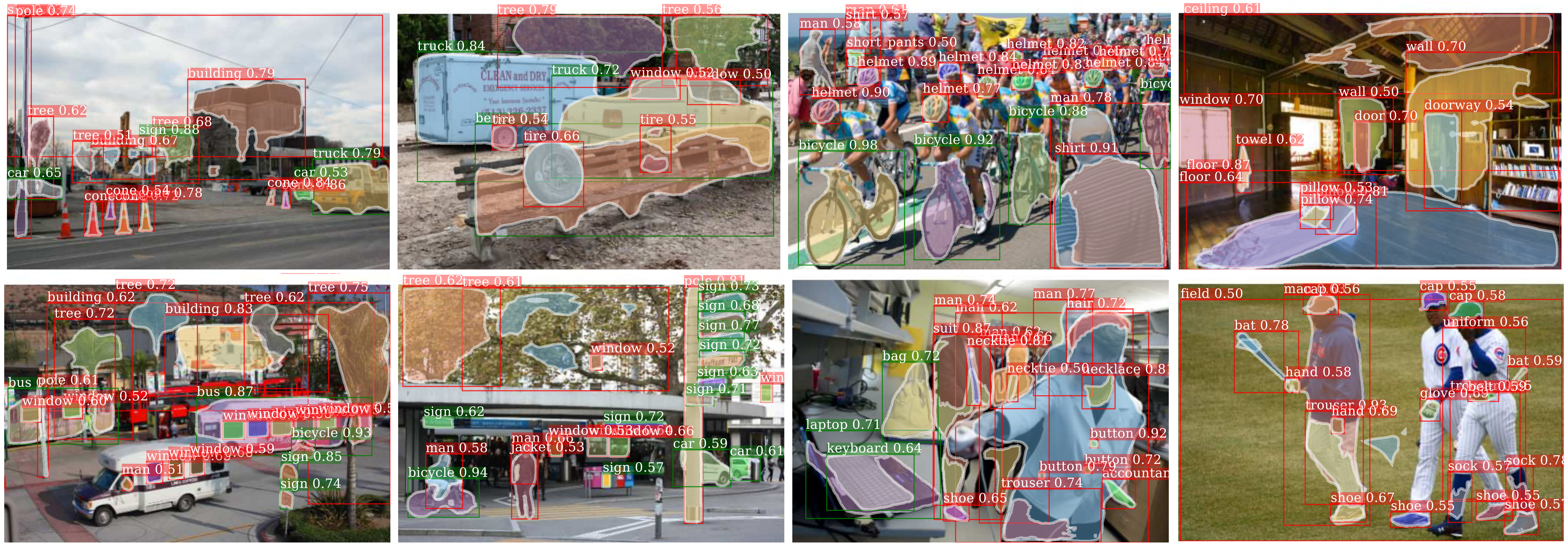
Detection: Label per region, e.g., PASCAL VOC



[Krizhevsky et al 2011][Ren et al 2016]

Segmentation

Segmentation: Label per pixel, e.g., MS COCO

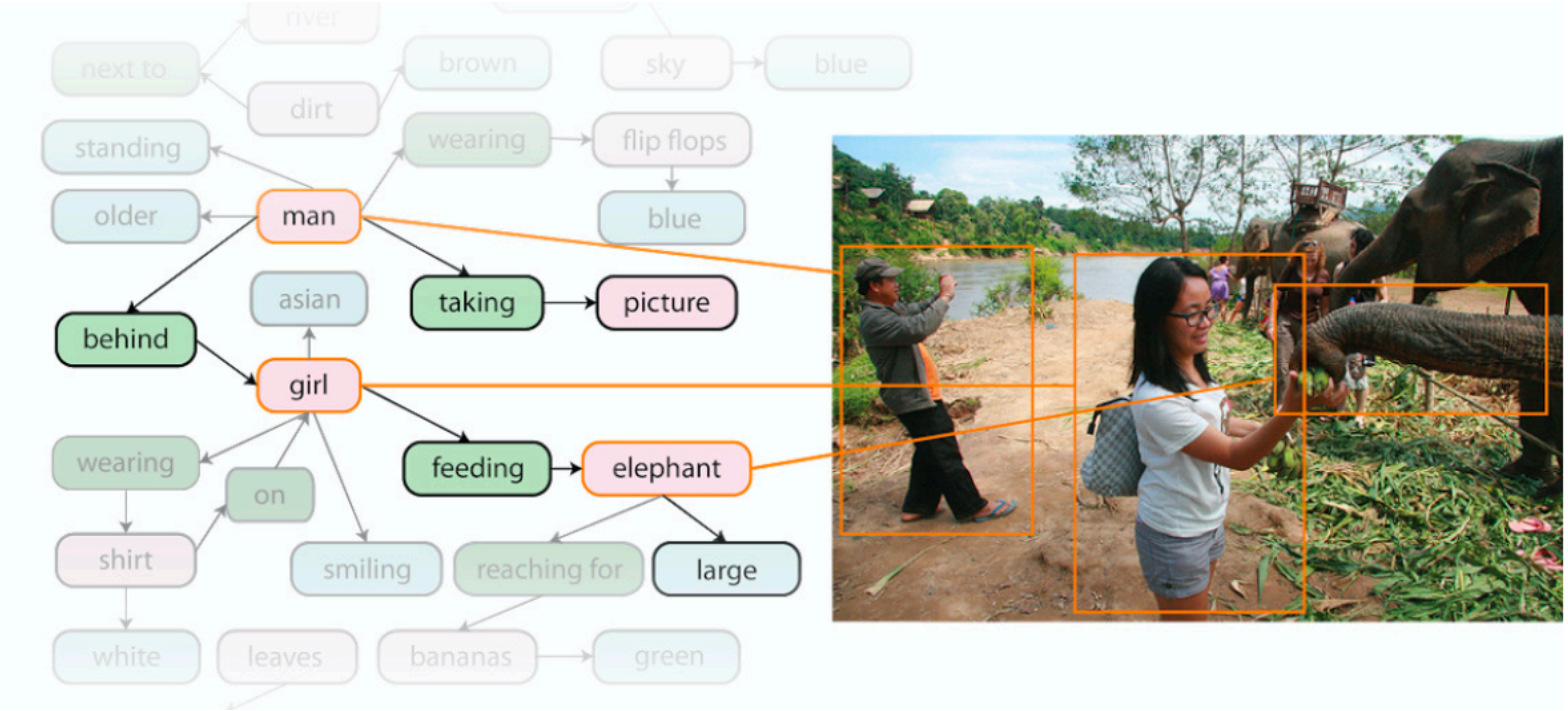


[Hu et al 2017]

Structured **Image Understanding**

"Girl feeding large elephant"

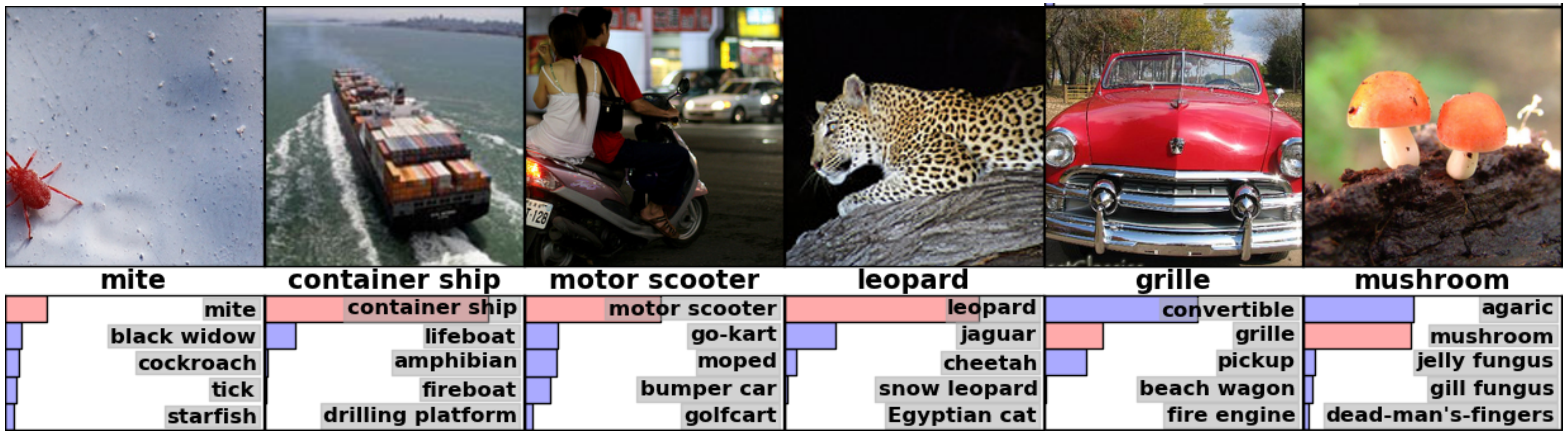
"A man taking a picture behind girl"



visualgenome.org [Krishna et al 2017]

Object **Classification**

Classification: Label per image, e.g., ImageNet



[Krizhevsky et al 2011][Ren et al 2016]

Classification: **Instance** vs. **Category**

Instance of Aeroplane (Wright Flyer)

Category of Aeroplane

[Caltech 101]

Classification: **Instance** vs. **Category**

Instance of a cat

Category of domestic cats

Taxonomy of Cats

- ↳ Mammals (Class Mammalia)
  - ↳ Therians (Subclass Theria)
    - ↳ Placental Mammals (Infraclass Placentalia)
      - ↳ Ungulates, Carnivorans, and Allies (Superorder Laurasiatheria)
        - ↳ Carnivorans (Order Carnivora)
          - ↳ Felines (Family Felidae)
            - ↳ Small Cats (Subfamily Felinae)
              - ↳ Genus *Felis*
                - ↳ Chinese Mountain Cat (*Felis bieti*)
                - ↳ Domestic Cat (*Felis catus*)
                - ↳ Jungle Cat (*Felis chaus*)
                - ↳ African Wildcat (*Felis lybica*)
                - ↳ Sand Cat (*Felis margarita*)
                - ↳ Black-footed Cat (*Felis nigripes*)
                - ↳ European Wildcat (*Felis silvestris*)

Bengal Tiger
[Omveer Choudhary]

Ocelot
[Jitze Couperus]

European Wildcat
[the wasp factory]

[inaturalist.org]

WordNet

We can use **language** to organize **visual categories**

This is the approach taken in **ImageNet** [Deng et al 2009], which uses the WordNet lexical database [wordnet.princeton.edu]

As in **language**, visual categories have **complex relationships**

e.g., a "sail" is part of a "sailboat" which is a "watercraft"

- S: (n) **sailboat**, sailing boat (a small sailing vessel; usually with a single mast)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) catboat (a sailboat with a single mast set far forward)
    - S: (n) sharpie (a shallow-draft sailboat with a sharp prow, flat bottom, and triangular sail; formerly used along the northern Atlantic coast of the United States)
    - S: (n) trimaran (a fast sailboat with 3 parallel hulls)
  - *part meronym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) sailing vessel, sailing ship (a vessel that is powered by the wind; often having several masts)

WordNet

We can use **language** to organize **visual categories**

This is the approach taken in **ImageNet** [Deng et al 2009], which uses the WordNet lexical database [wordnet.princeton.edu]

As in **language**, visual categories have **complex relationships**

e.g., a "sail" is part of a "sailboat" which is a "watercraft"

- S: (n) **sailboat**, sailing boat (a small sailing vessel; usually with a single mast)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) catboat (a sailboat with a single mast set far forward)
    - S: (n) sharpie (a shallow-draft sailboat with a sharp prow, flat bottom, and triangular sail; formerly used along the northern Atlantic coast of the United States)
    - S: (n) trimaran (a fast sailboat with 3 parallel hulls)
  - *part meronym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) sailing vessel, sailing ship (a vessel that is powered by the wind; often having several masts)

If we call a "**sailboat**" a **watercraft**, is this wrong? What if we call it a "**sail**"?

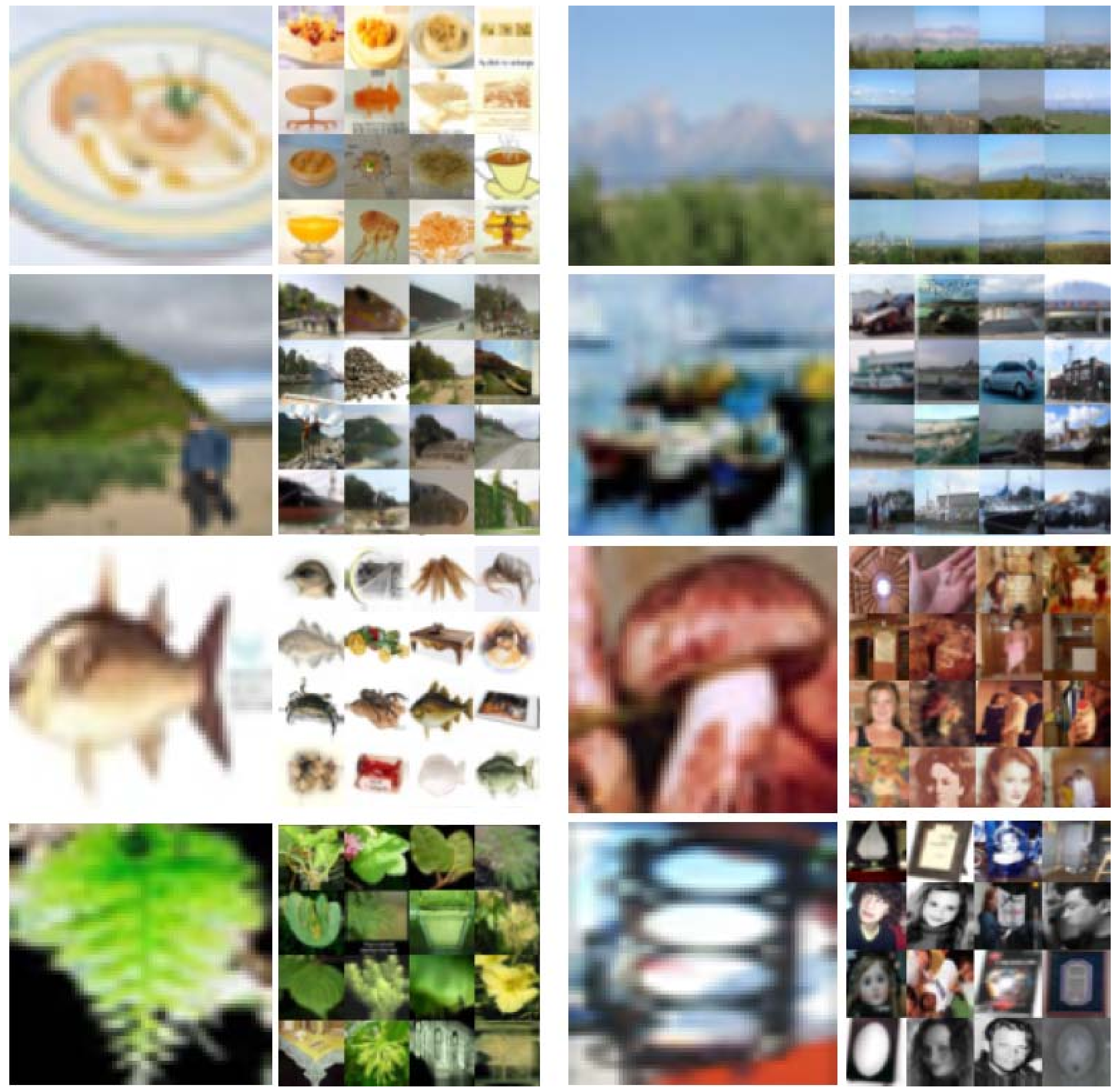
Tiny Image Dataset

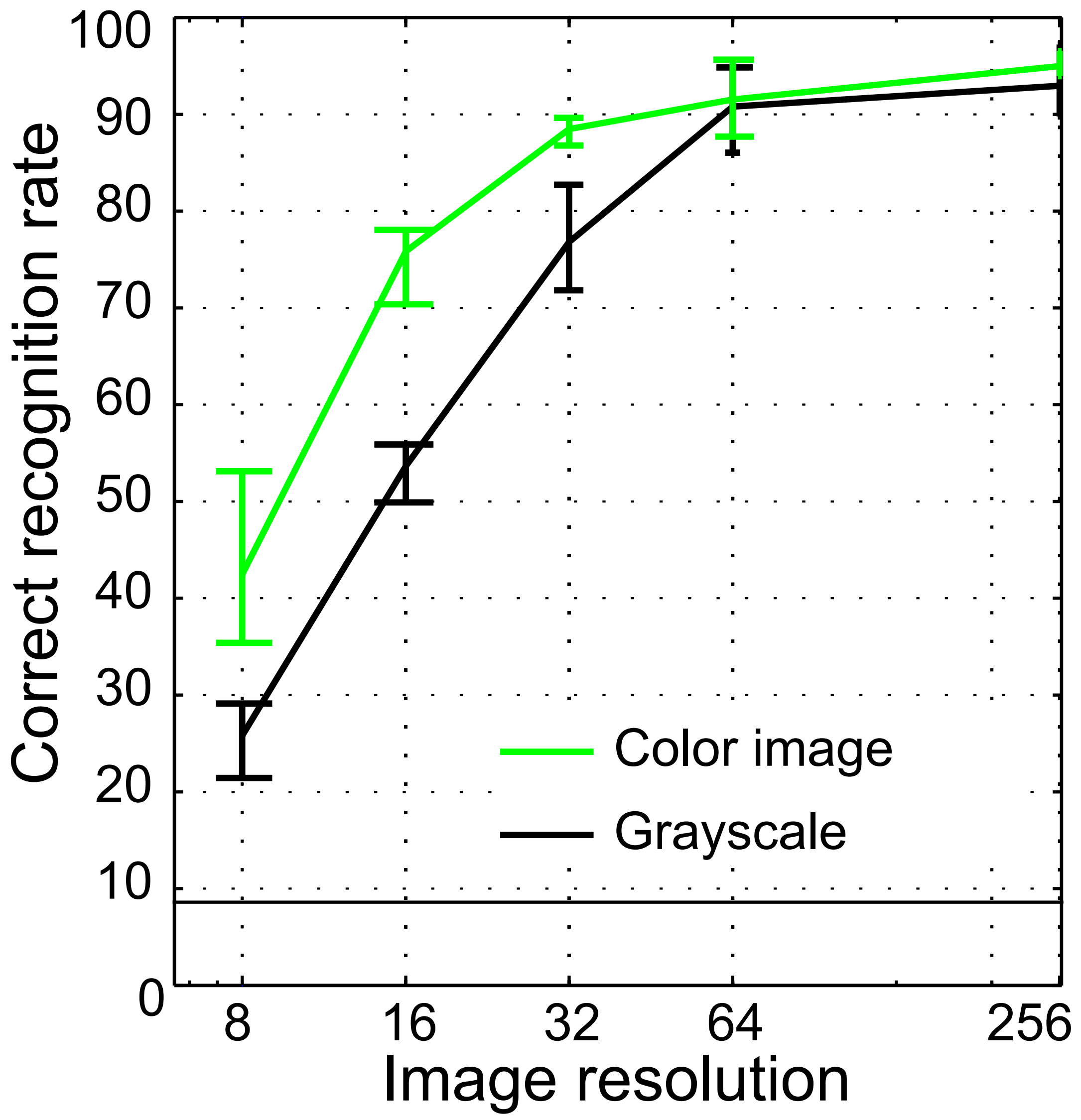
Precursor to ImageNet and CIFAR10/100

80 million images collected via image search circa 2008 using 75,062 noun synsets from WordNet (labels are noisy)

Very small images (32x32xRGB) used to minimise storage

Note human performance is still quite good at this scale!





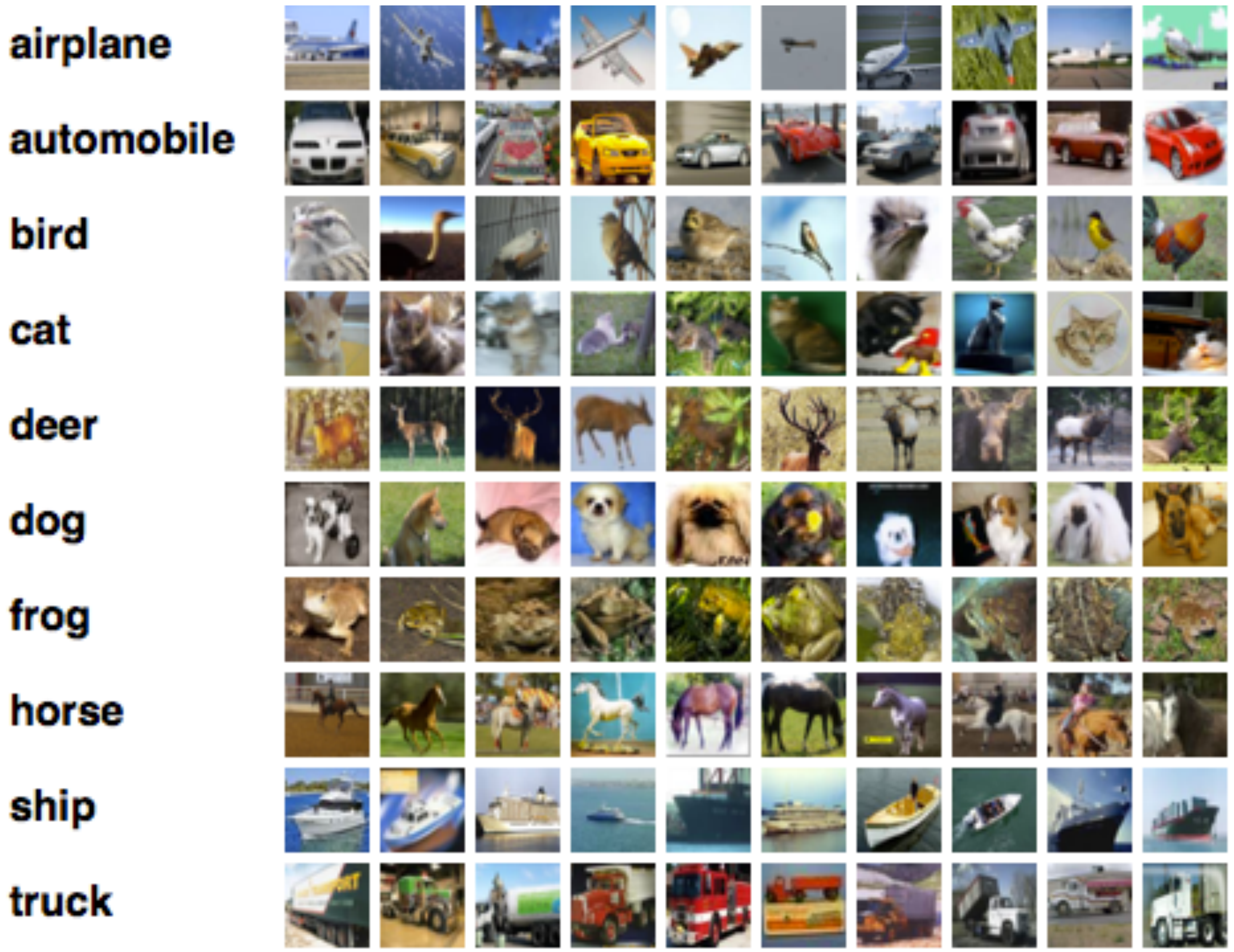
a) Scene recognition

[Torralba Freeman Fergus 2008]

CIFAR10 Dataset

Hand labelled set of 10 categories from Tiny Images dataset

60,000 32x32 images in 10 classes (50k train, 10k test)



Good test set for visual recognition problems

Classification

Problem:

Assign new observations into one of a fixed set of categories (classes)

Key Idea(s):

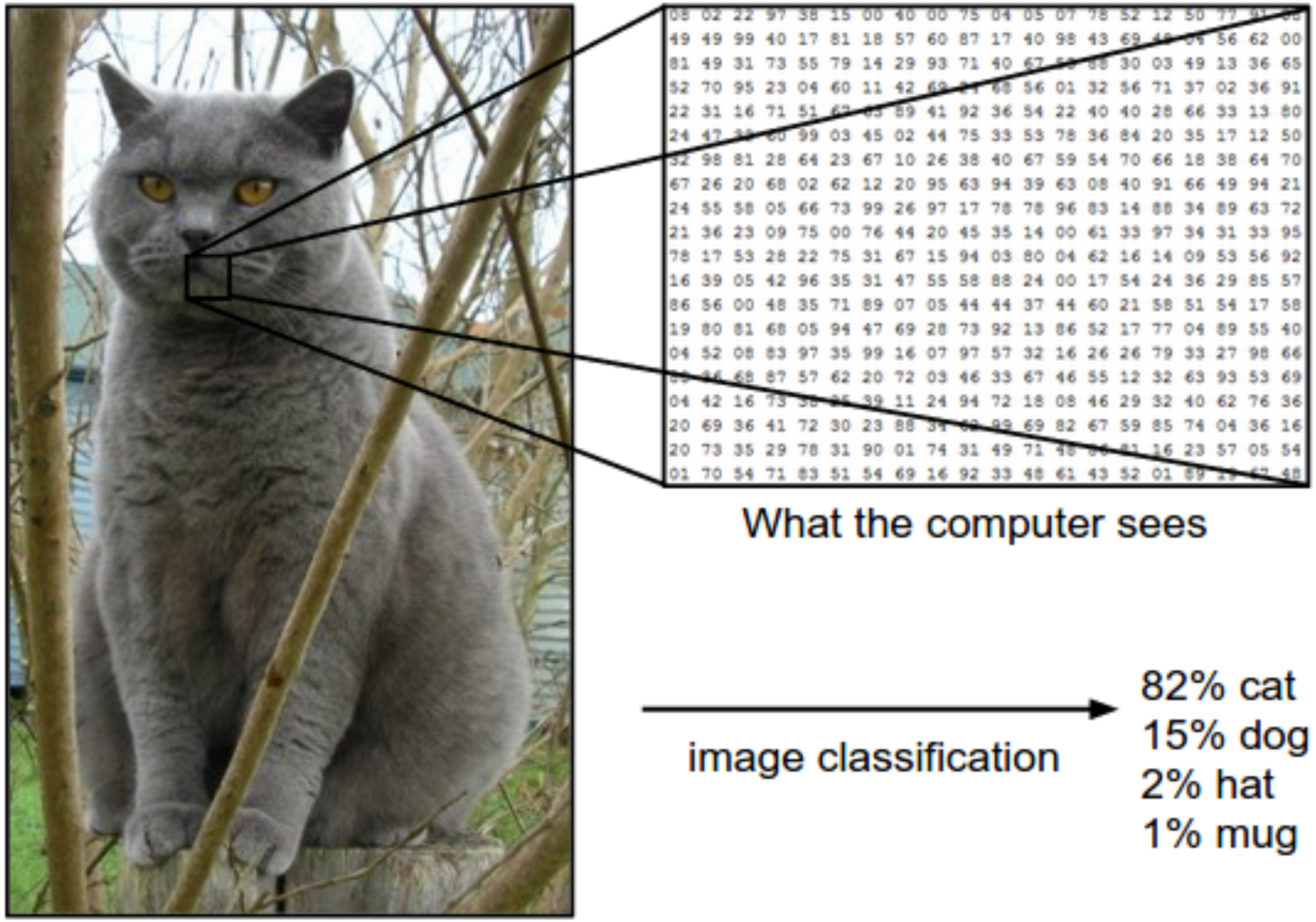
Build a model of data in a given category based on observations of instances in that category

Classification

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

$\longrightarrow$ cat

Classification



Classification

A **classifier** is a procedure that accepts as input a set of features and outputs a class **label** (probability over class labels)

Classification

A **classifier** is a procedure that accepts as input a set of features and outputs a class **label** (probability over class labels)

Classifiers can be **binary** (face vs. not-face) or **multi-class** (cat, dog, horse, ...).

Binary: $[0]/[1]$

Multi-class: $[1, 0, 0, 0, ...]$ (one-hot)

$[\ 9\ ]$ (label)

Classification

A **classifier** is a procedure that accepts as input a set of features and outputs a class **label** (probability over class labels)

Classifiers can be **binary** (face vs. not-face) or **multi-class** (cat, dog, horse, ...).

We build a classifier using a **training set** of labelled examples $\{(\mathbf{x}_i, y_i)\}$, where each $\mathbf{x}_i$ is a feature vector and each y_i is a class label.

Binary: $[0]/[1]$

Multi-class: $[1, 0, 0, 0, ...]$ (one-hot)

$[\ 9\ ]$ (label)

Classification

A **classifier** is a procedure that accepts as input a set of features and outputs a class **label** (probability over class labels)

Classifiers can be **binary** (face vs. not-face) or **multi-class** (cat, dog, horse, ...).

We build a classifier using a **training set** of labelled examples $\{(\mathbf{x}_i, y_i)\}$, where each $\mathbf{x}_i$ is a feature vector and each y_i is a class label.

Given a previously unseen observation, we use the classifier to predict its class label.

Binary: $[0]/[1]$

Multi-class: $[1, 0, 0, 0, ...]$ (one-hot)

$[\ 9\ ]$ (label)

Classification

— Collect a database of images with labels
— Use ML to train an image classifier
— Evaluate the classifier on test images

Example training set

Label

Feature vector computed from the image

Example 1: A Toy Classification Problem

Categorize images of fish
— “Atlantic salmon” vs “Pacific salmon”

Use **features** such as length, width, lightness, fin shape & number, mouth position, etc.

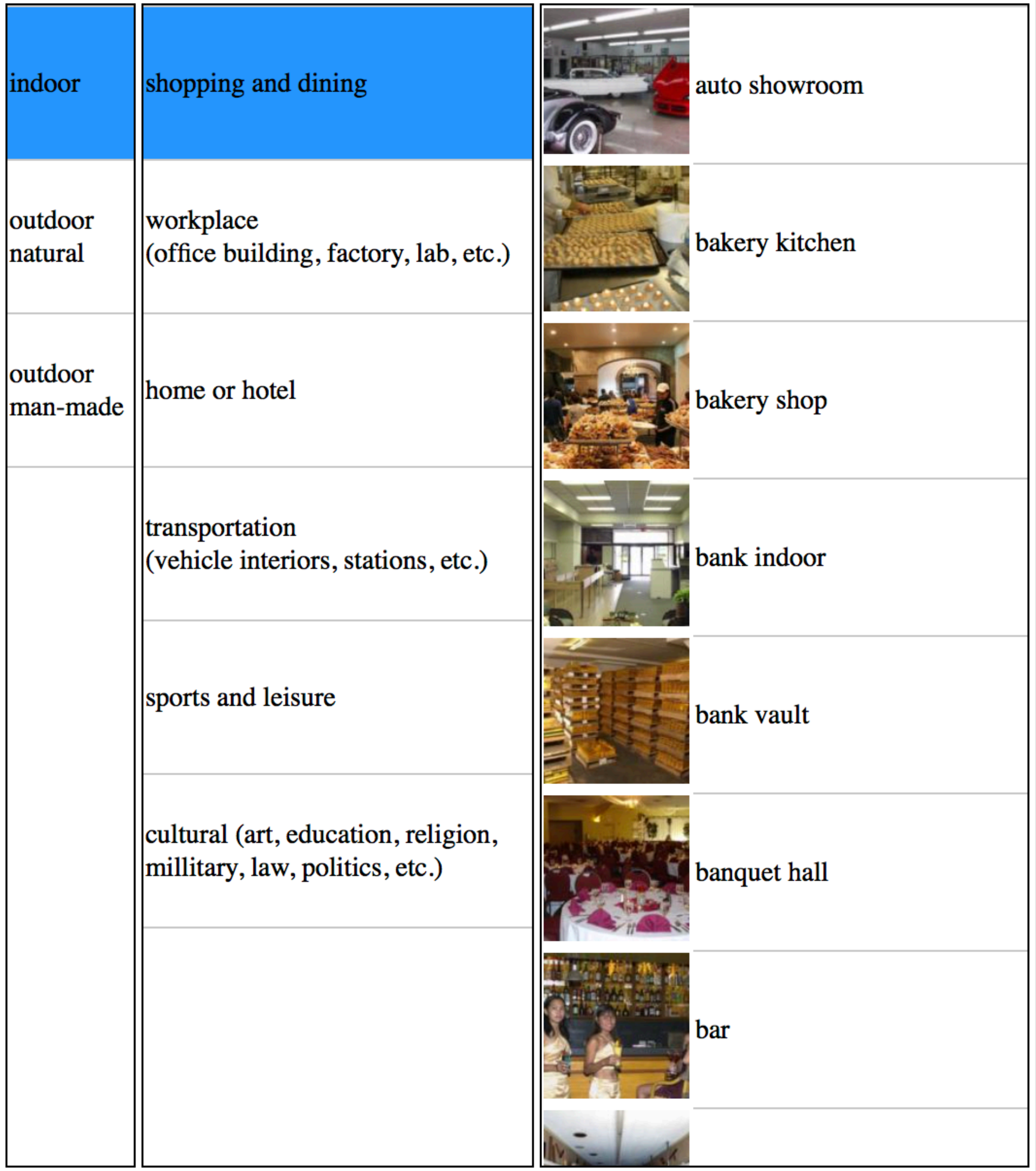
Given a previously unobserved image of a salmon, use the learned classifier to guess whether it is an Atlantic or Pacific salmon

Figure credit: Duda & Hart

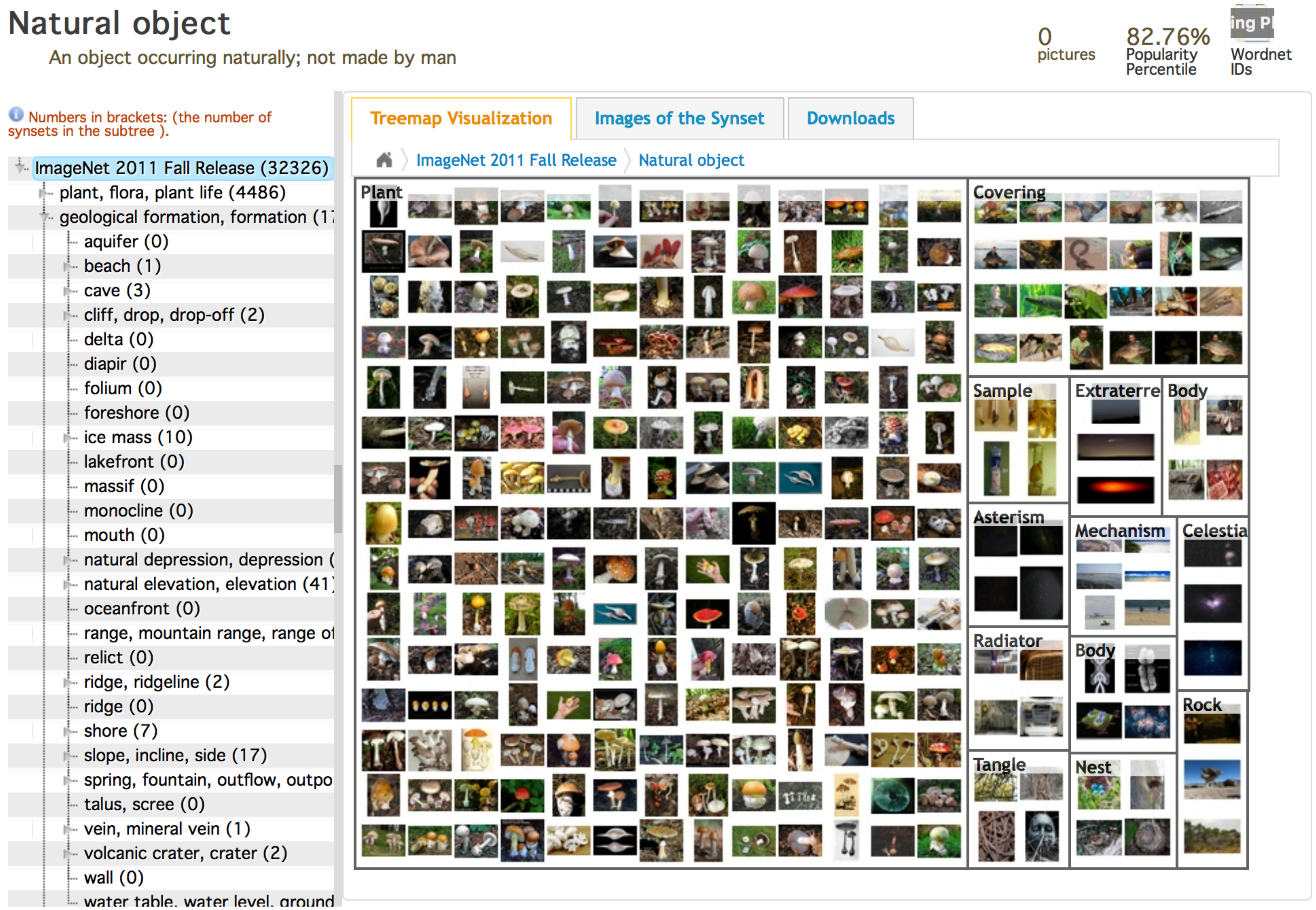
Example 2: Real Classification Problem

SUN Dataset

- 131K images
- 908 **scene** categories



Example 3: Real Classification Problem



ImageNet Dataset

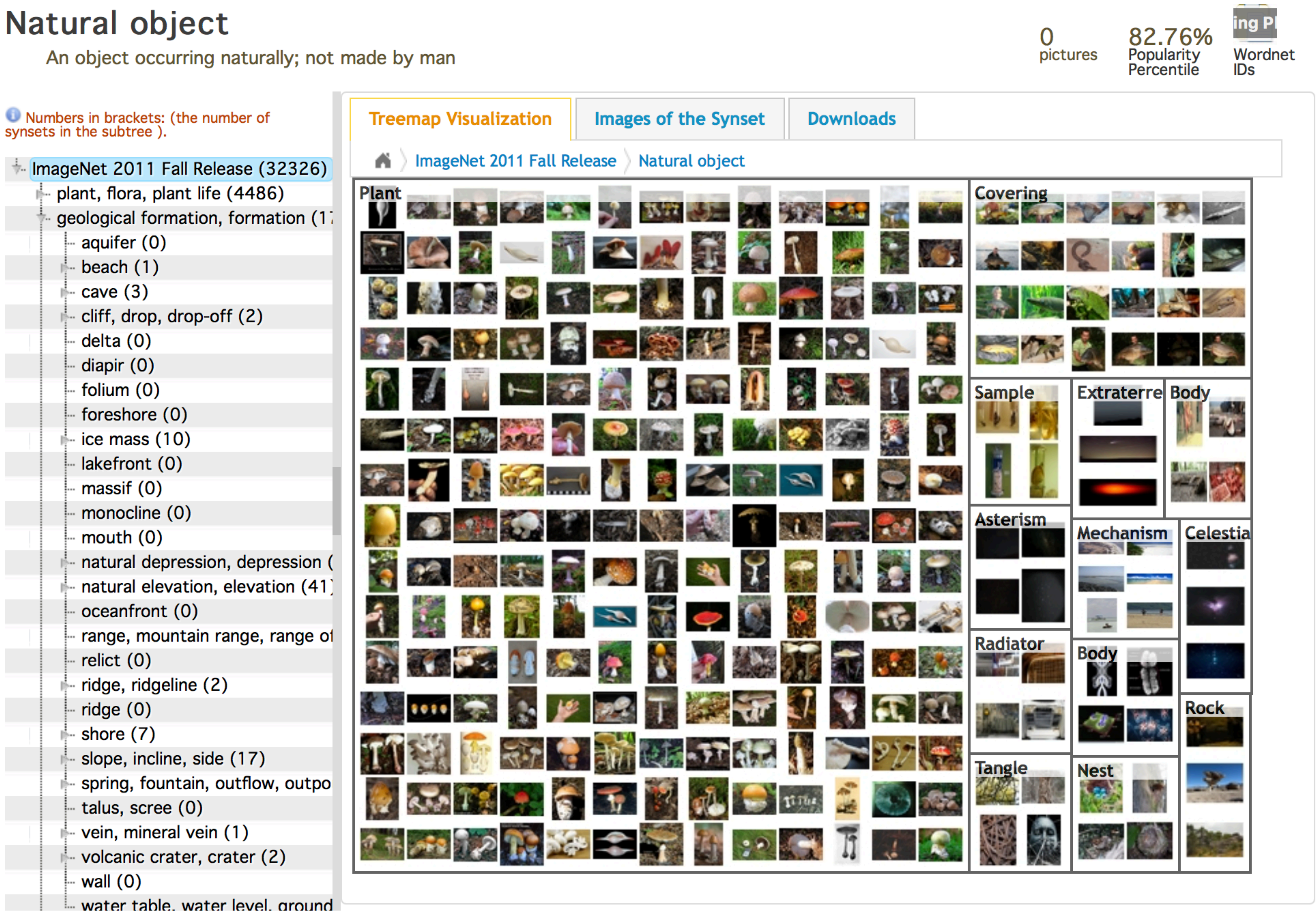
- 14 Million images
- 21K **object** categories

Example 3: Real Classification Problem

ImageNet Dataset

- 14 Million images

- 21K **object** categories

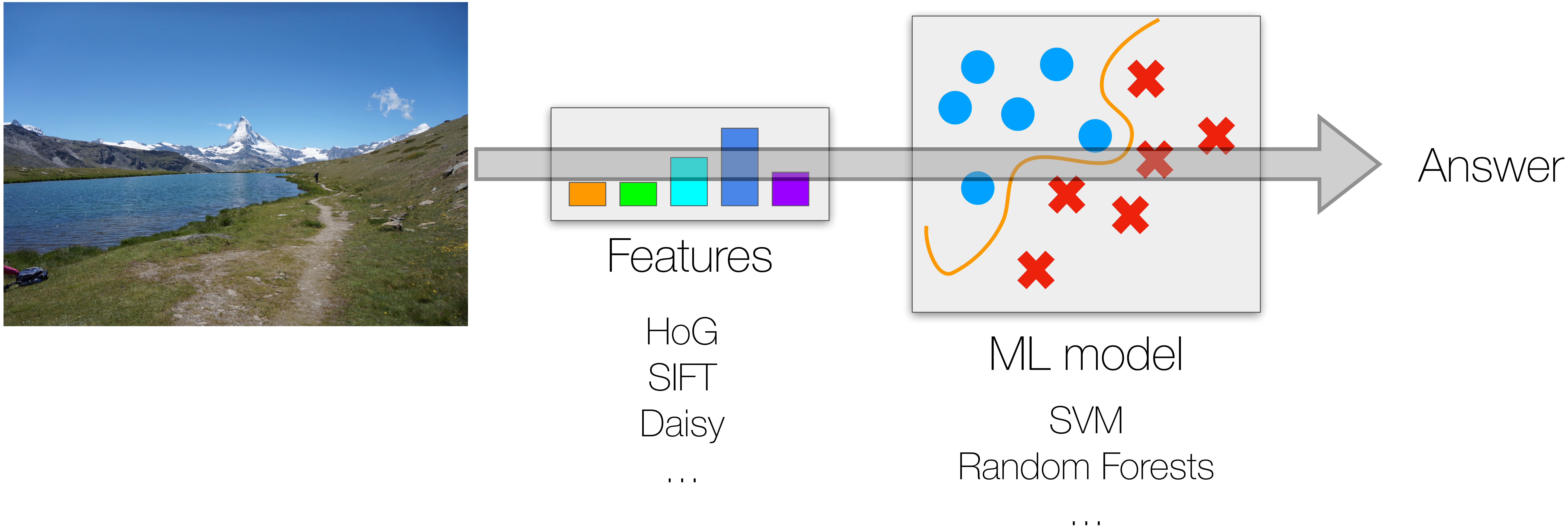


Closed-world problem

Issue: Classification assumes that incoming image belongs to one of k classes. However, in practice it is impossible to enumerate all relevant classes in the world, nor would doing so be useful. So how do we deal with images which don't belong?

Solution: Create an "unknown" or "irrelevant" class.

Traditional Image Classification Pipeline



Traditional Image Classification Pipeline

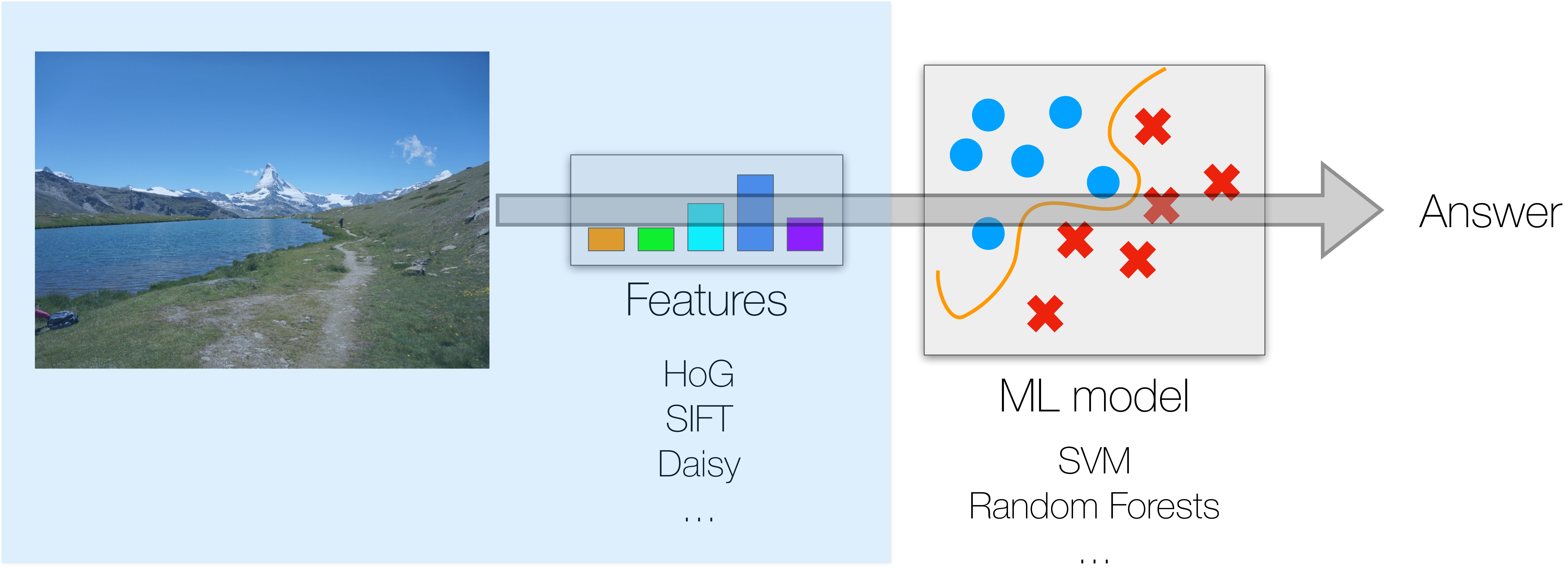


Image Classification

Representation of Images

— Image pixels directly

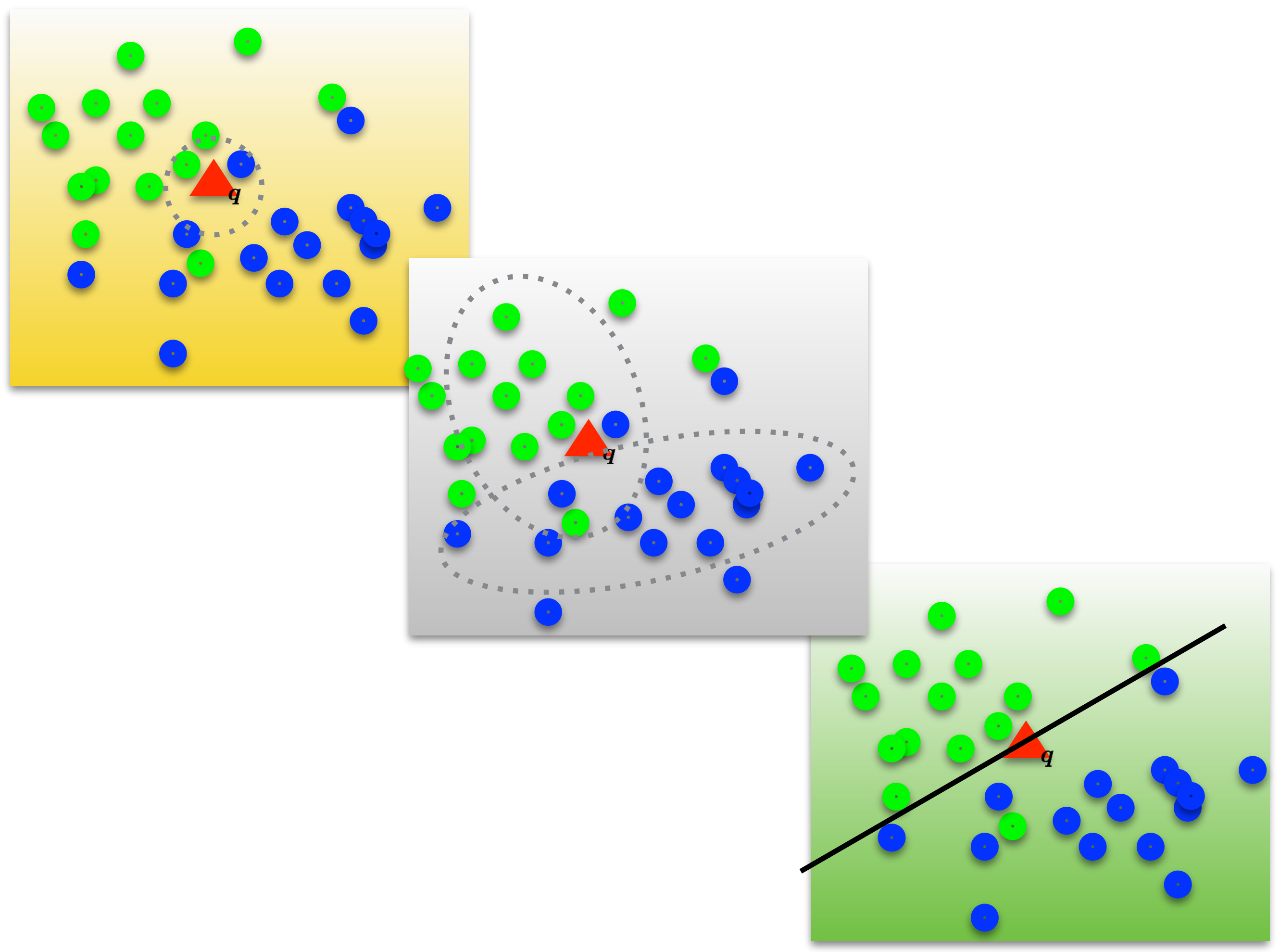
— Bag of Words

Classification Algorithms

— Bayes' Classifier

— Nearest Neighbor Classifier

— SVM Classifier



Visual **Words**

Many algorithms for image classification accumulate evidence on the basis of **visual words**.

To classify a text document (e.g. as an article on sports, entertainment, business, politics) we might find patterns in the occurrences of certain words.

Vector Space Model

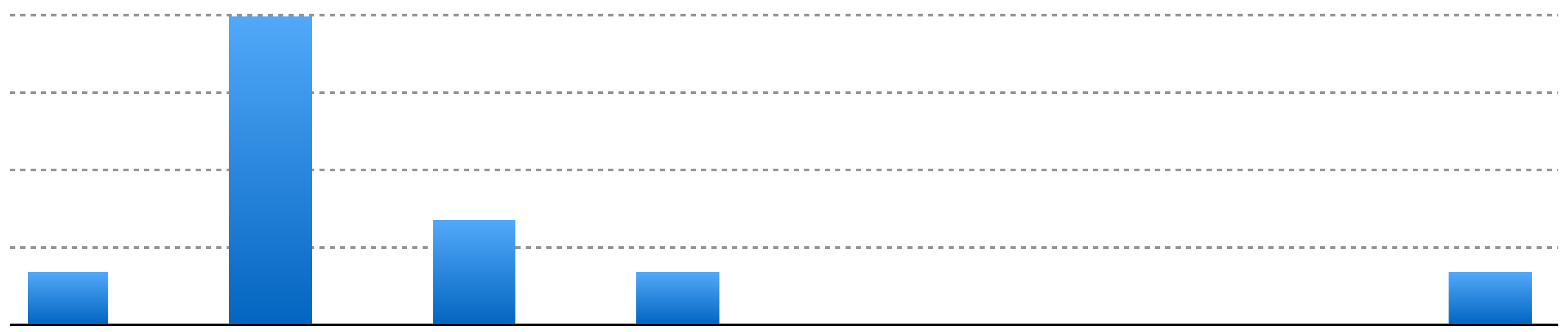
G. Salton. ‘Mathematics and Information Retrieval’ Journal of Documentation, 1979


The Newspa

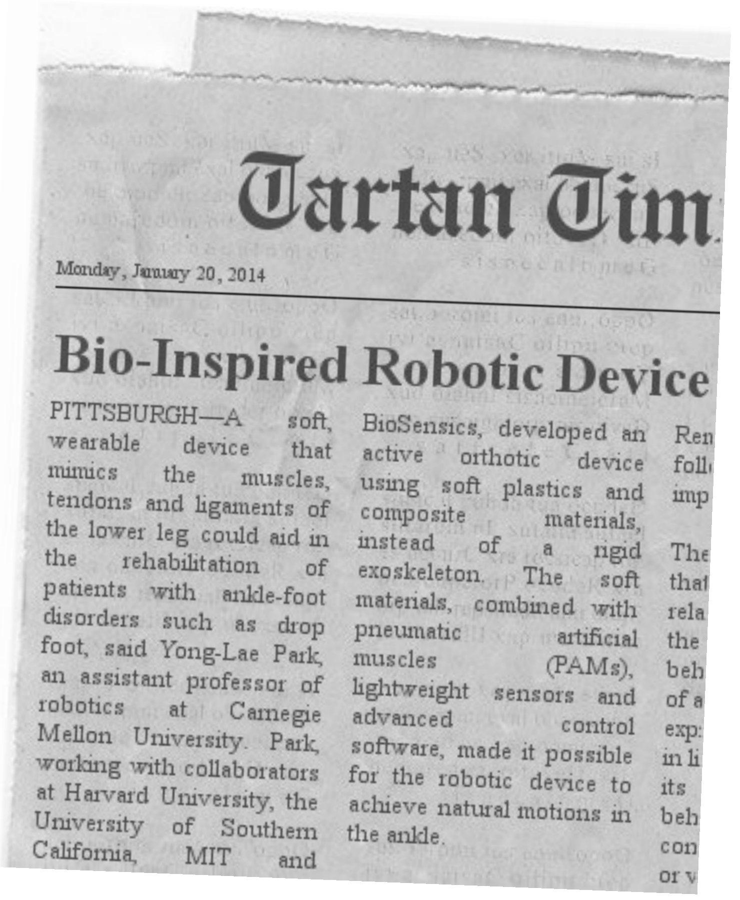
Sunday, December 22, 2013

DARPA Selects Carnegie Me

The Tartan Rescue Team from Carnegie Mellon University's National Robotics Engineering Center ranked third among teams competing in the Defense Advanced Research Projects Agency (DARPA) Robotics Challenge Trials this weekend in Homestead, Fla., and was selected by the agency as one of eight teams eligible for DARPA funding to prepare for next December's finals. The team's four-limbed CMU Highly Intelligent Mobile Platform, or CHIMP, robot scored 18 out of a possible 32 points during the two-day trials. It demonstrated its ability to perform such tasks as removing debris, cutting a hole through a wall and closing a series of valves.



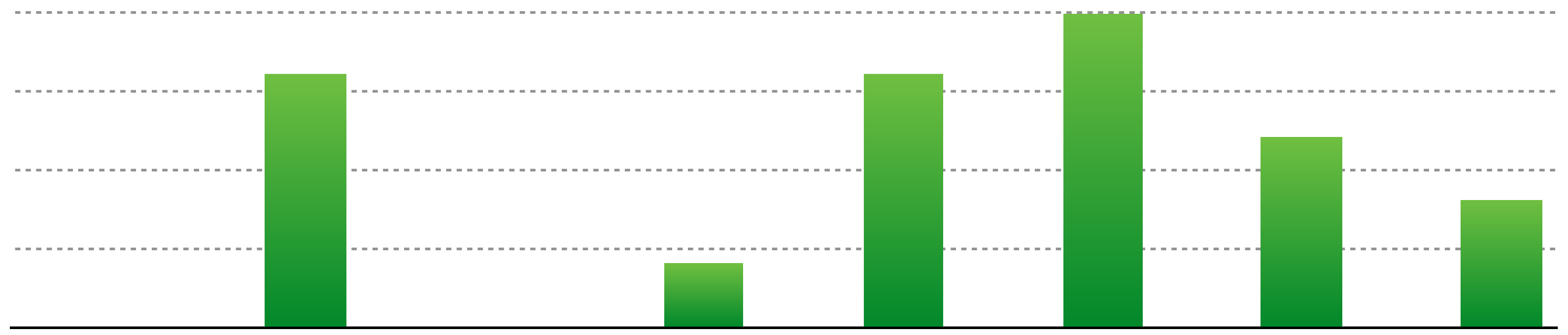
1	6	2	1	0	0	0	1
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor


Tartan Tim

Monday, January 20, 2014

Bio-Inspired Robotic Device

PITTSBURGH—A soft, wearable device that mimics the muscles, tendons and ligaments of the lower leg could aid in the rehabilitation of patients with ankle-foot disorders such as drop foot, said Yong-Lae Park, an assistant professor of robotics at Carnegie Mellon University. Park, working with collaborators at Harvard University, the University of Southern California, MIT and BioSensics, developed an active orthotic device using soft plastics and composite materials, instead of a rigid exoskeleton. The soft materials, combined with pneumatic artificial muscles (PAMs), lightweight sensors and advanced control software, made it possible for the robotic device to achieve natural motions in the ankle.



0	4	0	1	4	5	3	2
Tartan	robot	CHIMP	CMU	bio	soft	ankle	sensor

http://www.fodey.com/generators/newspaper/snippet.asp

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Vector Space Model

A document (datapoint) is a vector of counts over each word (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

$n(\cdot)$ counts the number of occurrences

just a histogram over words

What is the similarity between two documents?

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Vector Space Model

A document (datapoint) is a vector of counts over each word (feature)

$$\boldsymbol{v}_d = [n(w_{1,d}) \quad n(w_{2,d}) \quad \cdots \quad n(w_{T,d})]$$

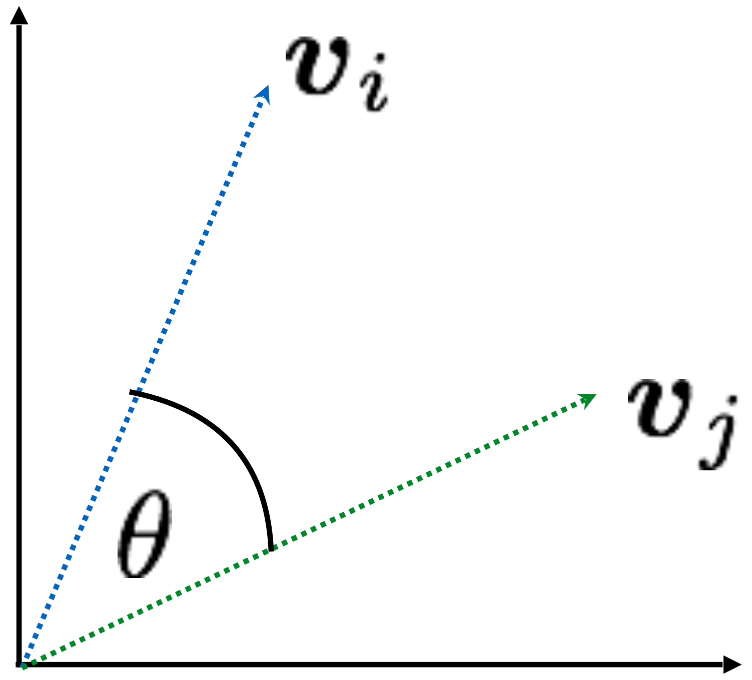
$n(\cdot)$ counts the number of occurrences

just a histogram over words

What is the similarity between two documents?

Use any distance you want but the cosine distance is fast and well designed for high-dimensional vector spaces:

$$\begin{aligned} d(\boldsymbol{v}_i, \boldsymbol{v}_j) &= \cos\theta \\ &= \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\|\|\boldsymbol{v}_j\|} \end{aligned}$$



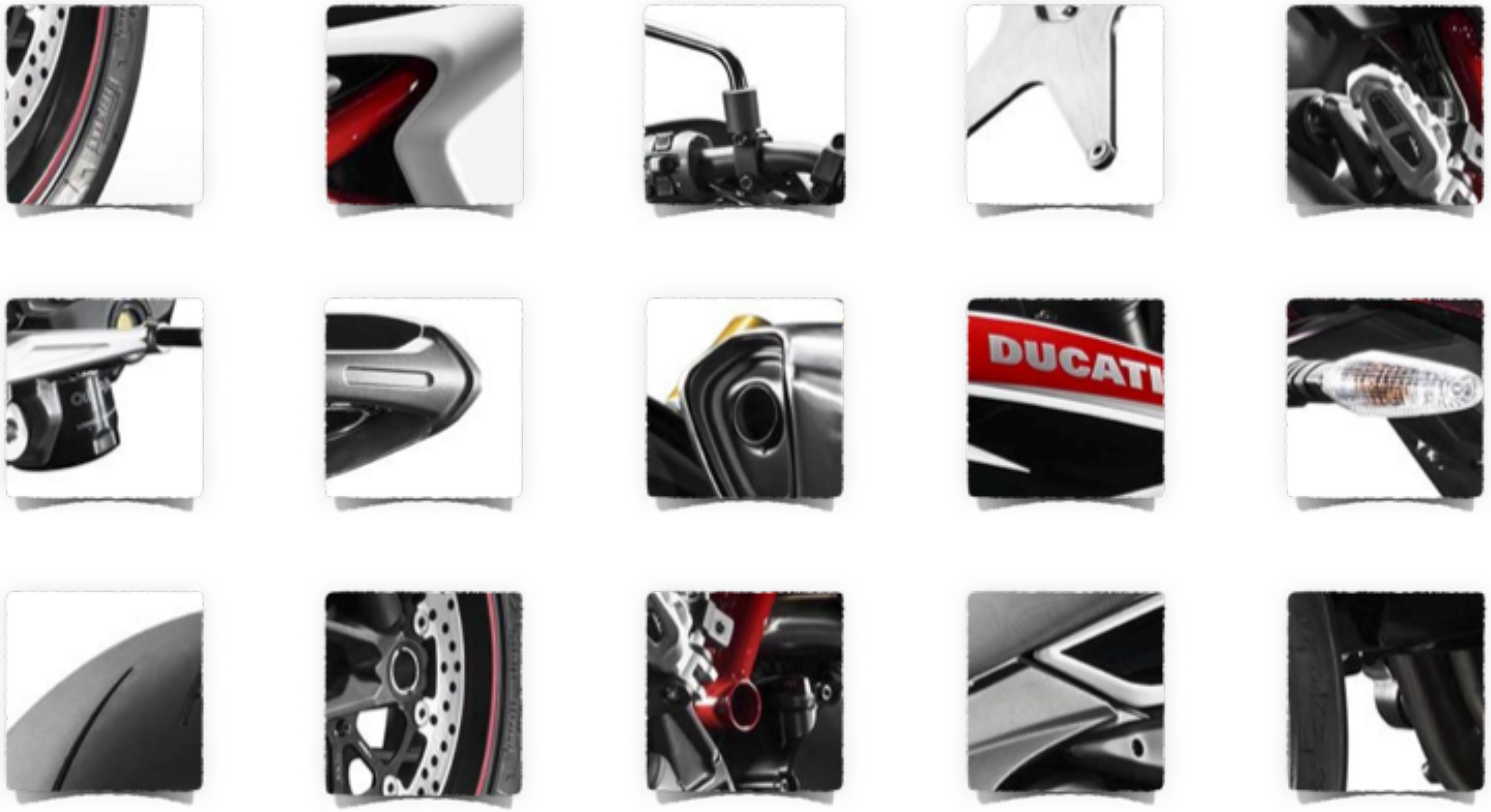
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Visual **Words**

In images, the equivalent of a **word** is a **local image patch**. The local image patch is described using a descriptor such as SIFT.

We construct a **vocabulary** or **codebook** of local descriptors, containing representative local descriptors.

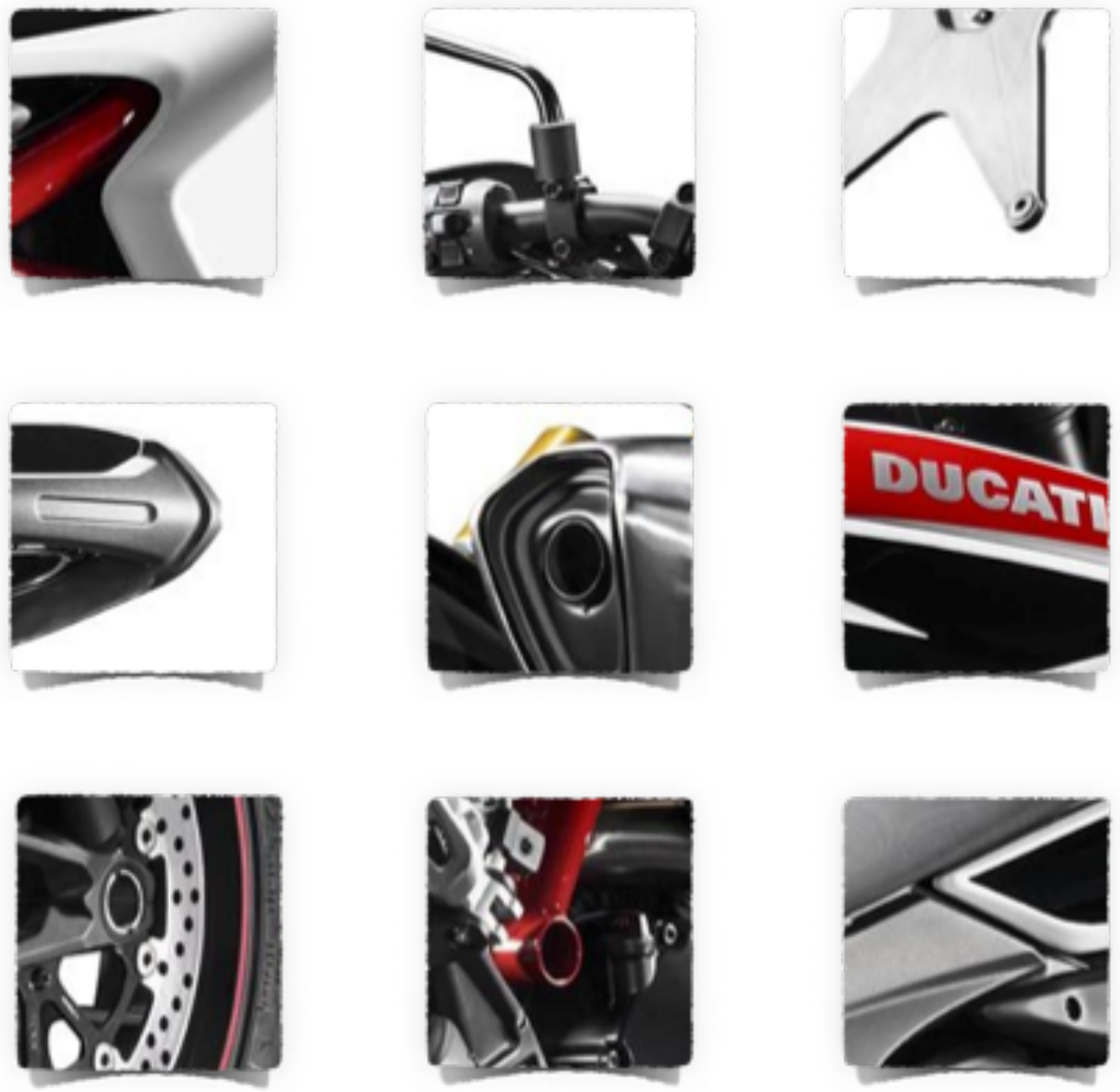
What **Objects** do These Parts Belong To?



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Some local feature are very informative

An object as



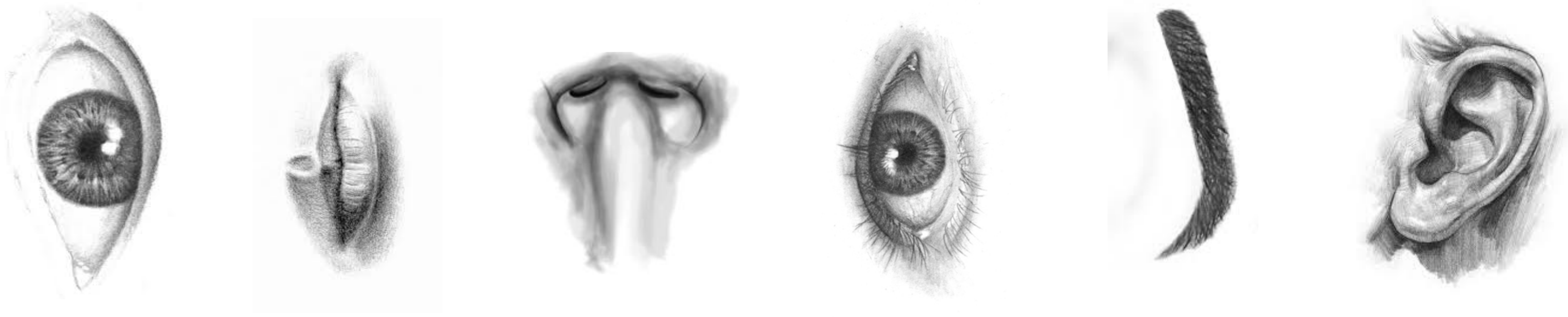
a collection of local features

(bag-of-features)

- deals well with occlusion
- scale invariant
- rotation invariant

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

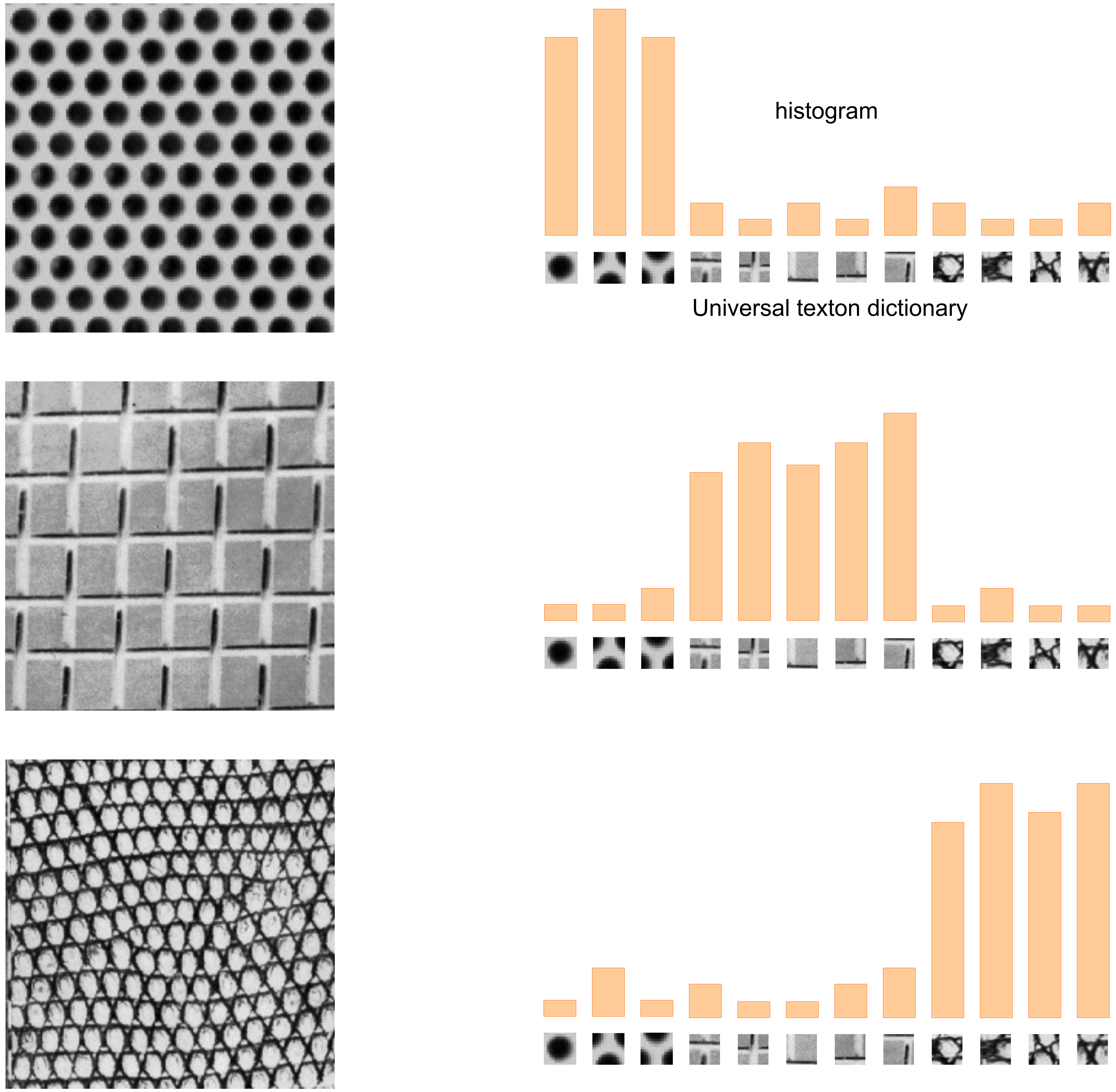
(**not so**) Crazy Assumption



spatial information of local features
can be ignored for object recognition (i.e., verification)

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Recall: Texture Representation



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Standard **Bag-of-Words** Pipeline (for image classification)

Dictionary Learning:
Learn Visual Words using clustering

Encode:
build Bags-of-Words (BOW) vectors
for each image

Classify:
Train and test data using BOWs

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Standard **Bag-of-Words** Pipeline (for image classification)

Dictionary Learning:
Learn Visual Words using clustering

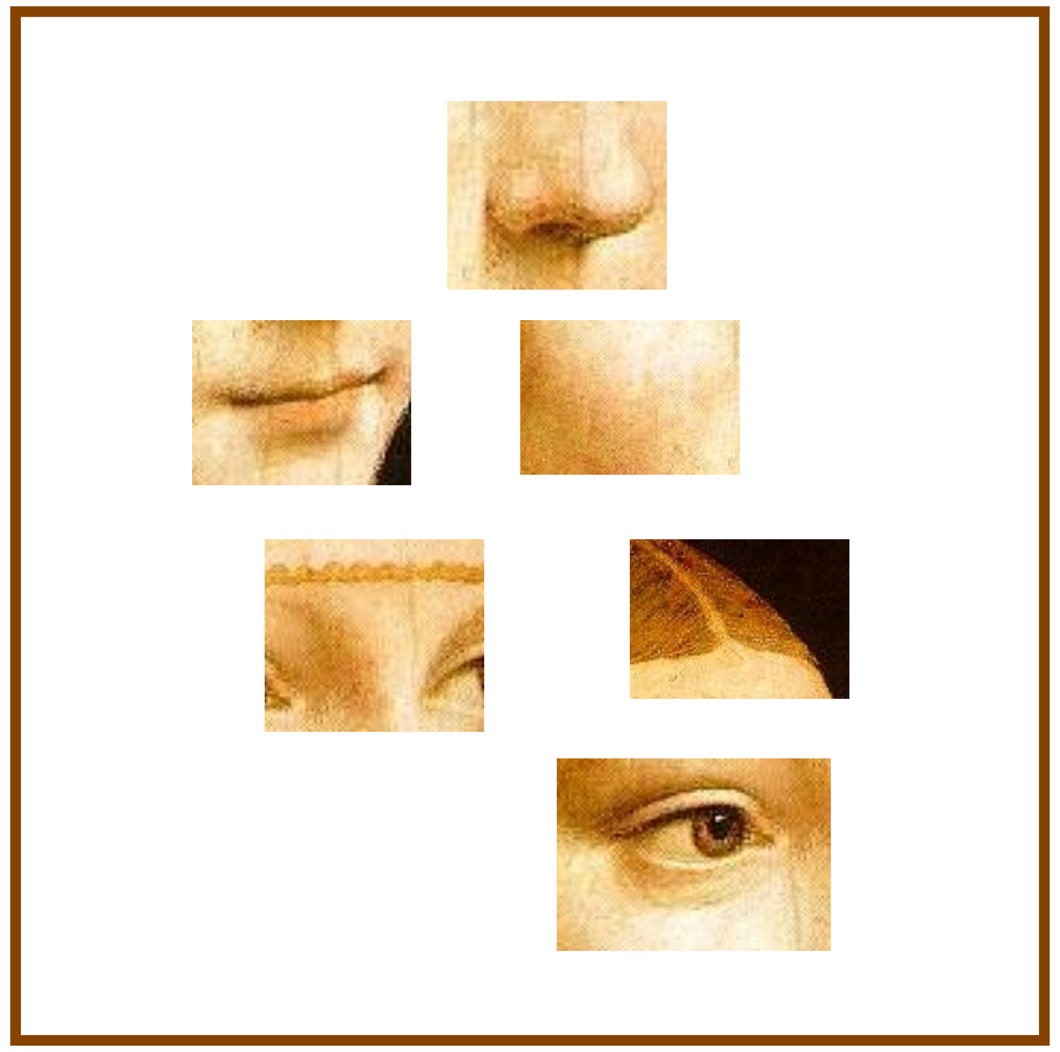
Encode:
build Bags-of-Words (BOW) vectors
for each image

Classify:
Train and test data using BOWs

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

1. **Dictionary Learning**: Learn Visual Words using Clustering

1. **Extract features** (e.g., SIFT) from images



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

1. **Dictionary Learning**: Learn Visual Words using Clustering

2. **Learn visual dictionary** (e.g., K-means clustering)

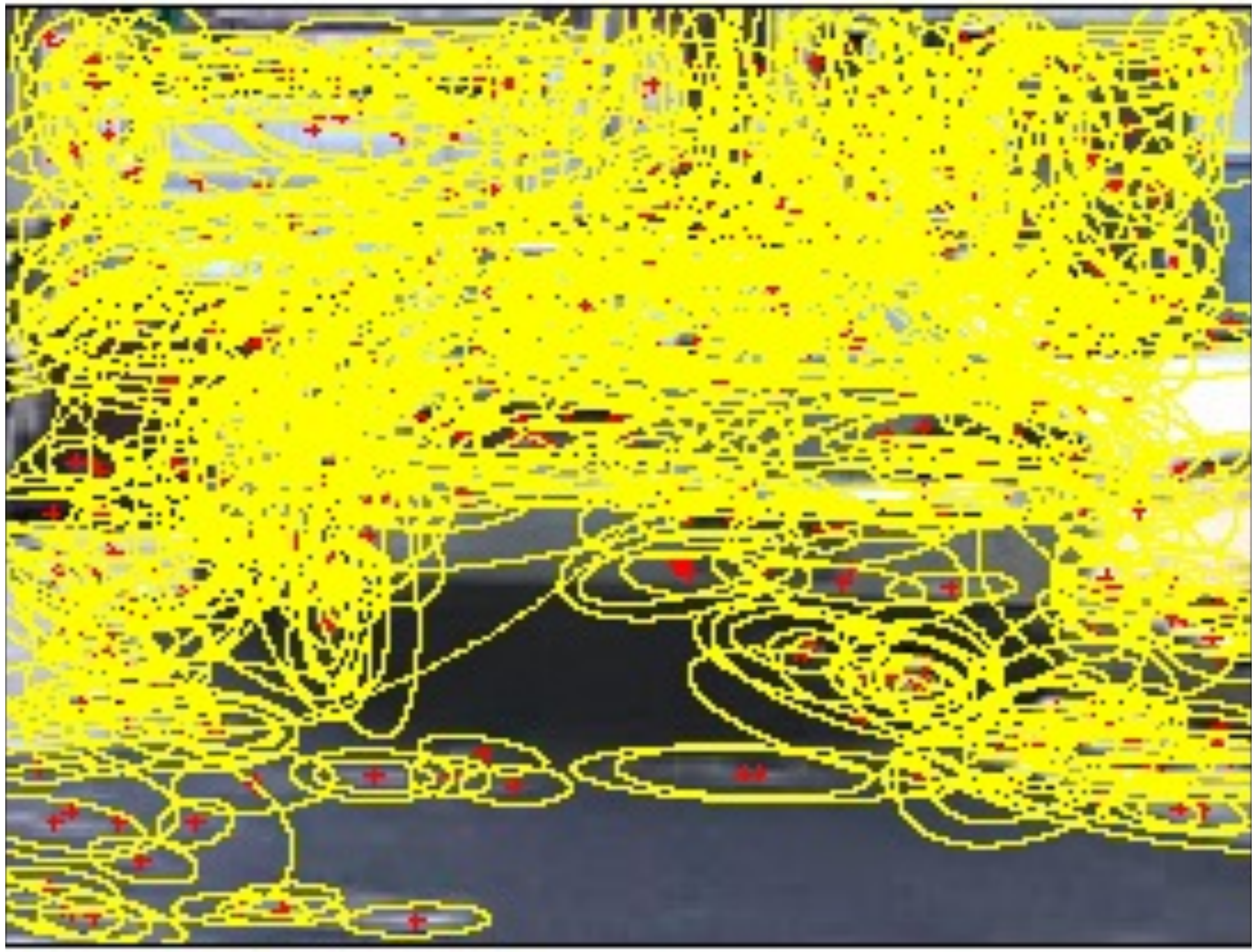
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

What **Features** Should We Extract?

- Regular grid
  Vogel & Schiele, 2003
  Fei-Fei & Perona, 2005

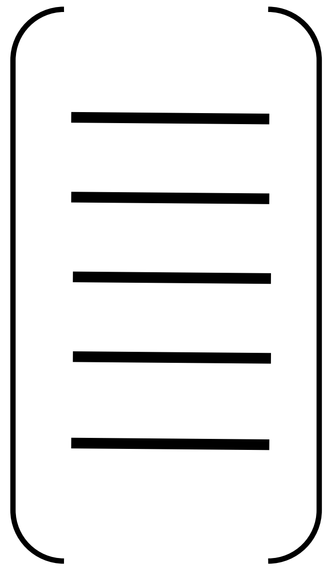
- Interest point detector
  Csurka et al. 2004
  Fei-Fei & Perona, 2005
  Sivic et al. 2005

- Other methods
  Random sampling (Vidal-Naquet & Ullman, 2002)
  Segmentation-based patches (Barnard et al. 2003)



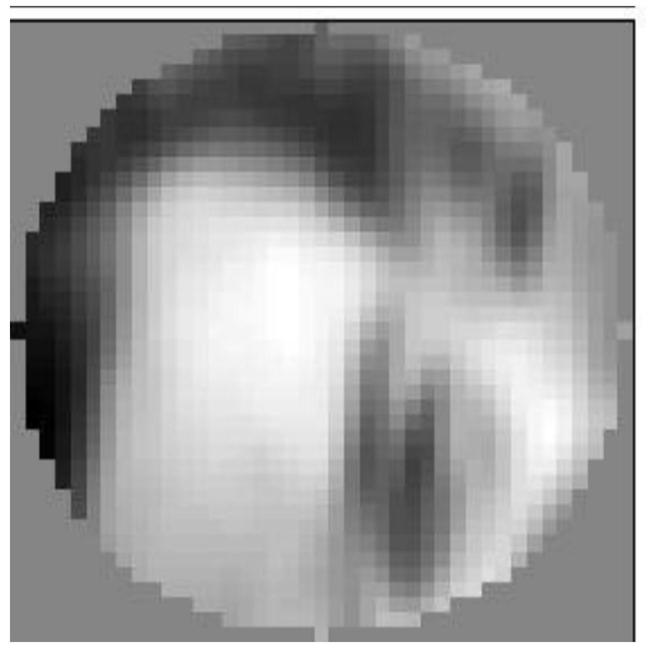
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Extracting **SIFT** Patches

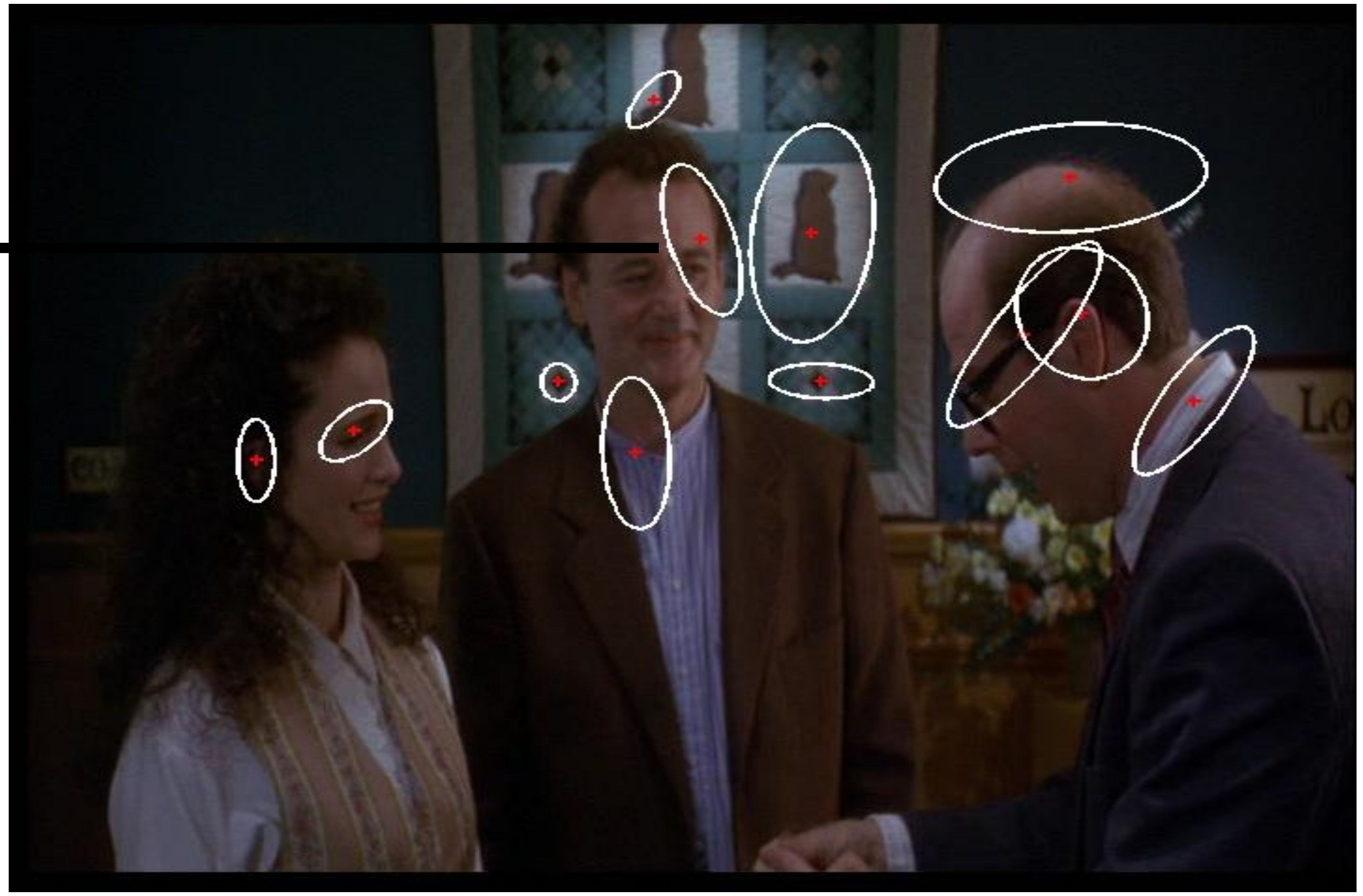


Compute SIFT descriptor

[Lowe'99]



Normalize patch



Detect patches

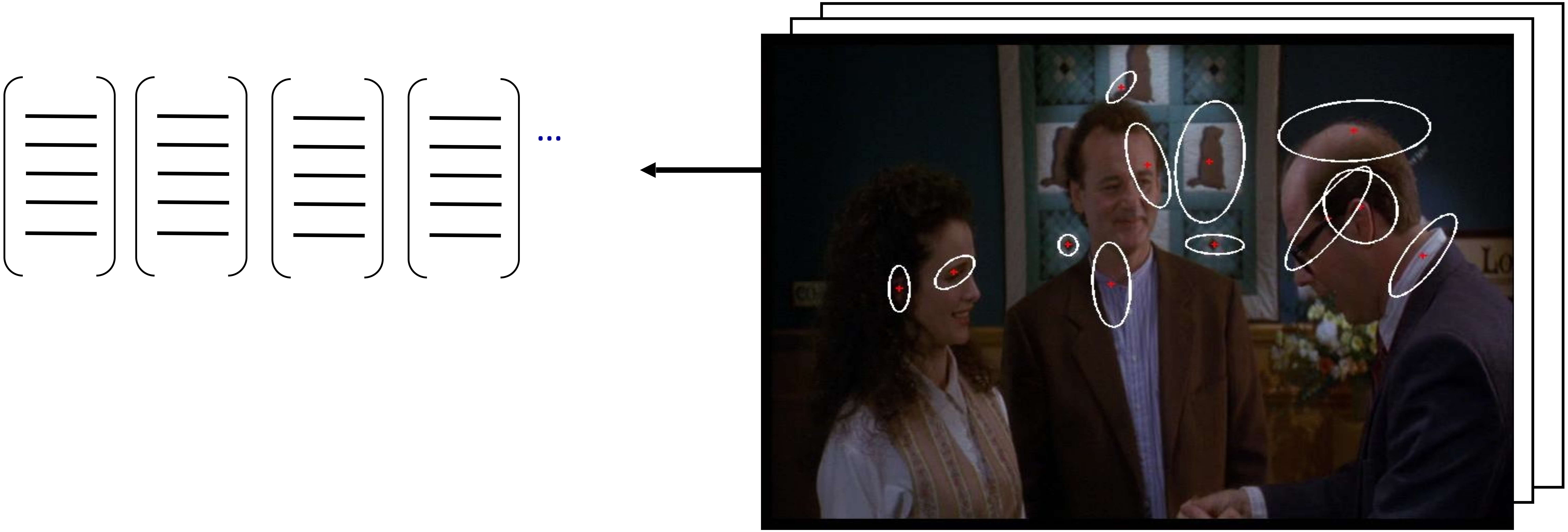
[Mikojaczyk and Schmid '02]

[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

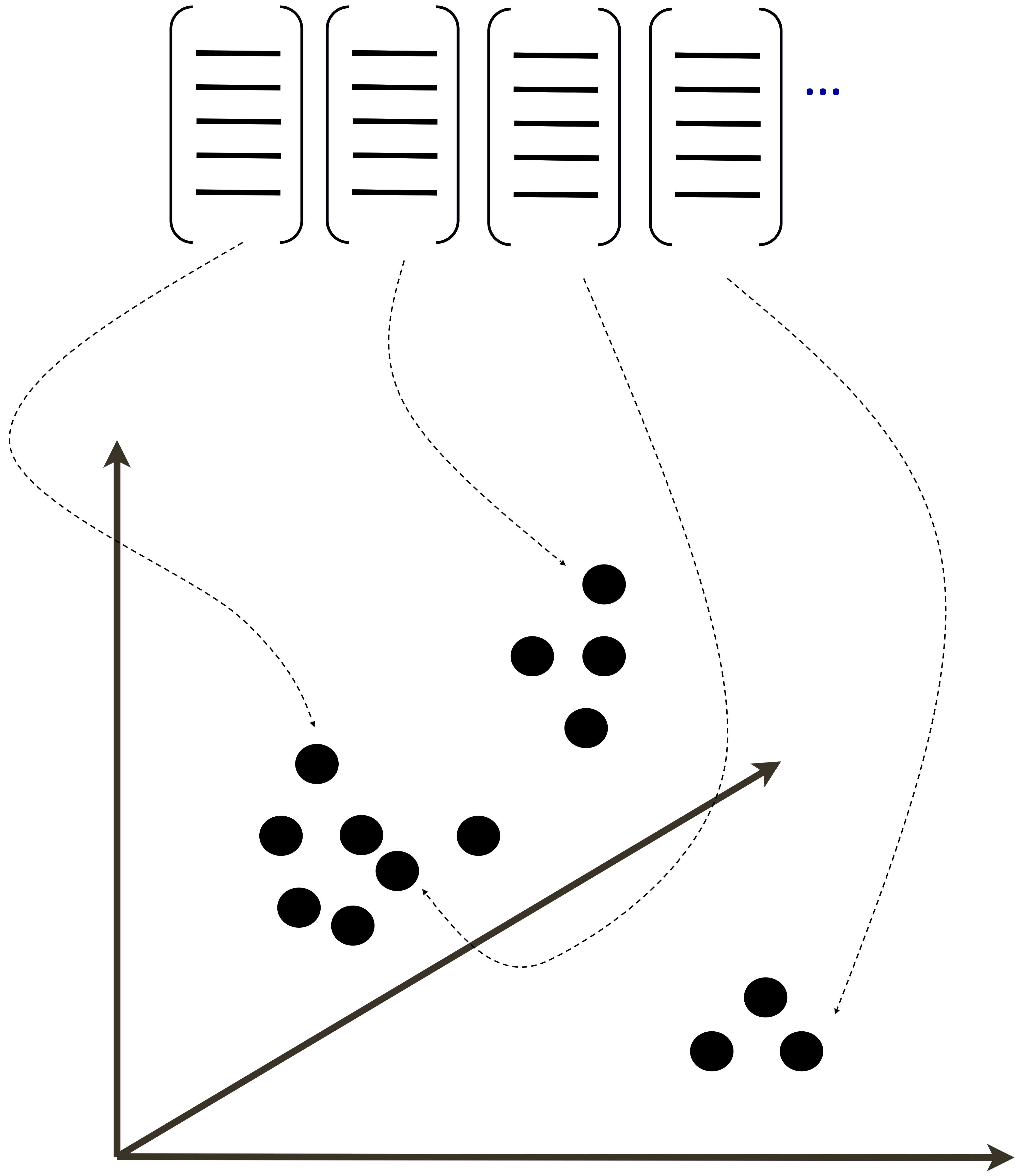
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Extracting **SIFT** Patches



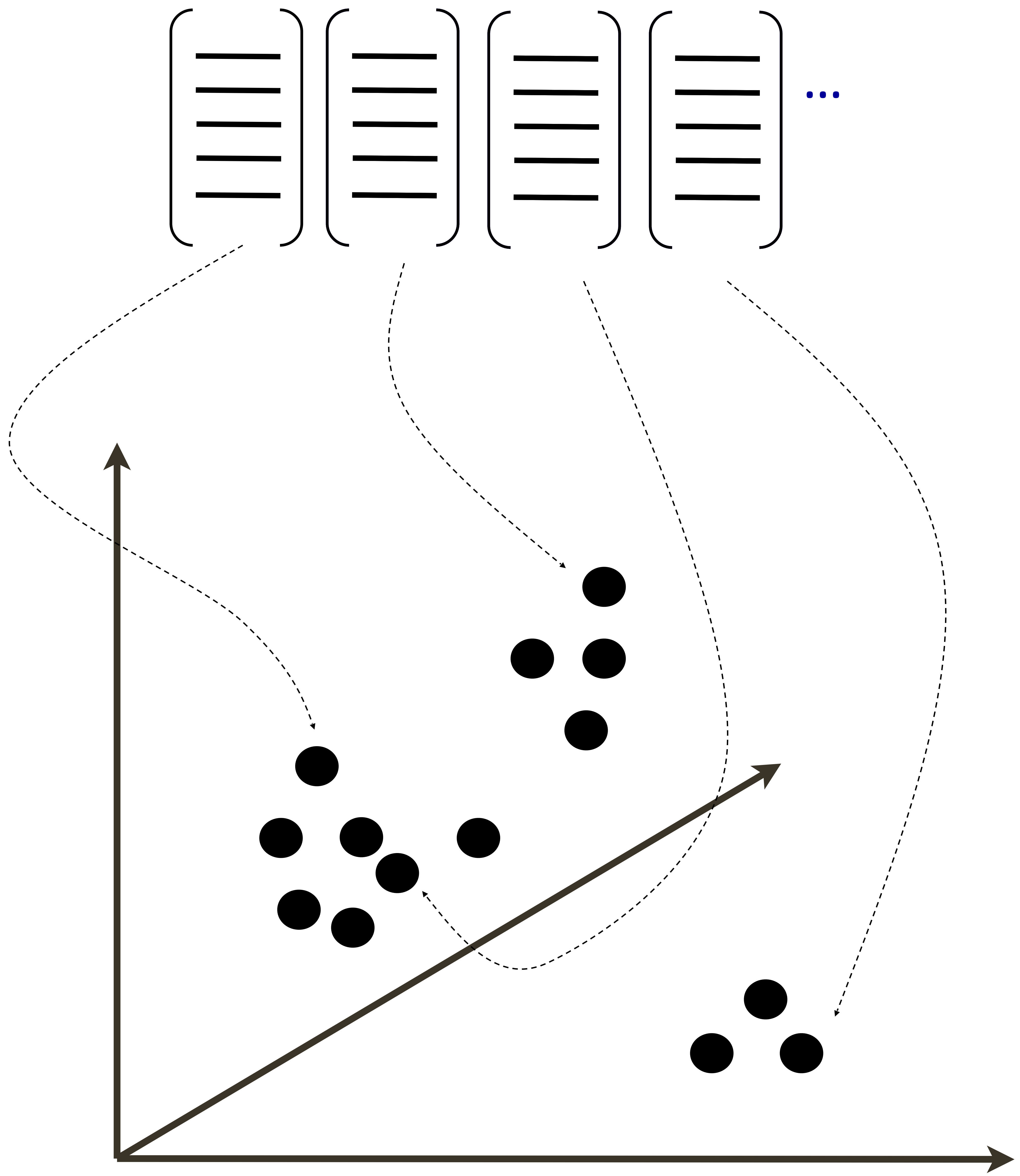
Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

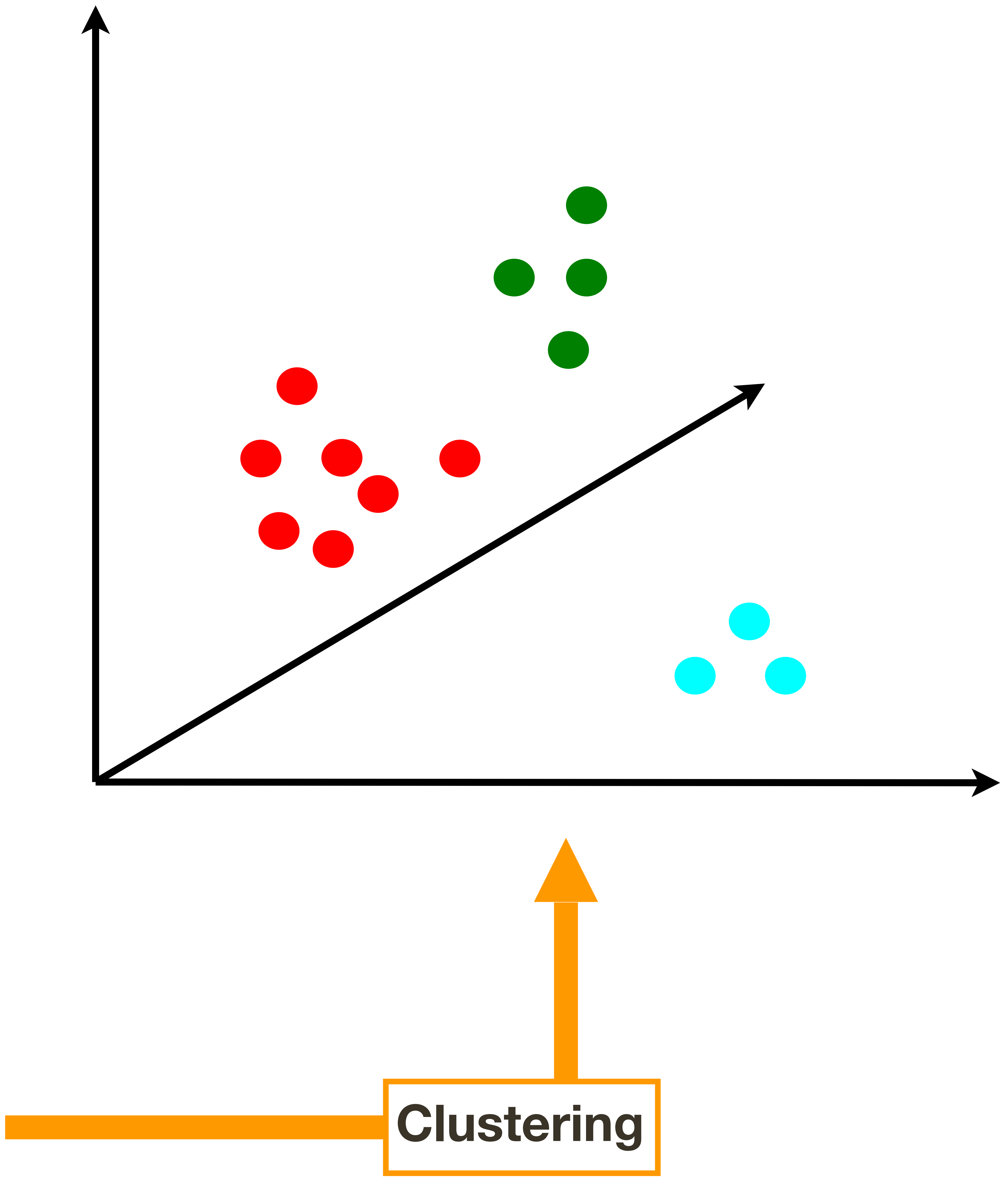
Creating **Dictionary**



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

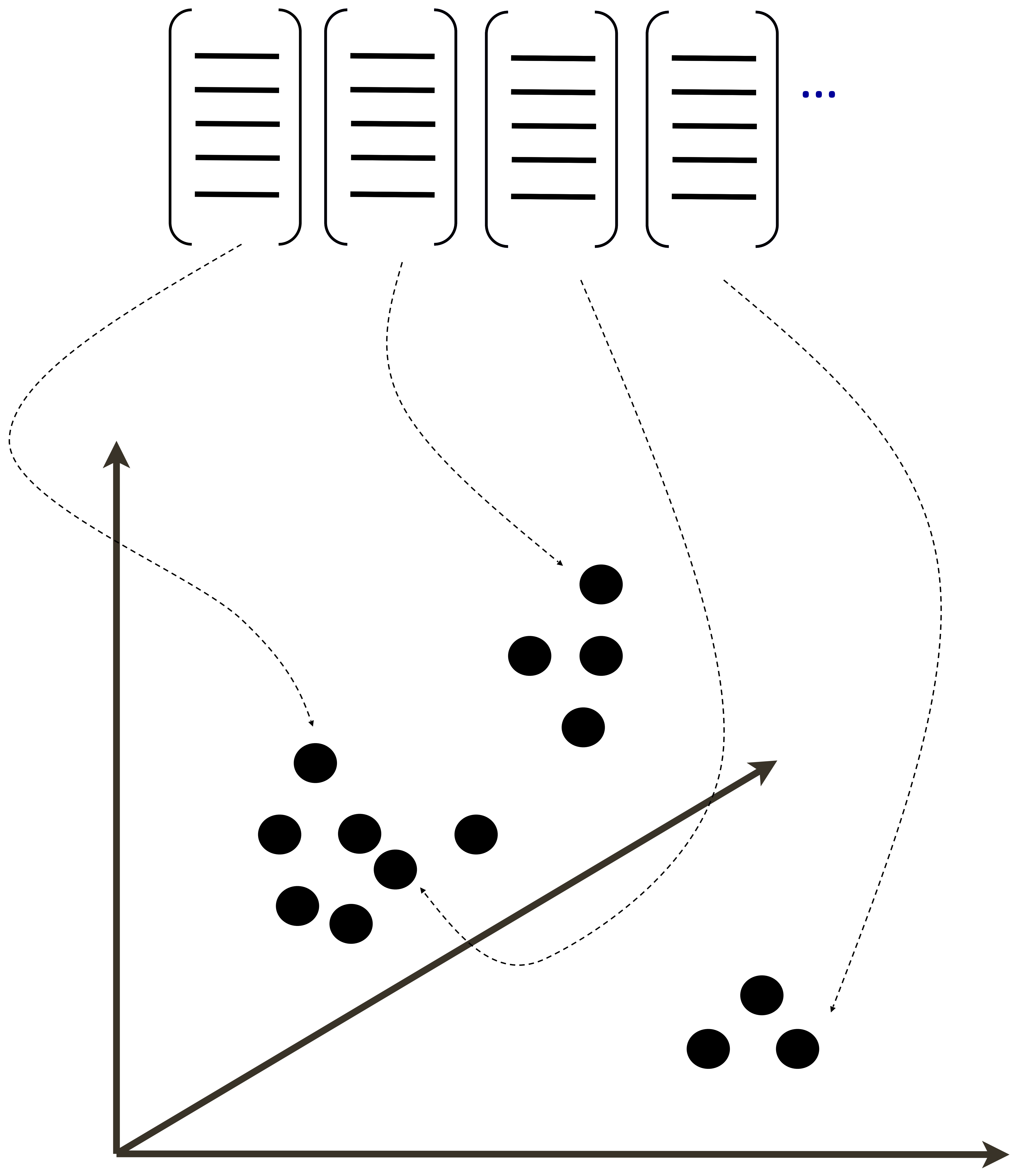
Creating **Dictionary**

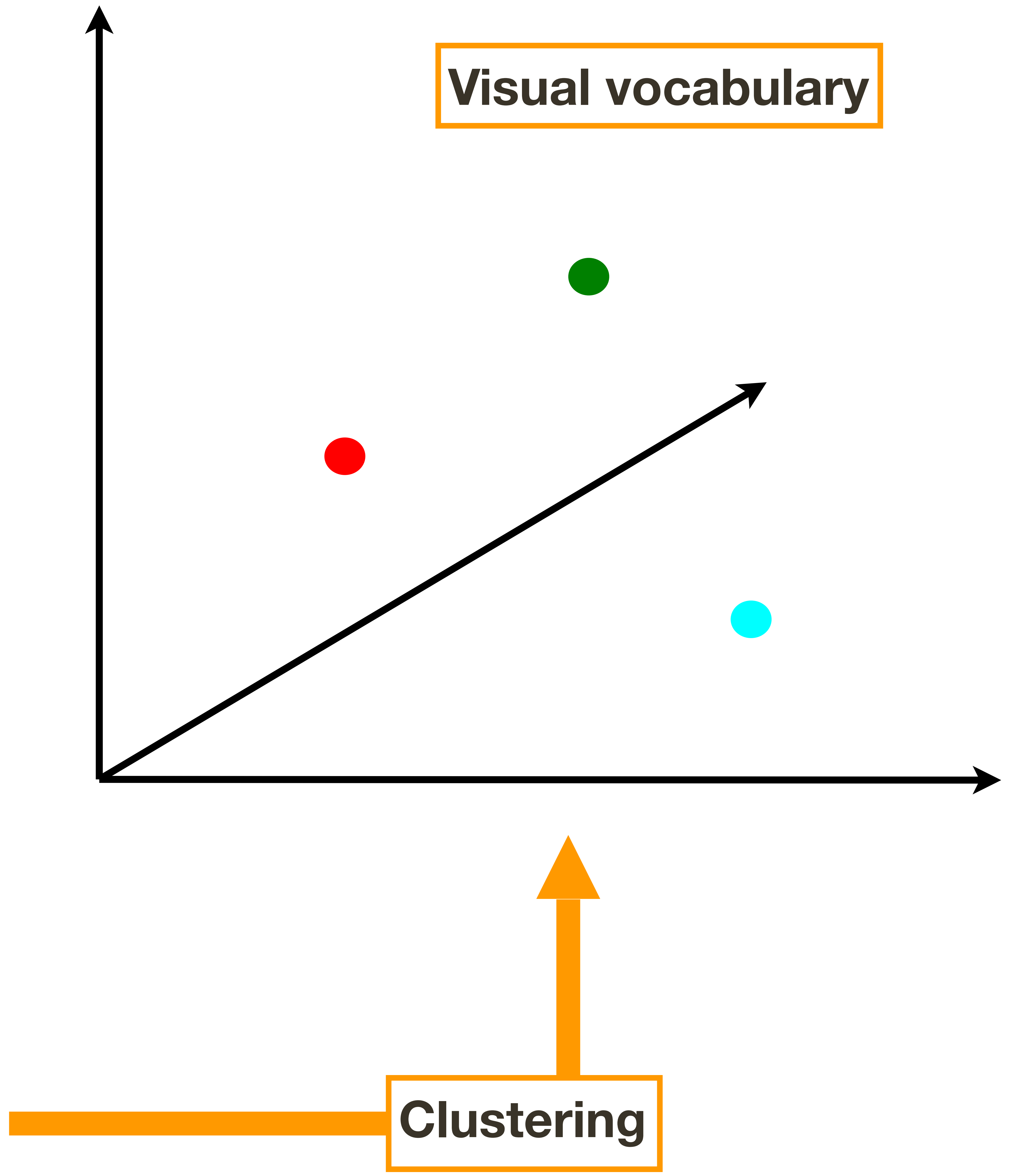




Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Creating **Dictionary**





Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

K-means clustering

K-Means Clustering

Assume we **know** how many clusters there are in the data - denote by K

Each **cluster** is represented by a **cluster center**, or mean

Our objective is to **minimize the representation error** (or quantization error) in letting each data point be represented by some cluster center

Minimize

$$\sum_{i \in clusters} \left\{ \sum_{j \in i^{th}\ cluster} ||x_j - \mu_i||^2 \right\}$$

K-Means Clustering

K-means clustering alternates between two steps:

1. Assume the cluster centers are known (fixed). Assign each point to the closest cluster center.

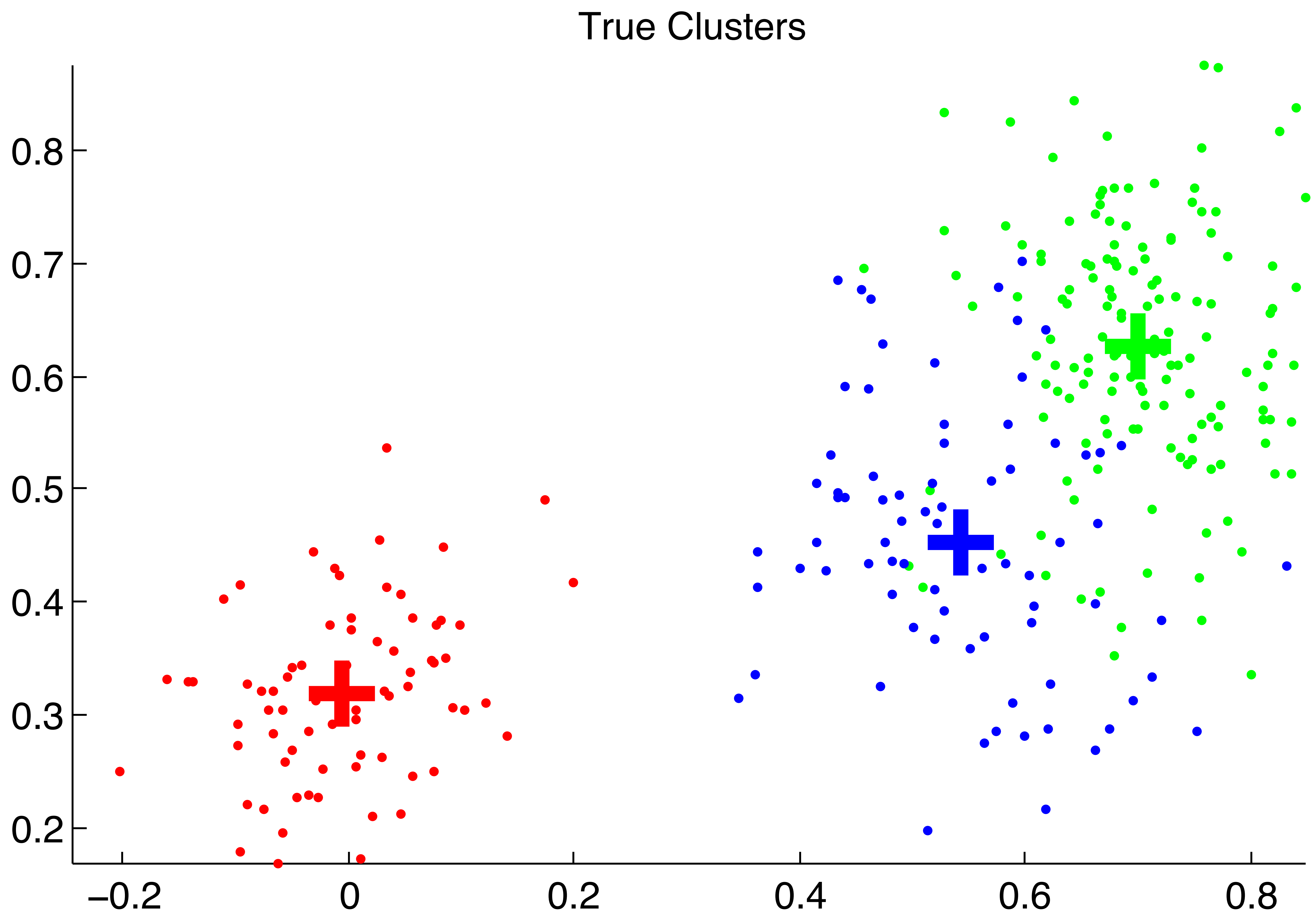
2. Assume the assignment of points to clusters is known (fixed). Compute the best center for each cluster, as the mean of the points assigned to the cluster.

The algorithm is initialized by choosing K random cluster centers

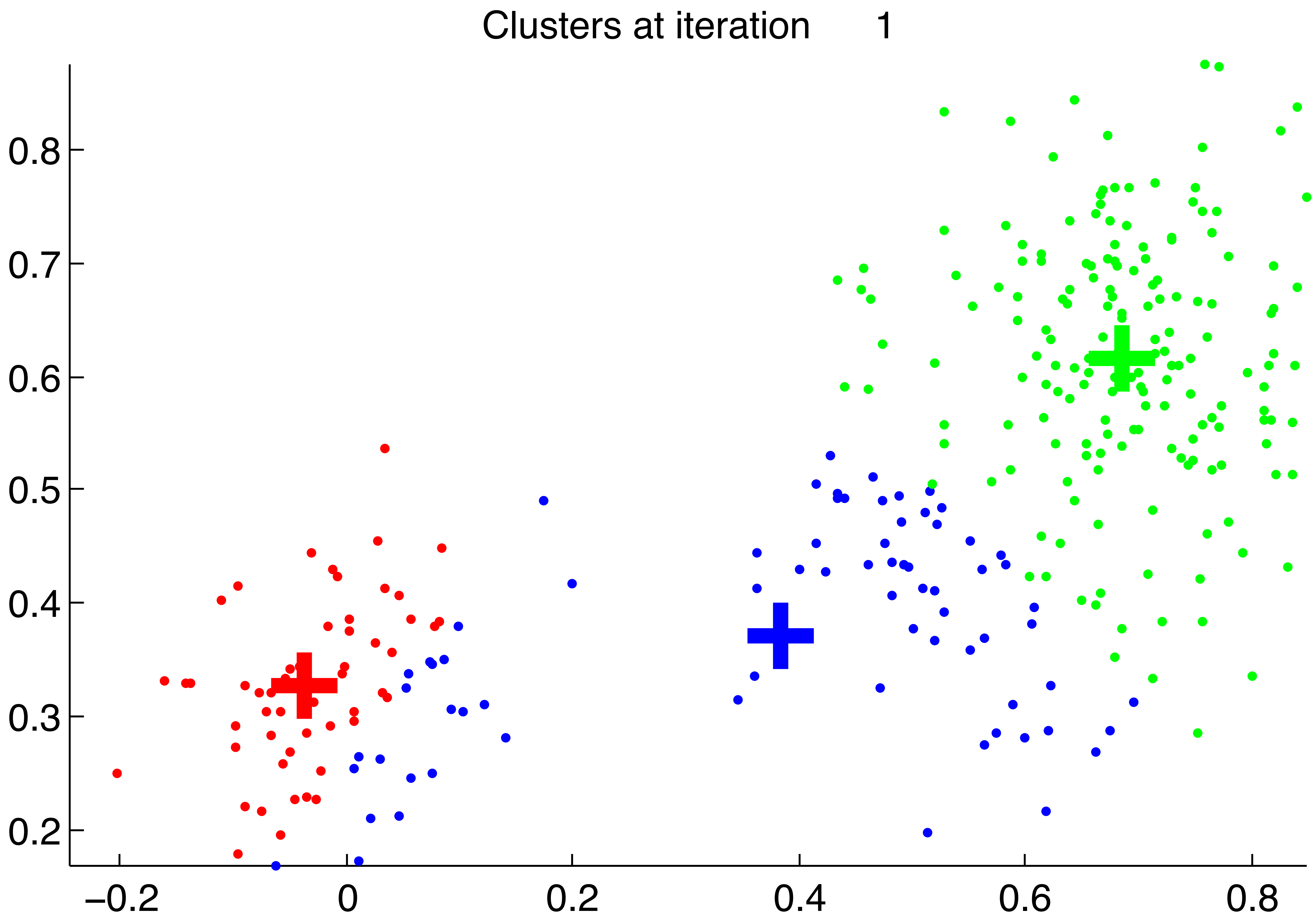
K-means converges to a local minimum of the objective function

— Results are initialization dependent

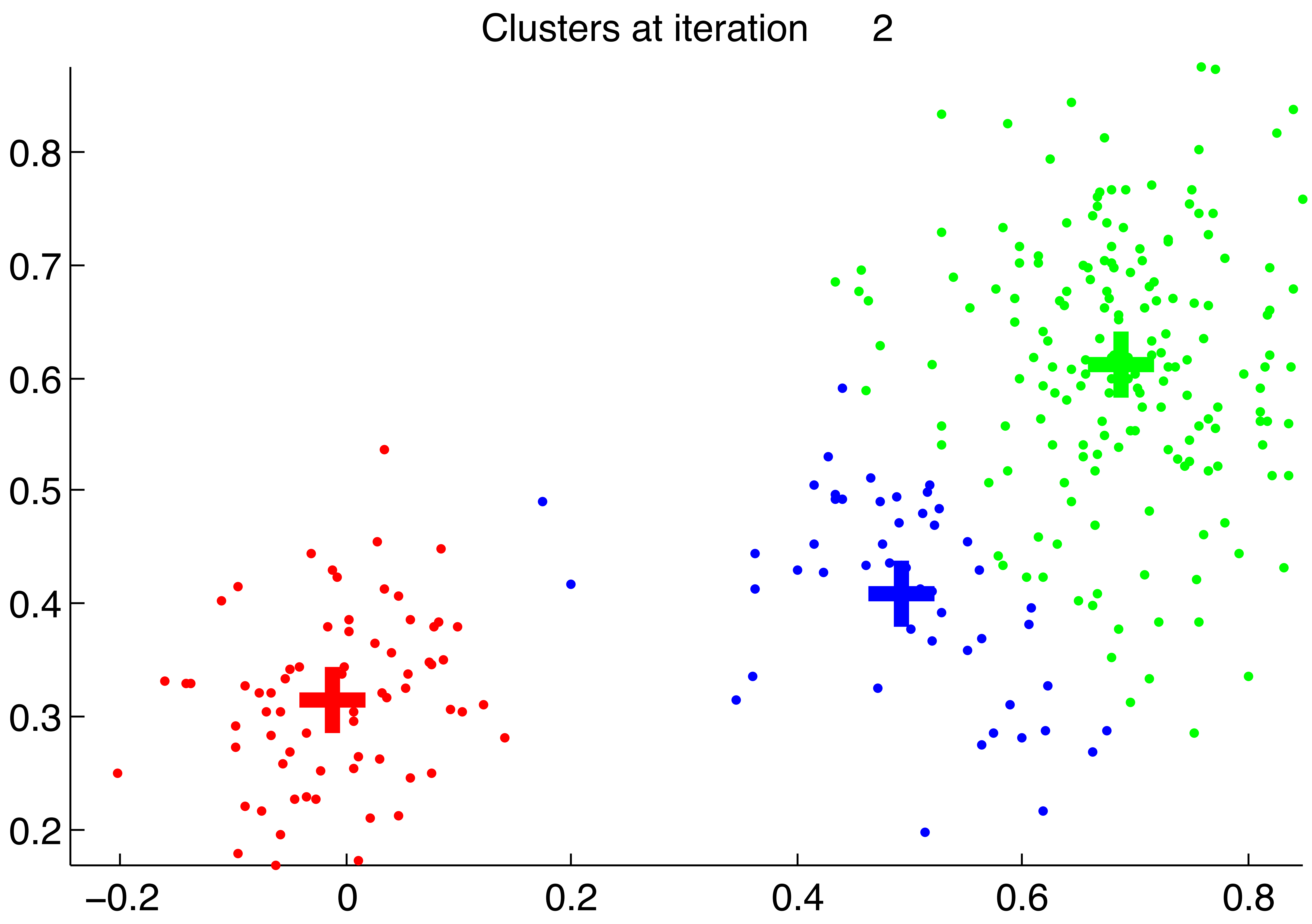
Example 1: K-Means Clustering



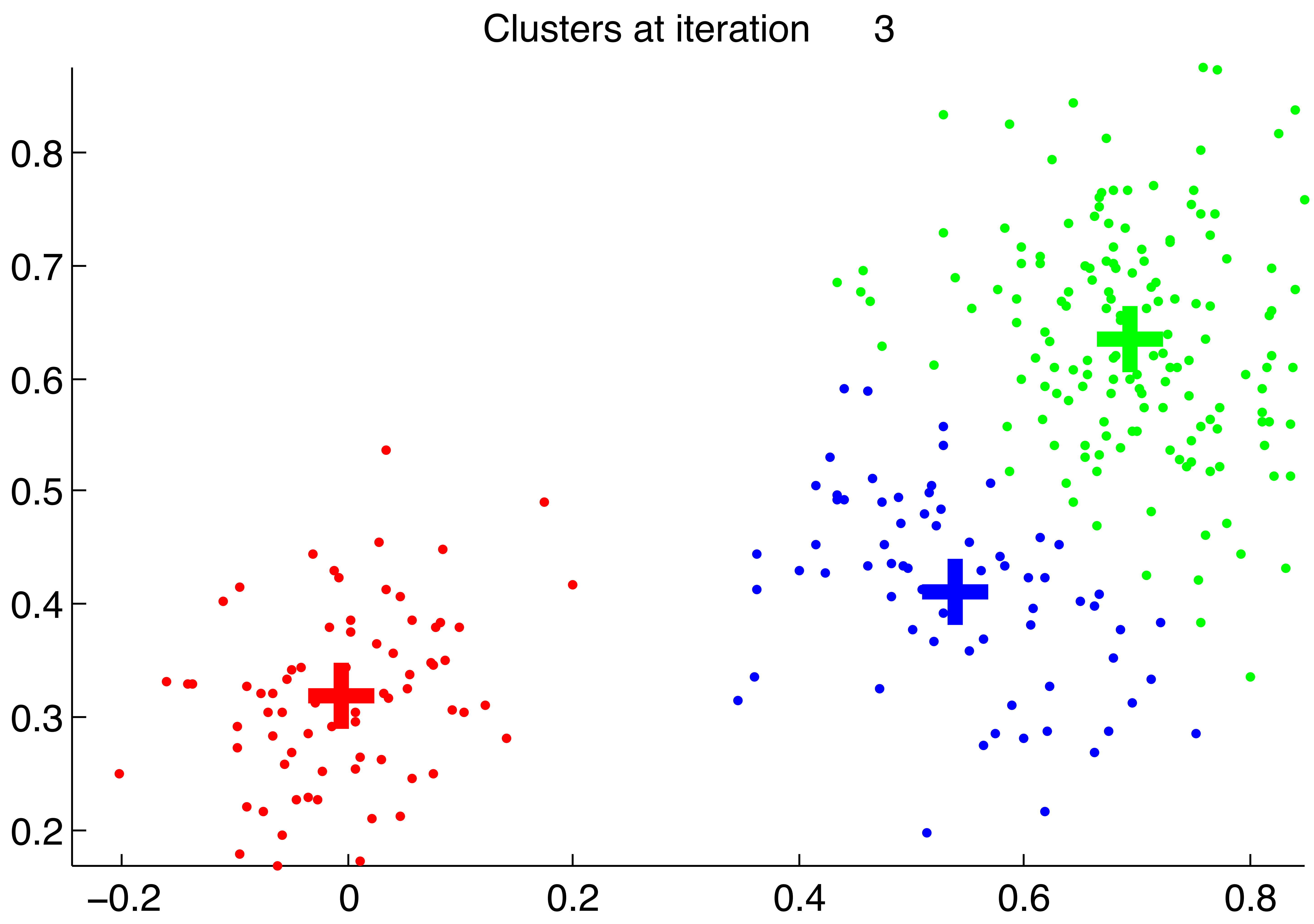
Example 1: K-Means Clustering



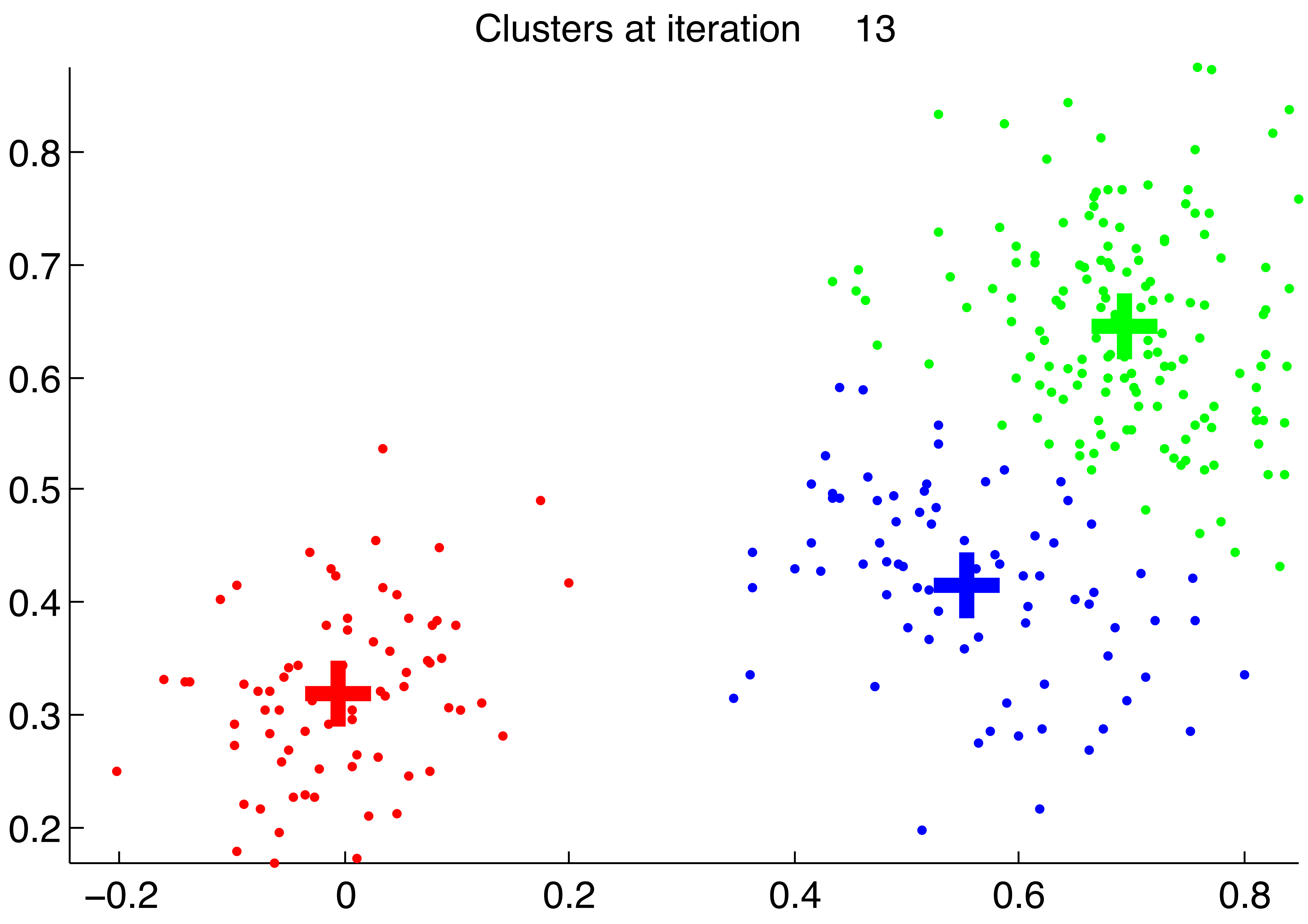
Example 1: K-Means Clustering



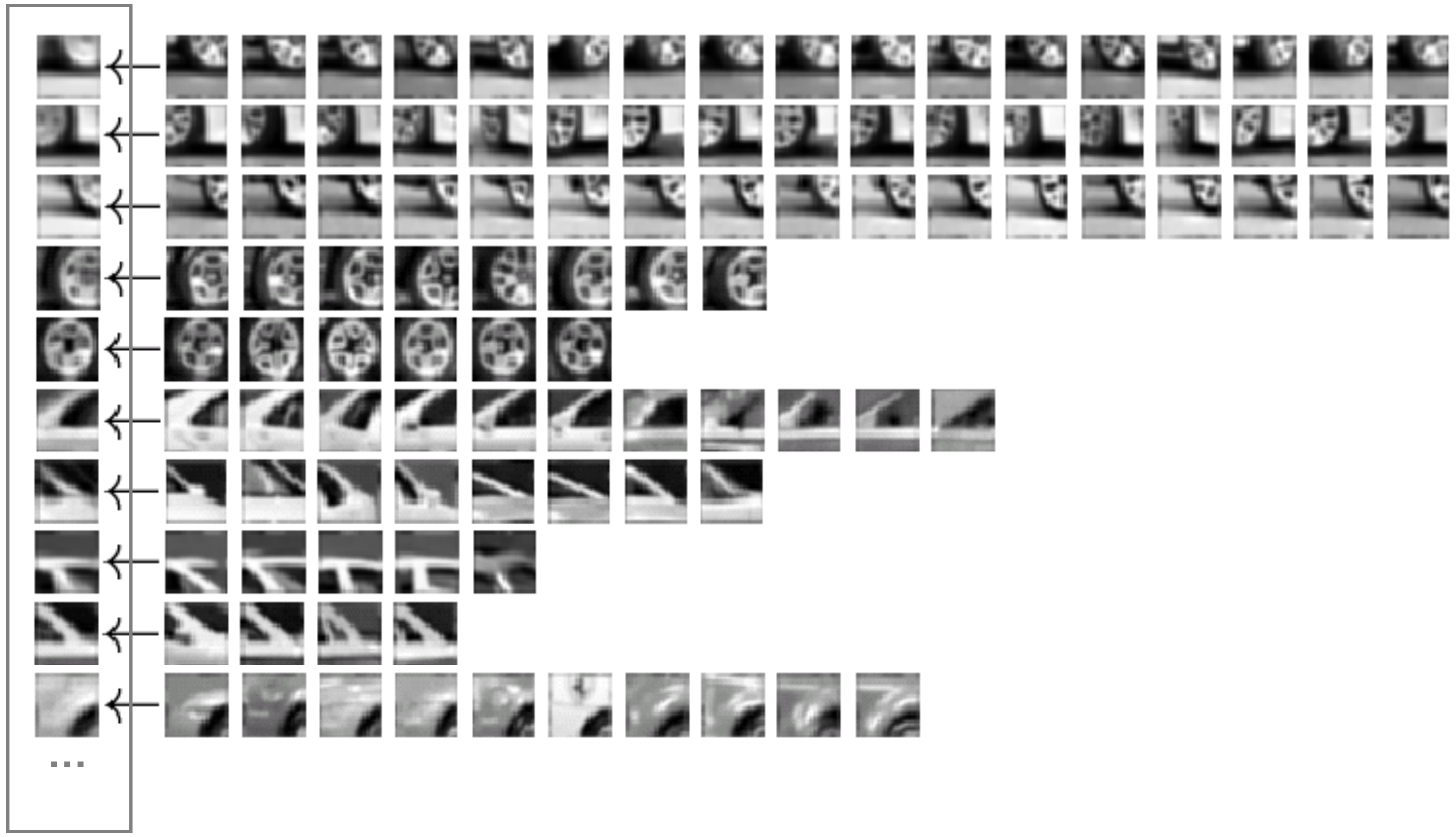
Example 1: K-Means Clustering



Example 1: K-Means Clustering

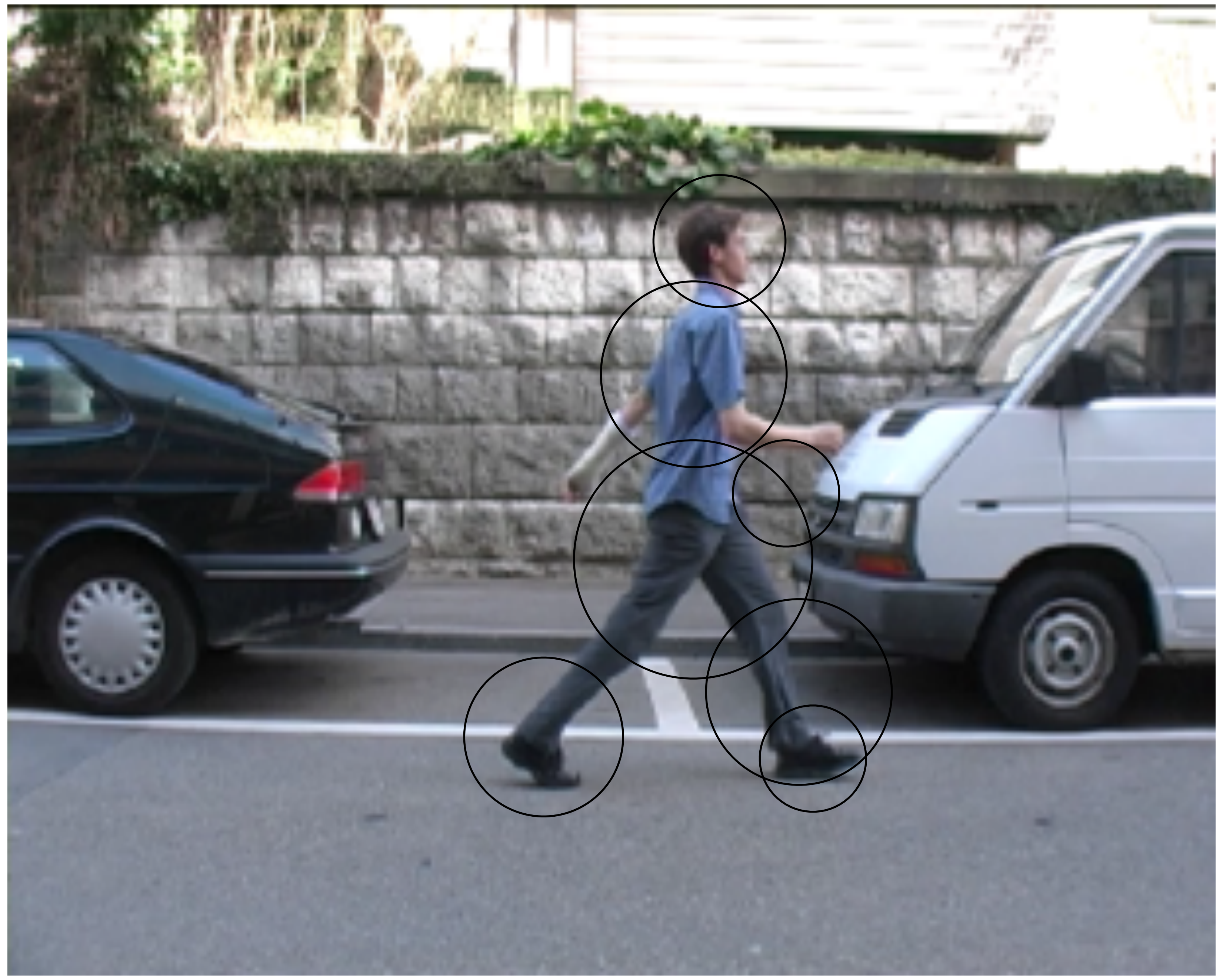


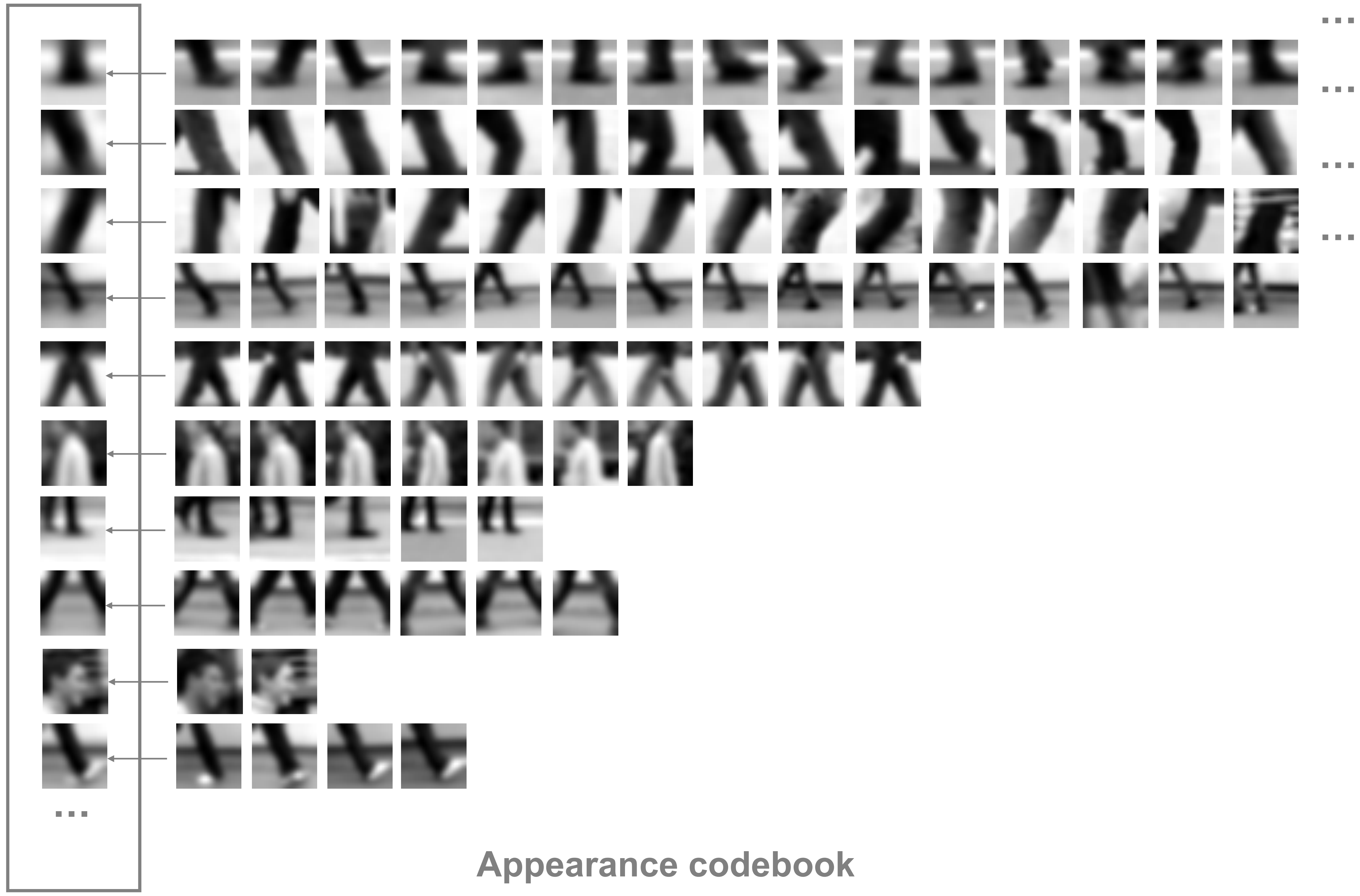
Example **Visual Dictionary**



Source: B. Leibe

Example **Visual Dictionary**





Source: B. Leibe

Standard **Bag-of-Words** Pipeline (for image classification)

Dictionary Learning:
Learn Visual Words using clustering

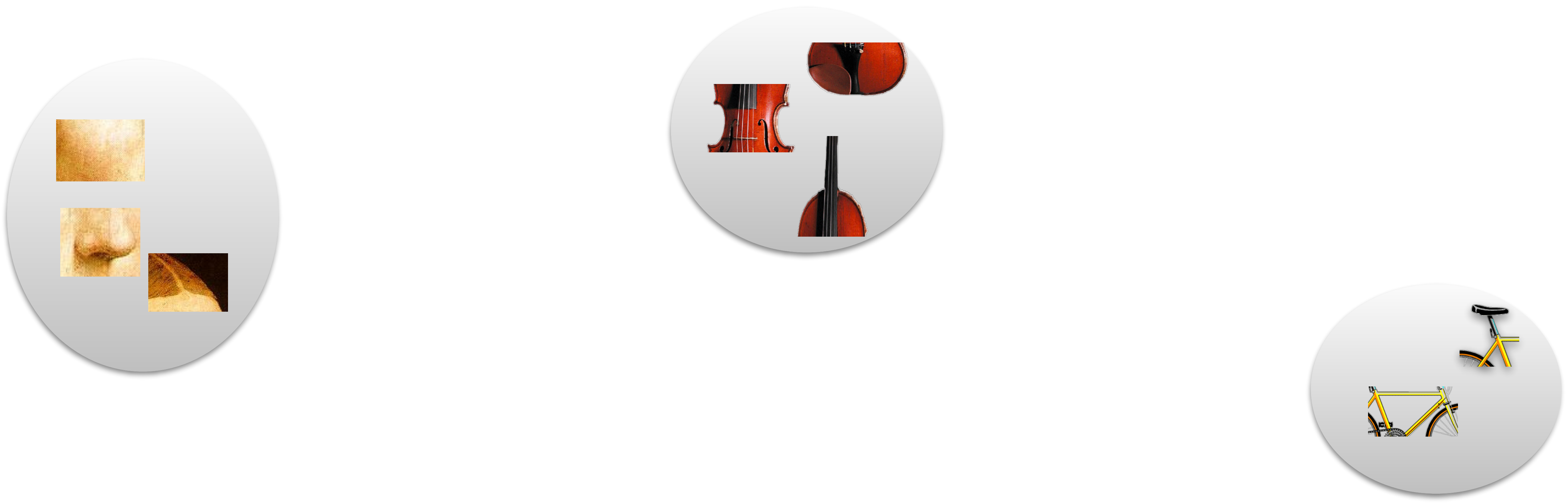
Encode:
build Bags-of-Words (BOW) vectors
for each image

Classify:
Train and test data using BOWs

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

2. **Encode:** build Bag-of-Words (BOW) vectors for each image

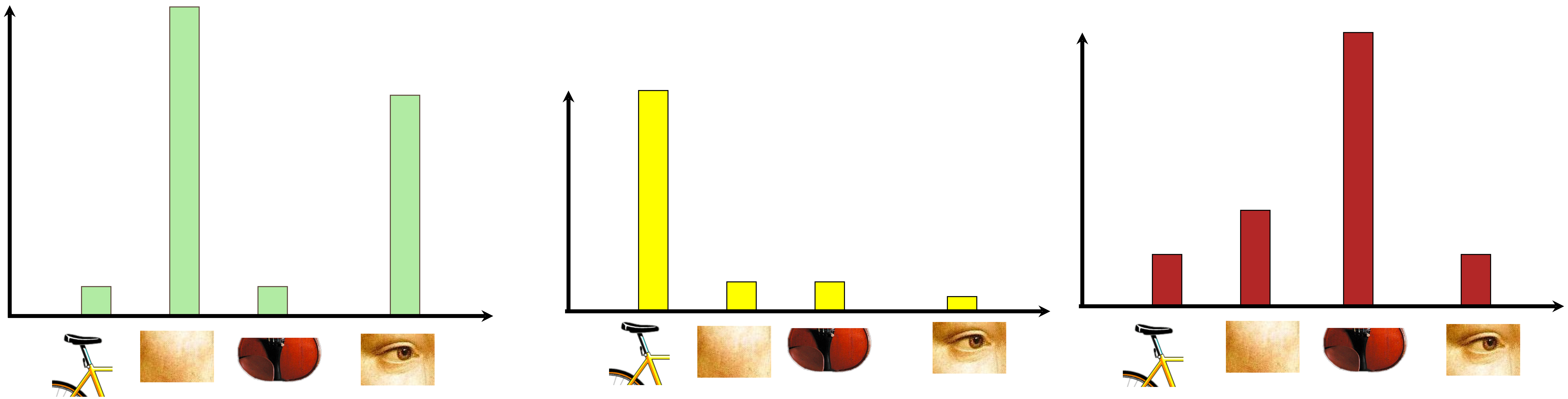
1. **Quantization**: image features gets associated to a visual word (nearest cluster center)



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

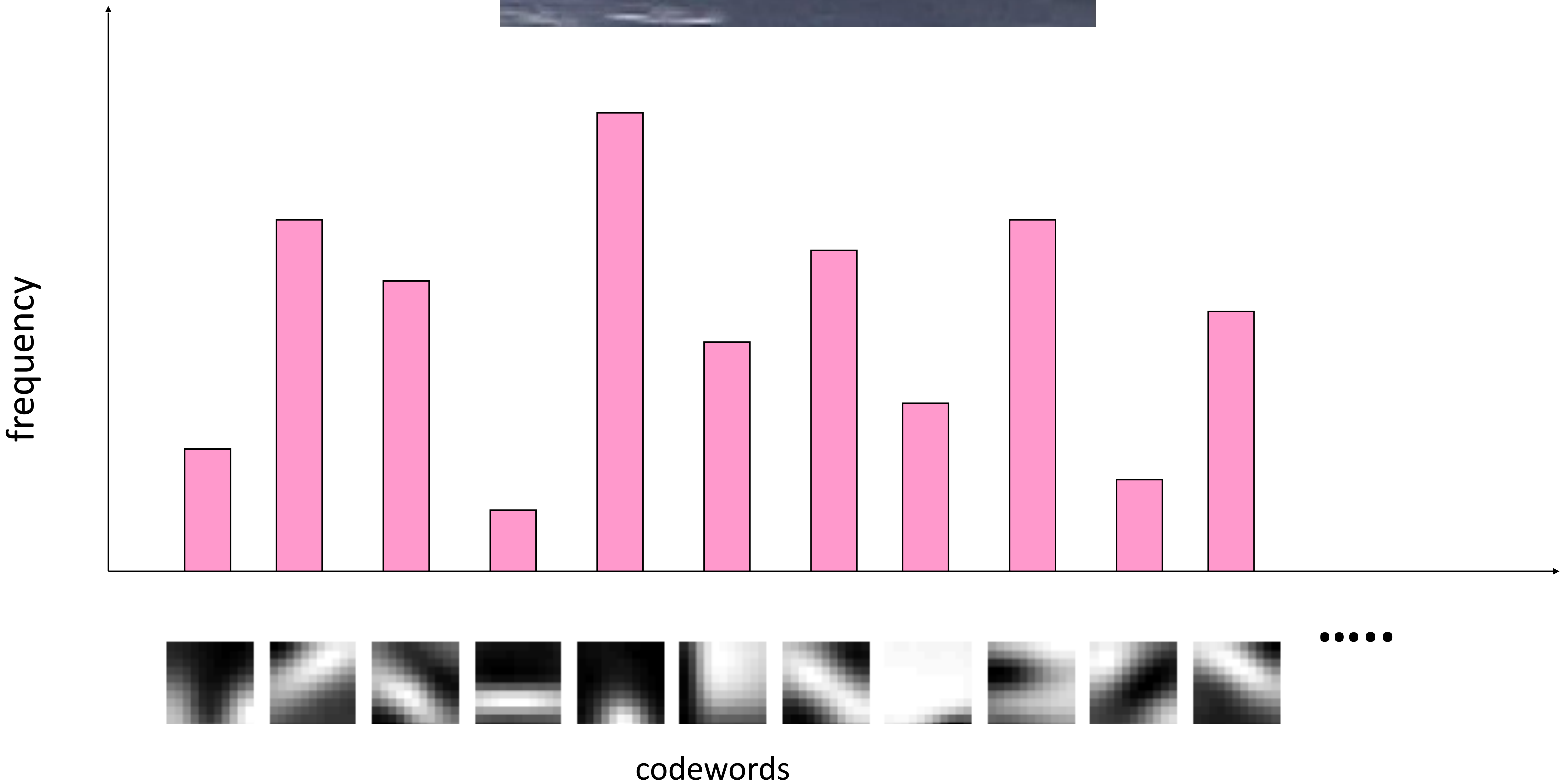
2. Encode: build Bag-of-Words (BOW) vectors for each image

2. **Histogram**: count the number of visual word occurrences



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

2. **Encode:** build Bag-of-Words (BOW) vectors for each image



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Standard **Bag-of-Words** Pipeline (for image classification)

Dictionary Learning:
Learn Visual Words using clustering

Encode:
build Bags-of-Words (BOW) vectors
for each image

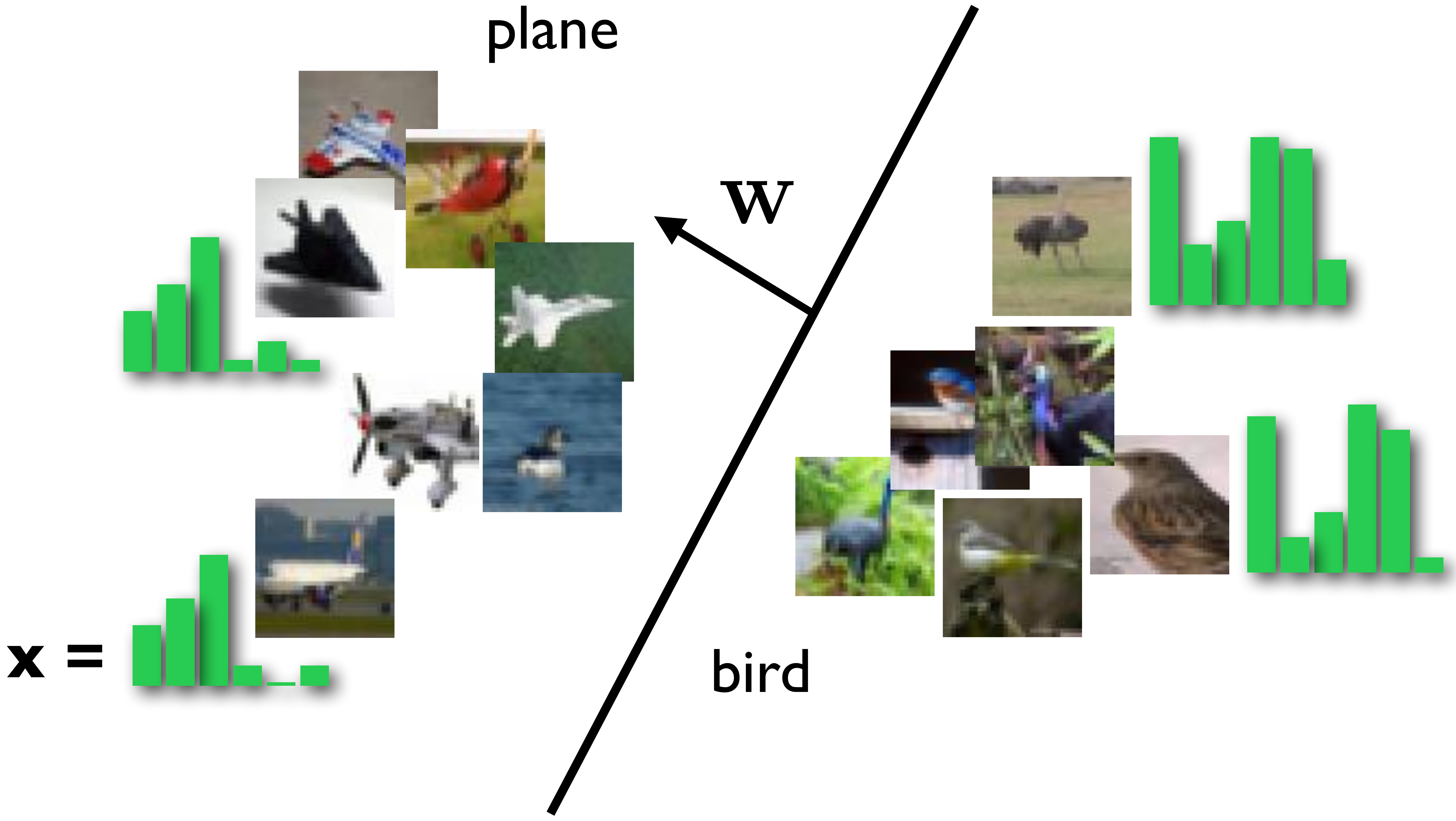
Classify:
Train and test data using BOWs

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Classify Visual Word Histograms

e.g., bird vs plane classifier as linear classifier in space of histograms

Histograms of visual word frequencies = vector **x**, linear classifier **w**



Bayes Rule (Review and Definitions)

Let c be the **class label** and let x be the **measurement** (i.e., evidence)

$$\underbrace{P(c|x)}_{\text{posterior probability}} = \frac{P(x|c)p(c)}{P(x)}$$

Bayes Rule (Review and Definitions)

Let c be the **class label** and let x be the **measurement** (i.e., evidence)

class–conditional probability
(a.k.a. likelihood)

prior probability

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)}$$

posterior probability

unconditional probability
(a.k.a. marginal likelihood)

Bayes Rule (Review and Definitions)

Let c be the **class label** and let x be the **measurement** (i.e., evidence)

Simple case:

— binary classification; i.e., $c \in \{1, 2\}$

— features are 1D; i.e., $x \in \mathbb{R}$

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)}$$

Bayes Rule (Review and Definitions)

Let c be the **class label** and let x be the **measurement** (i.e., evidence)

Simple case:

— binary classification; i.e., $c \in \{1, 2\}$

— features are 1D; i.e., $x \in \mathbb{R}$

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)}$$

Classify ***x*** as

1 if p(1|***x***) > p(2|***x***)

2 if p(1|***x***) < p(2|***x***)

Bayes Rule (Review and Definitions)

Let c be the **class label** and let x be the **measurement** (i.e., evidence)

Simple case:

— binary classification; i.e., $c \in \{1, 2\}$

— features are 1D; i.e., $x \in \mathbb{R}$

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)}$$

General case:

— multi-class; i.e., $c \in \{1, ..., 1000\}$

— features are high-dimensional; i.e., $x \in \mathbb{R}^{2,000+}$

Example: Discrete Bayes Classifier

Assume we have two classes: $c_1 = \textbf{male}$ $c_2 = \textbf{female}$

We have a person who's gender we don't know, who's name is *drew*

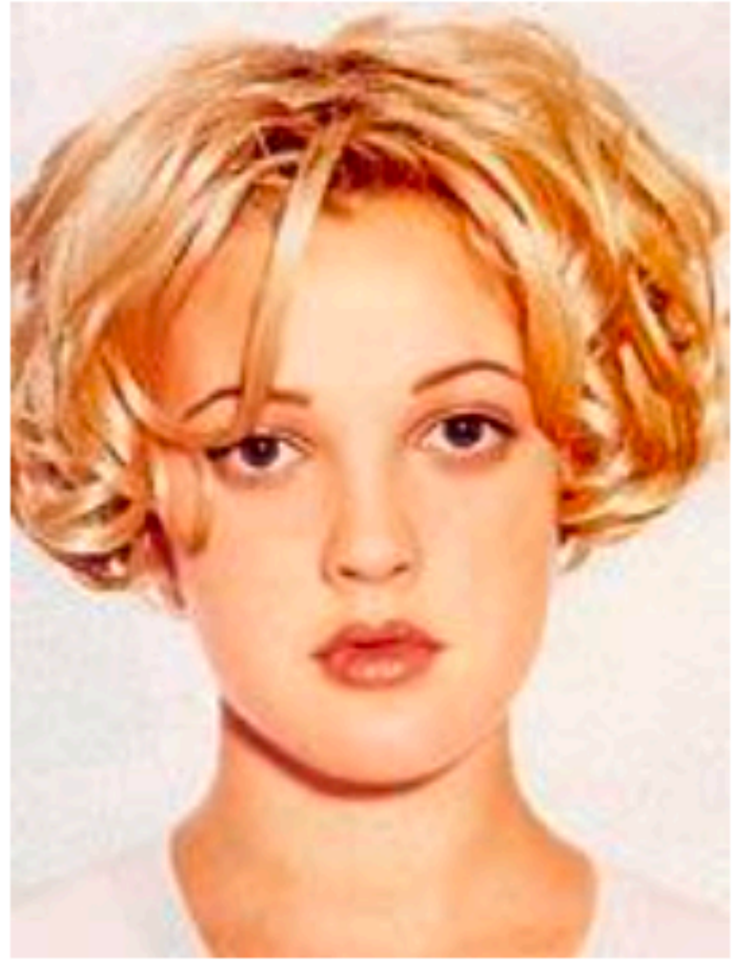
Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

Assume we have two classes: $c_1 = \textbf{male}$ $c_2 = \textbf{female}$

We have a person who's gender we don't know, who's name is *drew*

Drew Carey



Drew Barrymore

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

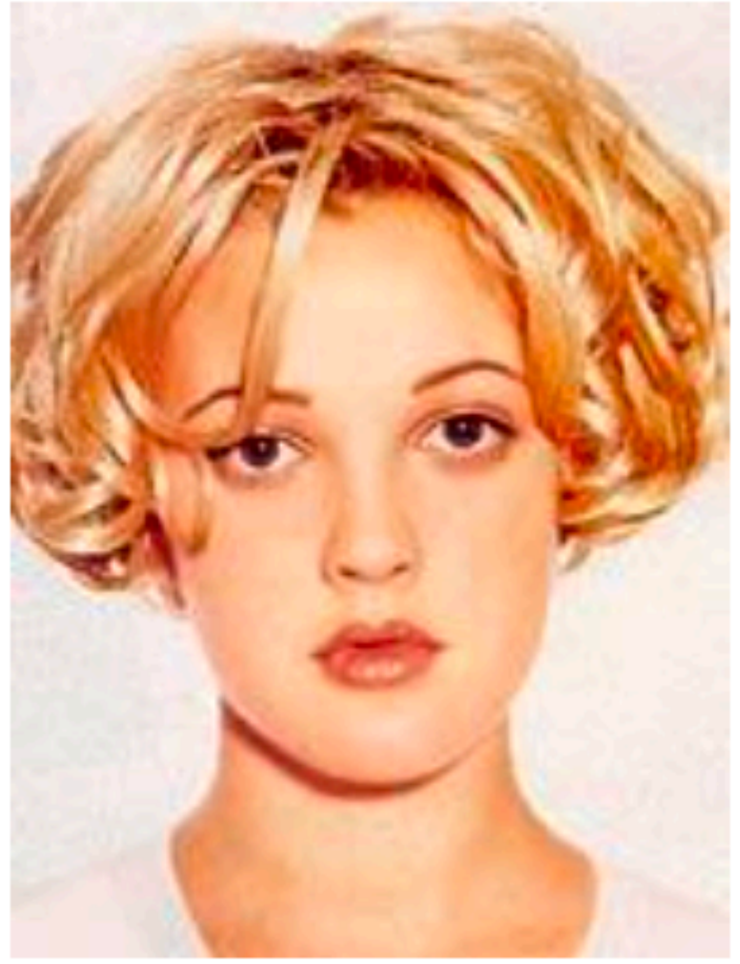
Assume we have two classes: $c_1 = \textbf{male}$ $c_2 = \textbf{female}$

We have a person who's gender we don't know, who's name is *drew*

Classifying drew as being male or female is equivalent to asking is it more probable that *drew* is male or female, i.e. which is greater $p(\textbf{male}|drew)$

$p(\textbf{female}|drew)$

Drew Carey



Drew Barrymore

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

Assume we have two classes: $c_1 = \textbf{male}$ $c_2 = \textbf{female}$

We have a person who's gender we don't know, who's name is *drew*

Classifying drew as being male or female is equivalent to asking is it more probable that *drew* is male or female, i.e. which is greater $p(\textbf{male}|drew)$

$p(\textbf{female}|drew)$

$$p(\textbf{male}|drew) = \frac{p(drew|\textbf{male})p(\textbf{male})}{p(drew)}$$

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\mathbf{male}|drew) = \frac{p(drew|\mathbf{male})p(\mathbf{male})}{p(drew)}$$

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\mathbf{male}) =$$

$$p(drew|\mathbf{male}) =$$

$$p(drew) =$$

$$p(\mathbf{male}|drew) = \frac{p(drew|\mathbf{male})p(\mathbf{male})}{p(drew)}$$

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

$$p(\textbf{male}) = \frac{3}{8}$$

$$p(drew|\textbf{male}) =$$

$$p(drew) =$$

$$p(\textbf{male}|drew) = \frac{p(drew|\textbf{male})p(\textbf{male})}{p(drew)}$$

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

$$p(\mathbf{male}) = \frac{3}{8}$$

$$p(drew|\mathbf{male}) = \frac{1}{3}$$

$$p(drew) =$$

$$p(\mathbf{male}|drew) = \frac{p(drew|\mathbf{male})p(\mathbf{male})}{p(drew)}$$

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

$$p(\mathbf{male}) = \frac{3}{8}$$

$$p(drew|\mathbf{male}) = \frac{1}{3}$$

$$p(drew) = \frac{3}{8}$$

$$p(\mathbf{male}|drew) = \frac{p(drew|\mathbf{male})p(\mathbf{male})}{p(drew)}$$

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\mathbf{male}) = \frac{3}{8}$$

$$p(drew|\mathbf{male}) = \frac{1}{3}$$

$$\cancel{p(drew) = \frac{3}{8}}$$

$$p(\mathbf{male}|drew) = \frac{p(drew|\mathbf{male})p(\mathbf{male})}{\cancel{p(drew)}} = 0.125$$

Example from: Eamonn Keogh

Example: Discrete Bayes Classifier

Name	Gender
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\mathbf{male}) = \frac{3}{8} \qquad p(\mathbf{female}) = \frac{5}{8}$$

$$p(drew|\mathbf{male}) = \frac{1}{3} \qquad p(drew|\mathbf{female}) = \frac{2}{5}$$

$$\cancel{p(drew) = \frac{3}{8}}$$

$$p(\mathbf{male}|drew) = \frac{p(drew|\mathbf{male})p(\mathbf{male})}{\cancel{p(drew)}} = 0.125$$

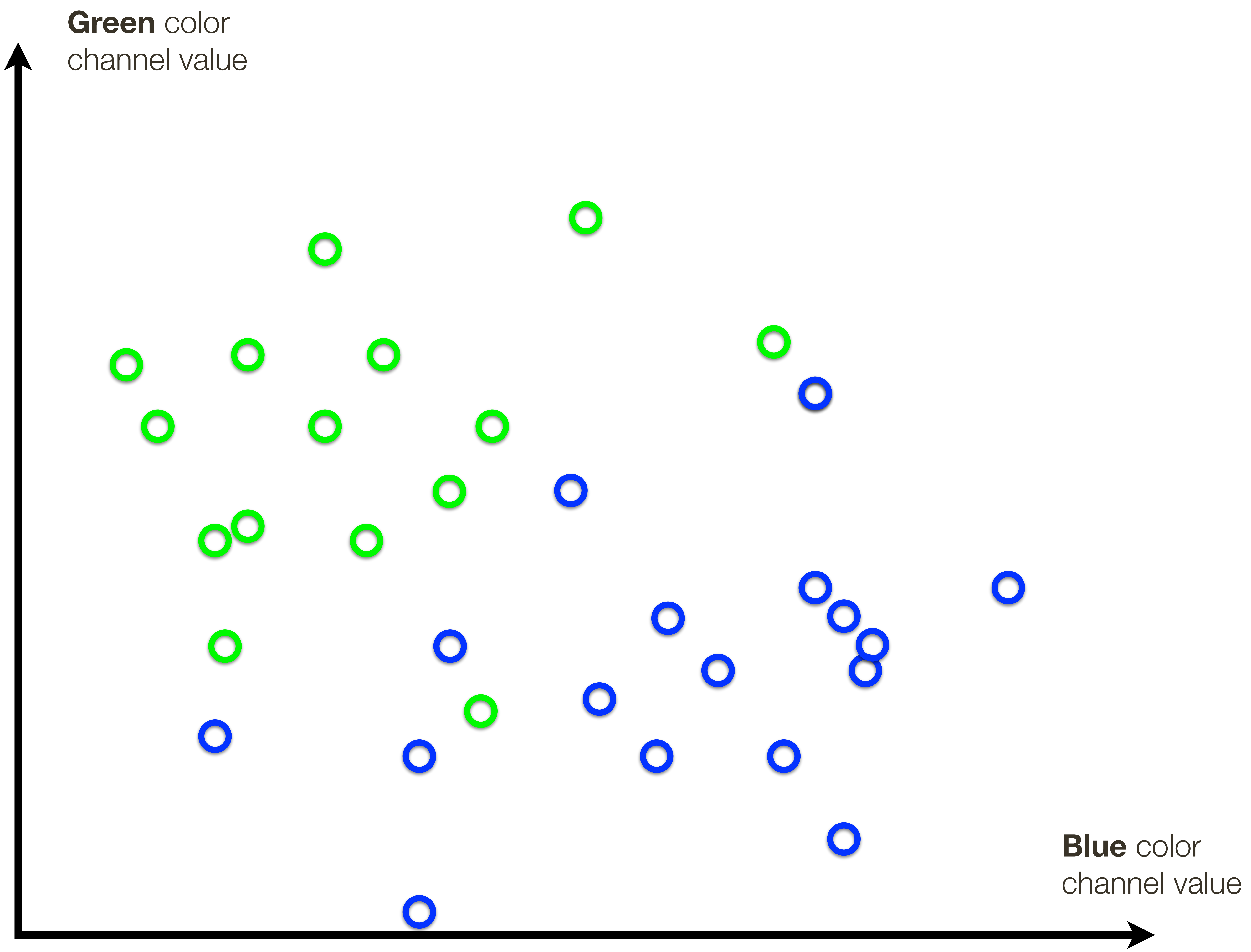
$$p(\mathbf{female}|drew) = \frac{p(drew|\mathbf{female})p(\mathbf{female})}{\cancel{p(drew)}} = 0.25$$

Example from: Eamonn Keogh

Example: 2D Bayes Classifier

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

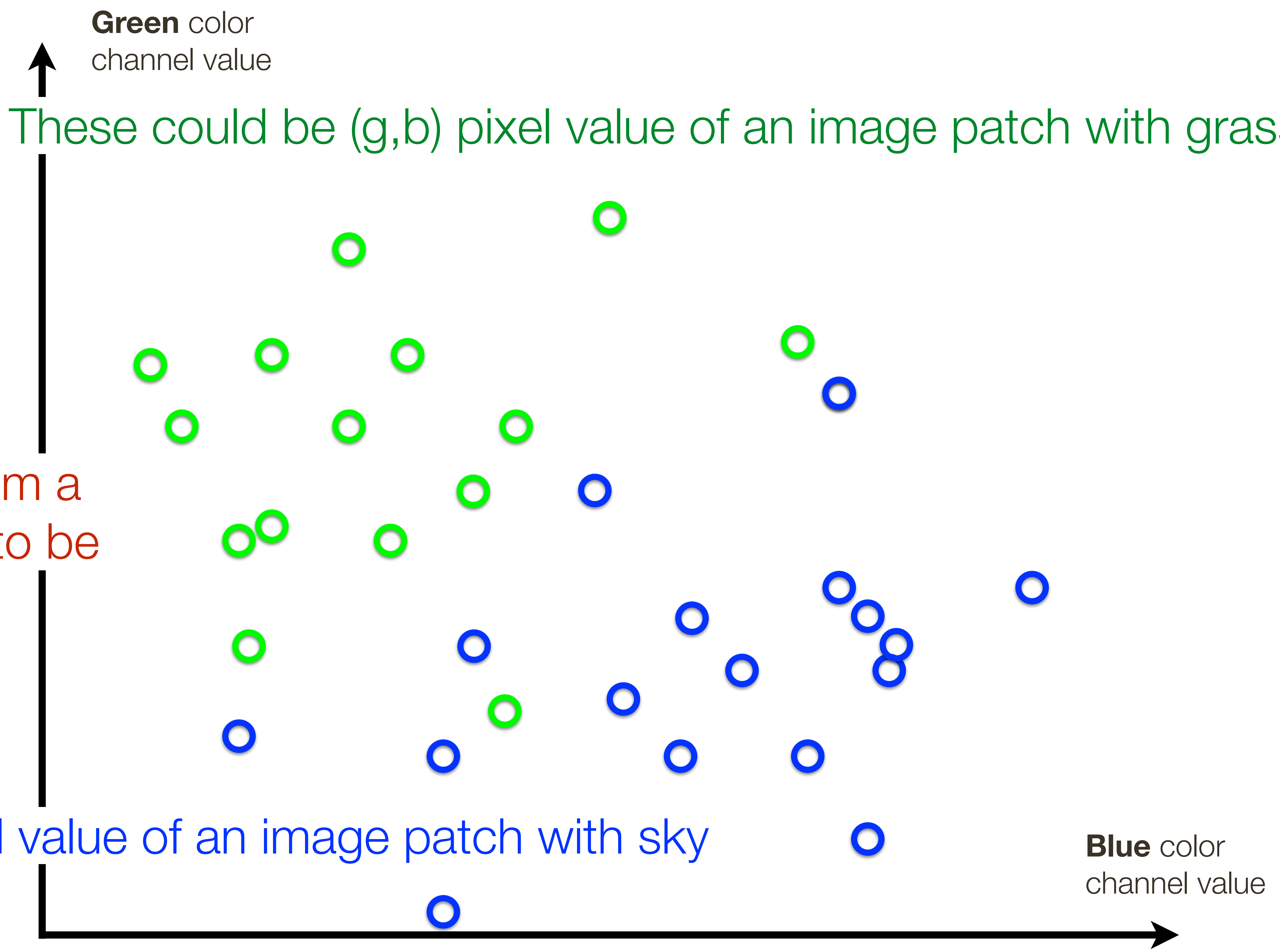
- 17 samples of grass
- 15 samples of sky



Example: 2D Bayes Classifier

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

- 17 samples of grass
- 15 samples of sky



Given a (g,b) pixel value from a new patch is it more likely to be be grass or sky?

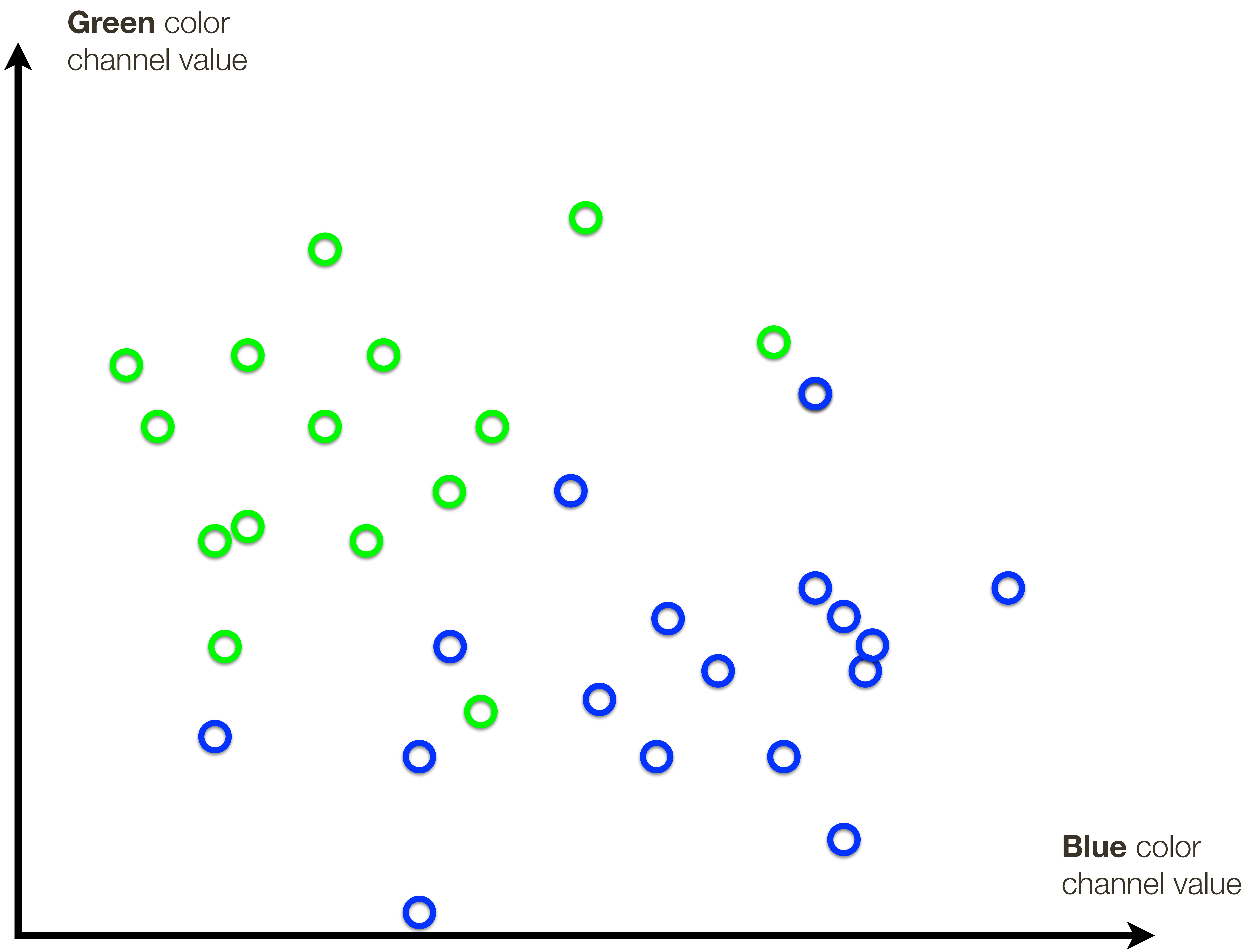
Example: 2D Bayes Classifier

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

- 17 samples of grass
- 15 samples of sky

$$p(blue) = \frac{17}{17+15}$$

$$p(green) = \frac{15}{17+15}$$



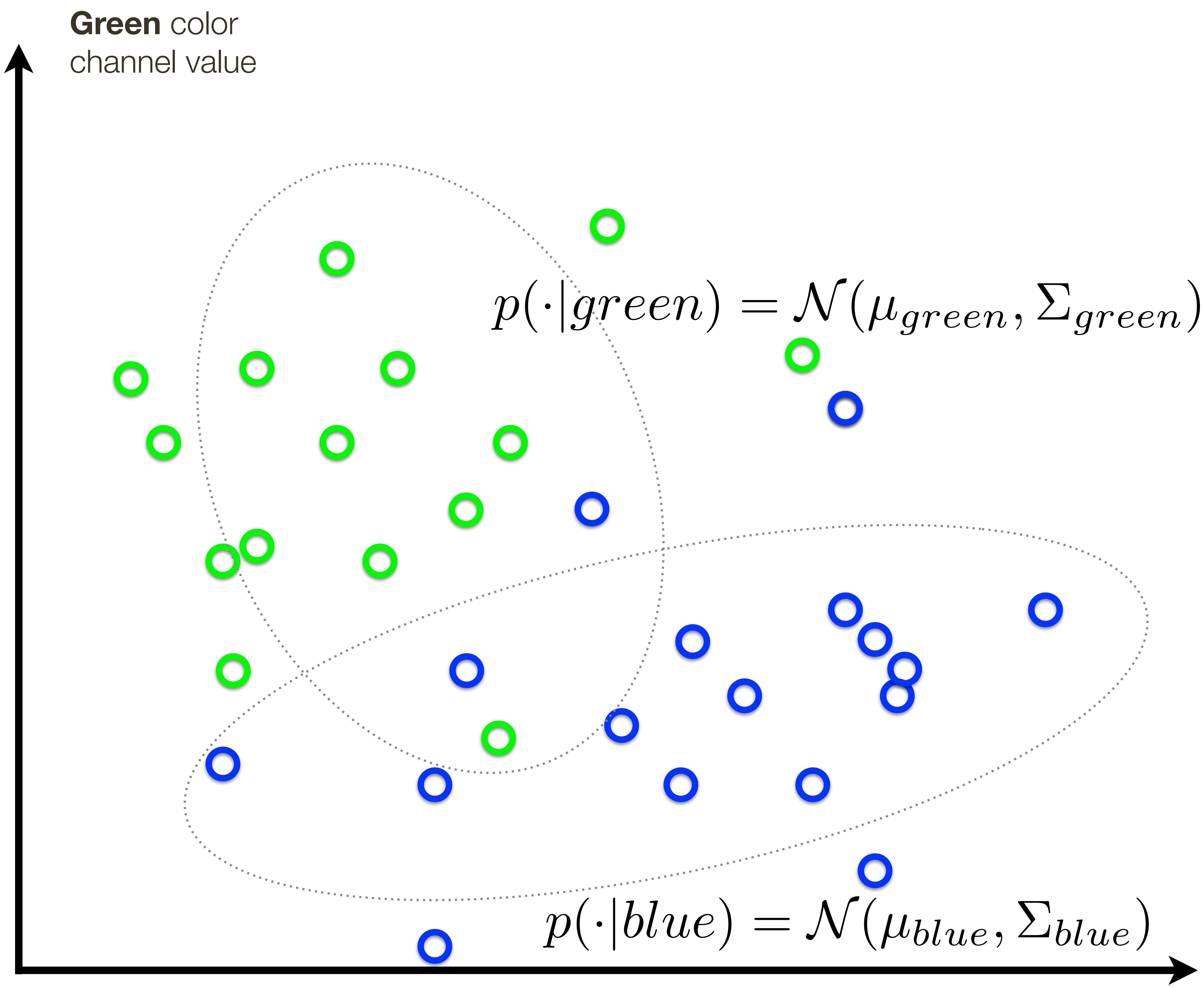
Example: 2D Bayes Classifier

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

- 17 samples of grass
- 15 samples of sky

$$p(blue) = \frac{17}{17+15}$$

$$p(green) = \frac{15}{17+15}$$



Example: 2D Bayes Classifier

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

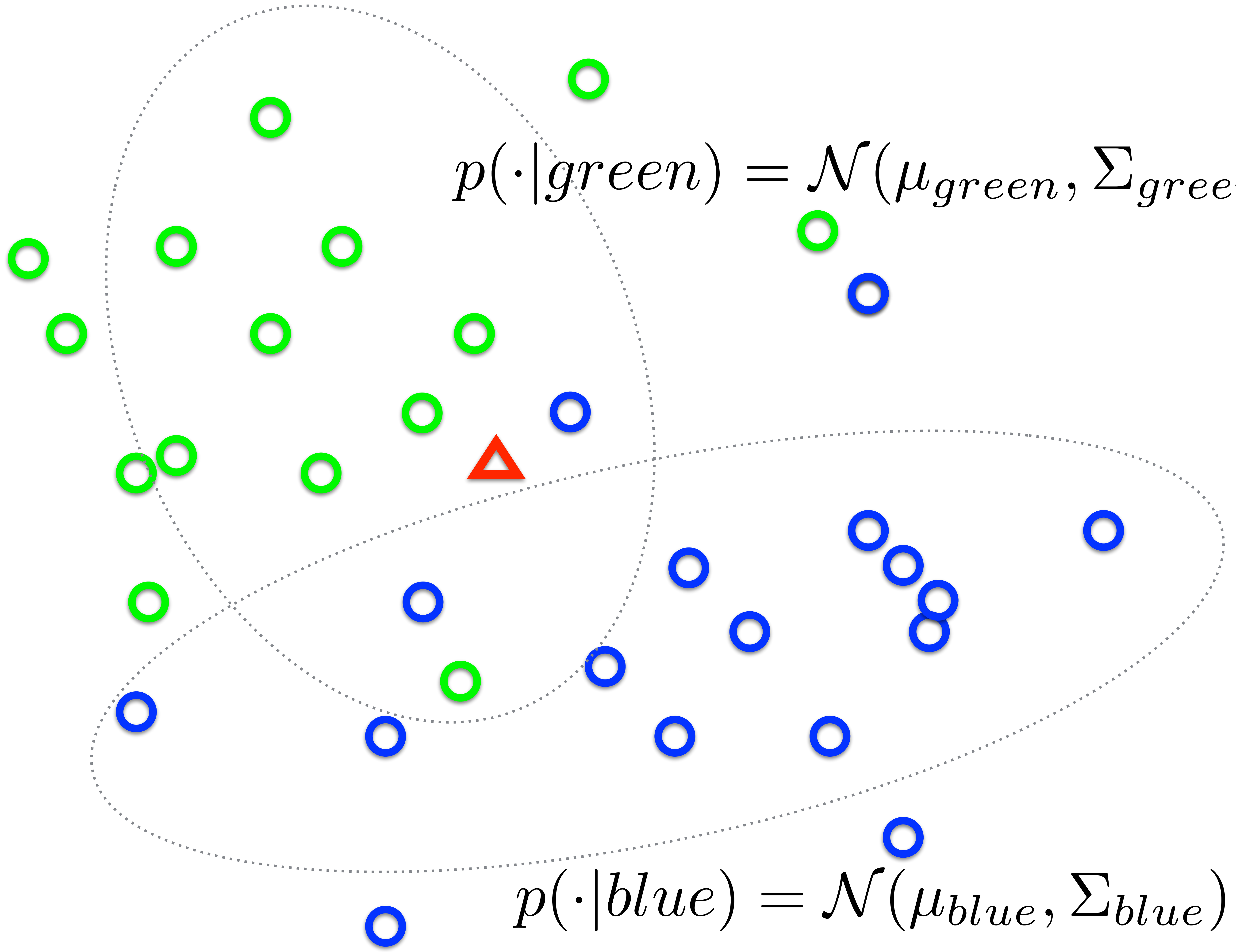
$$p(green|\triangle) \propto \mathcal{N}(\triangle; \mu_{green}, \Sigma_{green})p(green)$$

$$p(blue|\triangle) \propto \mathcal{N}(\triangle; \mu_{blue}, \Sigma_{blue})p(blue)$$

- 17 samples of grass
- 15 samples of sky

$$p(blue) = \frac{17}{17+15}$$

$$p(green) = \frac{15}{17+15}$$



$$p(\cdot|green) = \mathcal{N}(\mu_{green}, \Sigma_{green})$$

$$p(\cdot|blue) = \mathcal{N}(\mu_{blue}, \Sigma_{blue})$$

Bayes Rule (Review and Definitions)

Let c be the **class label** and let x be the **measurement** (i.e., evidence)

Simple case:

— binary classification; i.e., $c \in \{1, 2\}$

— features are 1D; i.e., $x \in \mathbb{R}$

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)}$$

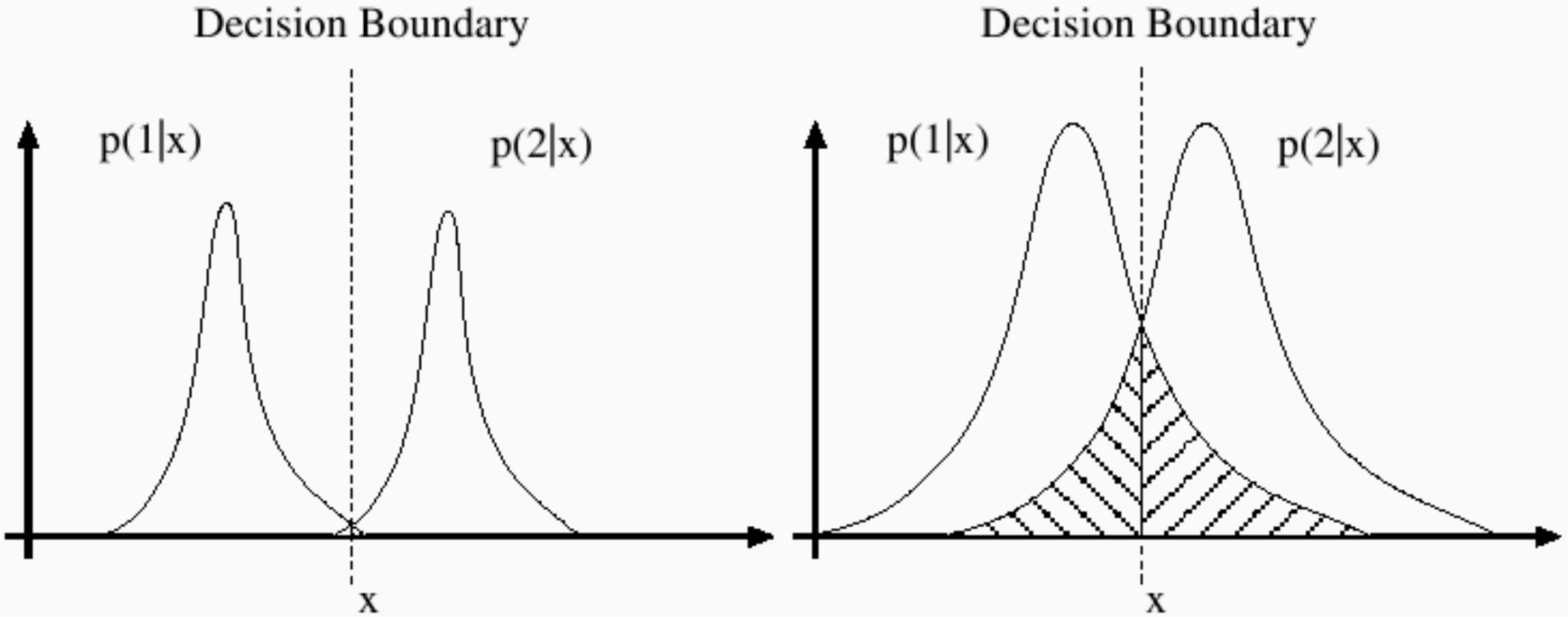
General case:

— multi-class; i.e., $c \in \{1, ..., 1000\}$

— features are high-dimensional; i.e., $x \in \mathbb{R}^{2,000+}$

Bayes' Risk

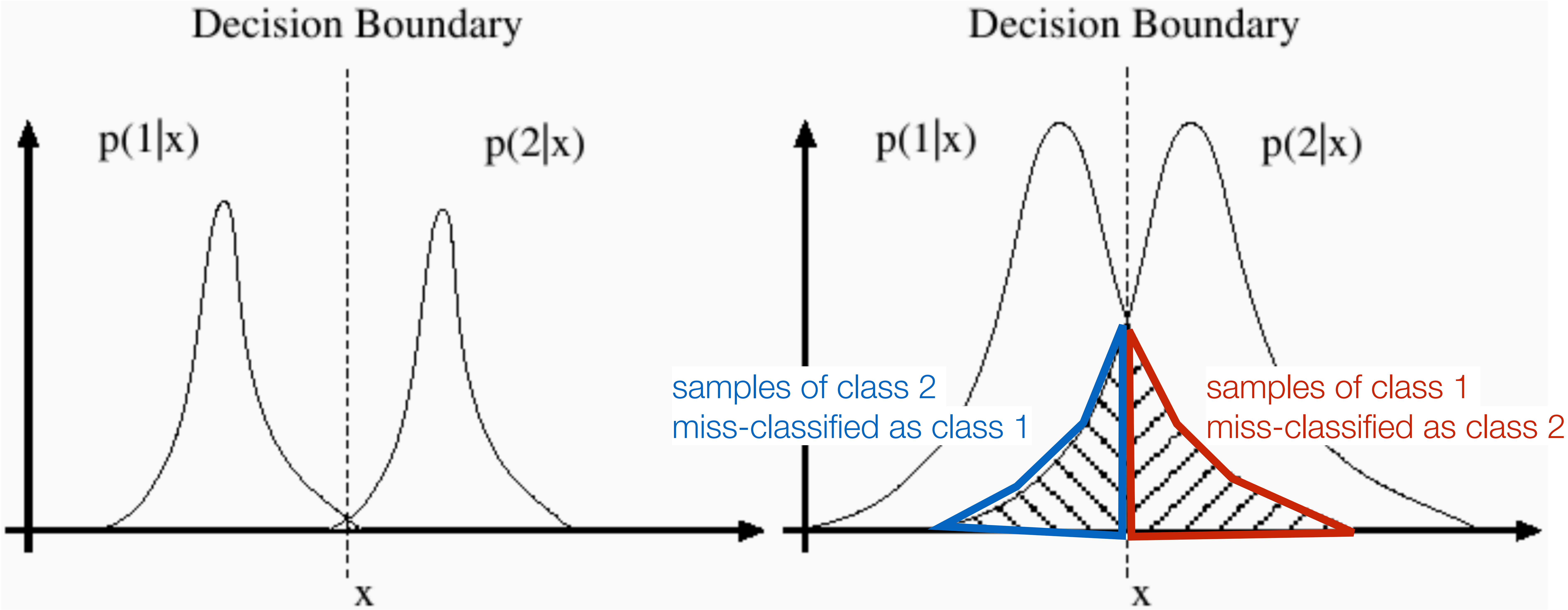
Some errors may be inevitable: the minimum risk (shaded area) is called the **Bayes' risk**



Forsyth & Ponce (2nd ed.) Figure 15.1

Bayes' Risk

Some errors may be inevitable: the minimum risk (shaded area) is called the **Bayes' risk**



Forsyth & Ponce (2nd ed.) Figure 15.1

Loss Functions and Classifiers

Loss

— Some errors may be more expensive than others

Example: A fatal disease that is easily cured by a cheap medicine with no side-effects. Here, false positives in diagnosis are better than false negatives

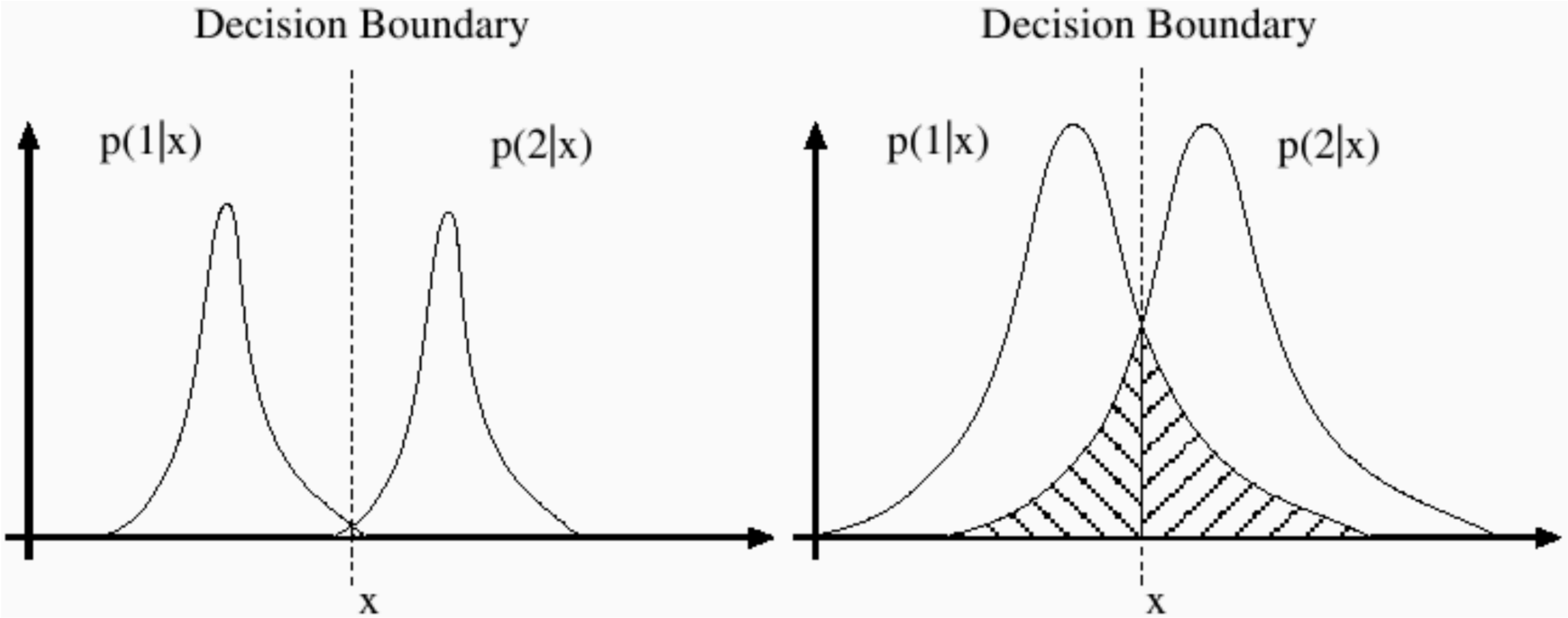
— We discuss two class classification:
$L(1 \rightarrow 2)$ is the loss caused by calling 1 a 2

Total risk of using classifier ***s*** is

$$R(s) = \underbrace{\Pr\{1 \rightarrow 2 \mid \text{using } \boldsymbol{s}\}}_{\text{Probability of Miss-classification}} \underbrace{L(1 \rightarrow 2)}_{\substack{\text{Loss} \\ \text{(i.e. cost of miss-classification)}}} + \underbrace{\Pr\{2 \rightarrow 1 \mid \text{using } \boldsymbol{s}\}}_{\text{Probability of Miss-classification}} \underbrace{L(2 \rightarrow 1)}_{\substack{\text{Loss} \\ \text{(i.e. cost of miss-classification)}}}$$

Bayes' Risk

Some errors may be inevitable: the minimum risk (shaded area) is called the **Bayes' risk**



Forsyth & Ponce (2nd ed.) Figure 15.1

Classifier Strategies

Classification strategies fall under two broad types: **parametric** and **non-parametric**.

Classifier Strategies

Classification strategies fall under two broad types: **parametric** and **non-parametric**.

Parametric classifiers are **model driven**. The parameters of the model are learned from training examples. New data points are classified by the learned model.

— fast, compact

— flexibility and accuracy depend on model assumptions

Classifier Strategies

Classification strategies fall under two broad types: **parametric** and **non-parametric**.

Parametric classifiers are **model driven**. The parameters of the model are learned from training examples. New data points are classified by the learned model.

— fast, compact

— flexibility and accuracy depend on model assumptions

Non-parametric classifiers are **data driven**. New data points are classified by comparing to the training examples directly. "The data is the model".

— slow

— highly flexible decision boundaries

Nearest Neighbor Classifier

Given a new data point, assign the label of nearest training example in feature space.

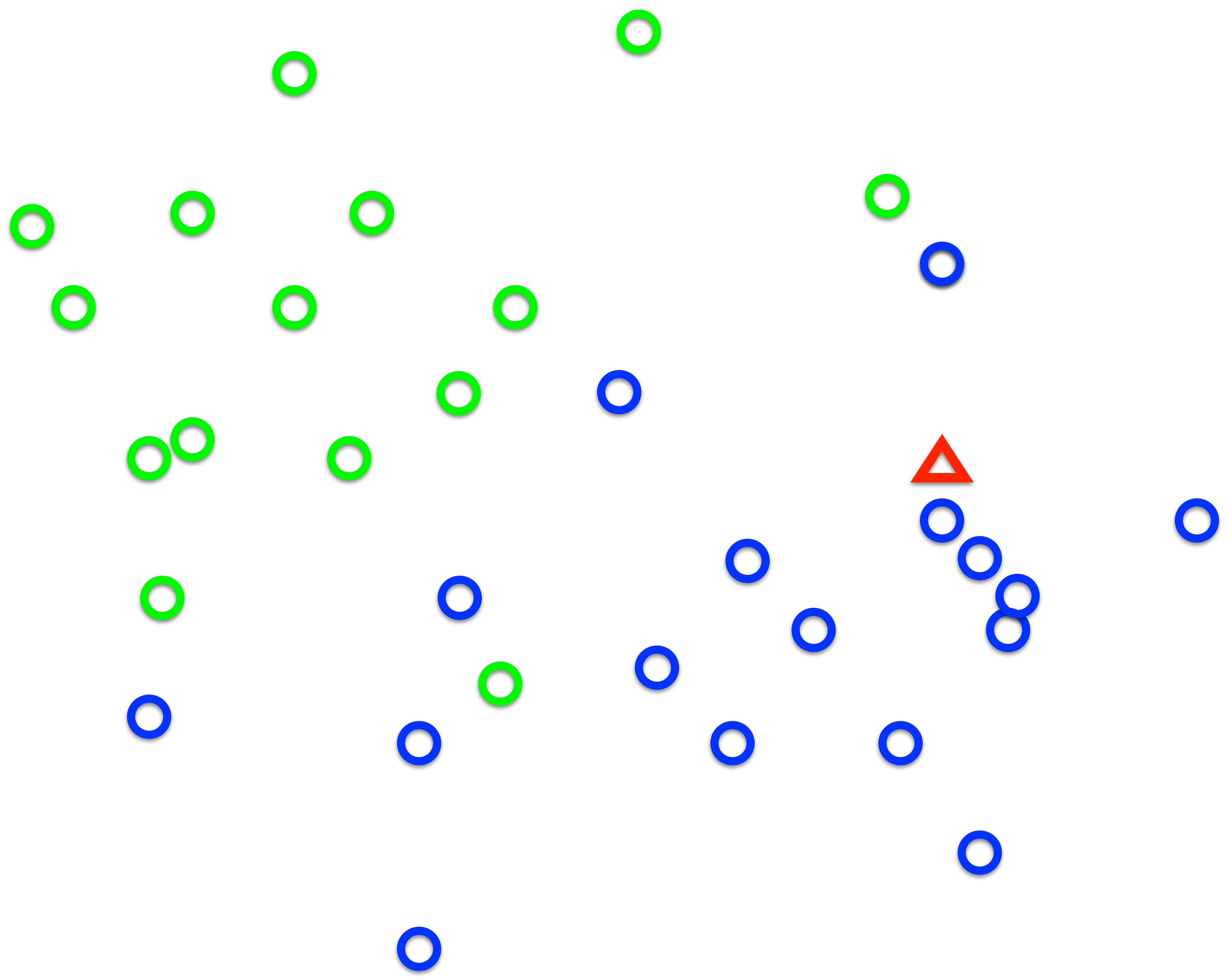


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Nearest Neighbor Classifier

Given a new data point, assign the label of nearest training example in feature space.

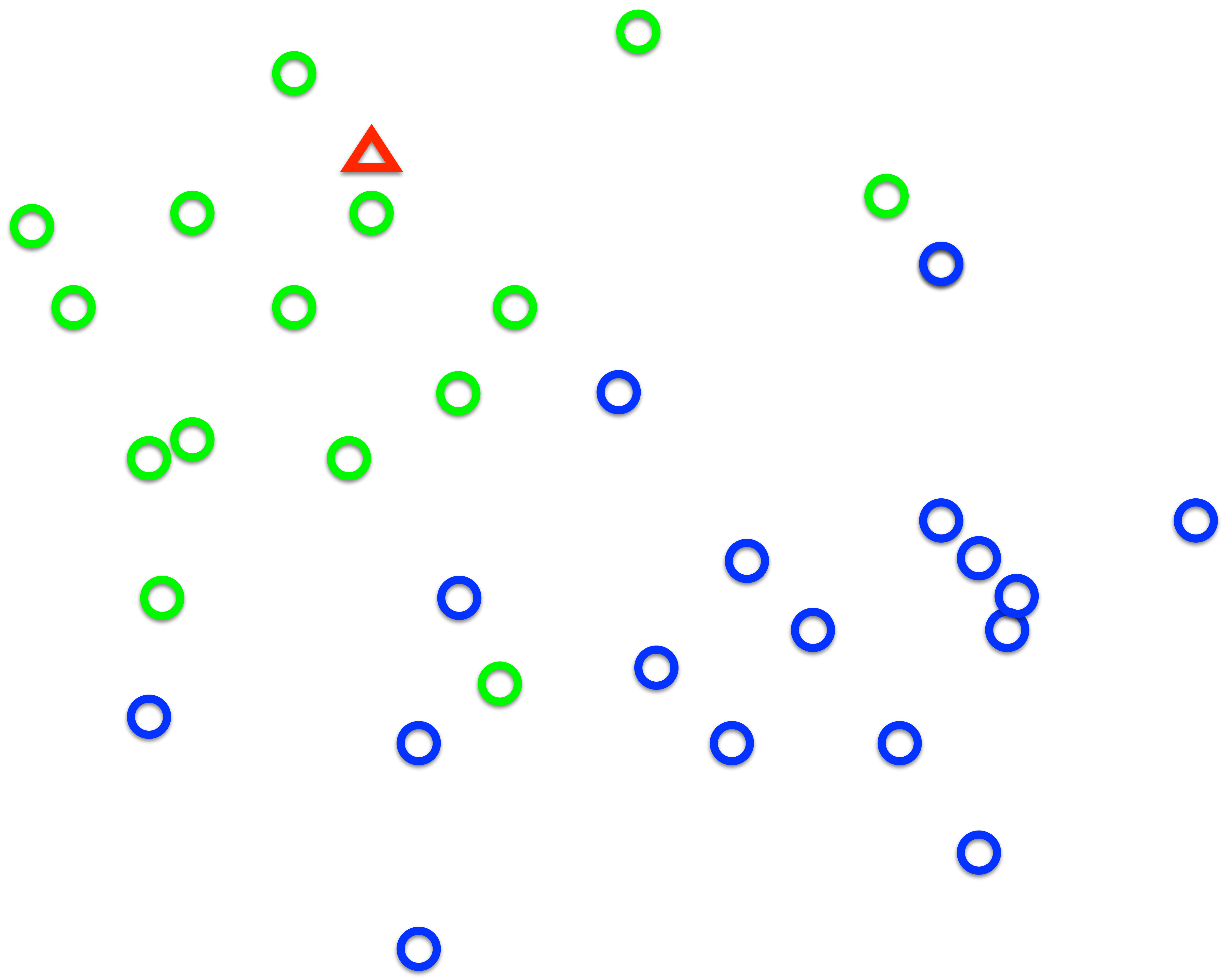


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

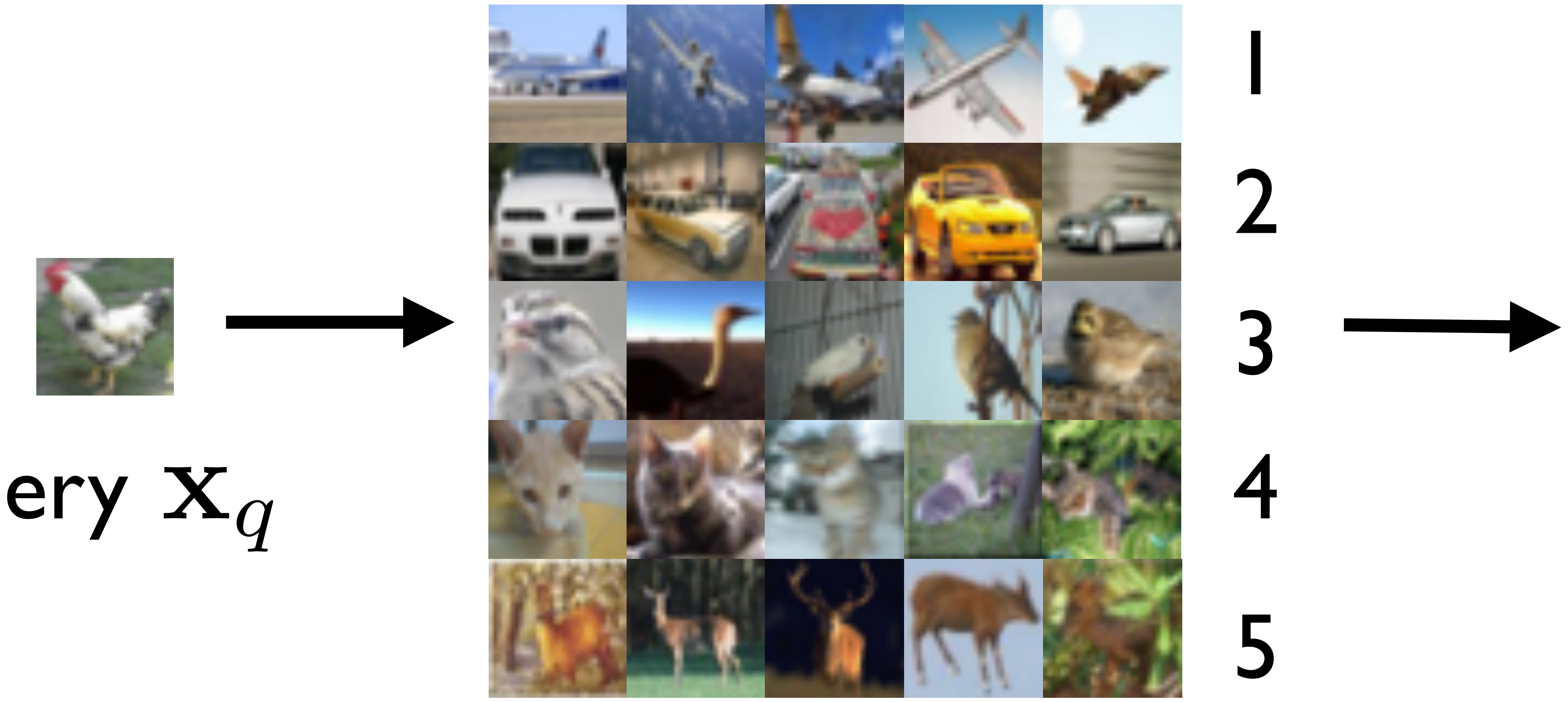
Nearest Neighbor Classifier

Find nearest neighbour in training set

$$i_{NN} = \arg\min_i |\mathbf{x}_q - \mathbf{x}_i|$$

Assign class to class of the nearest neighbour

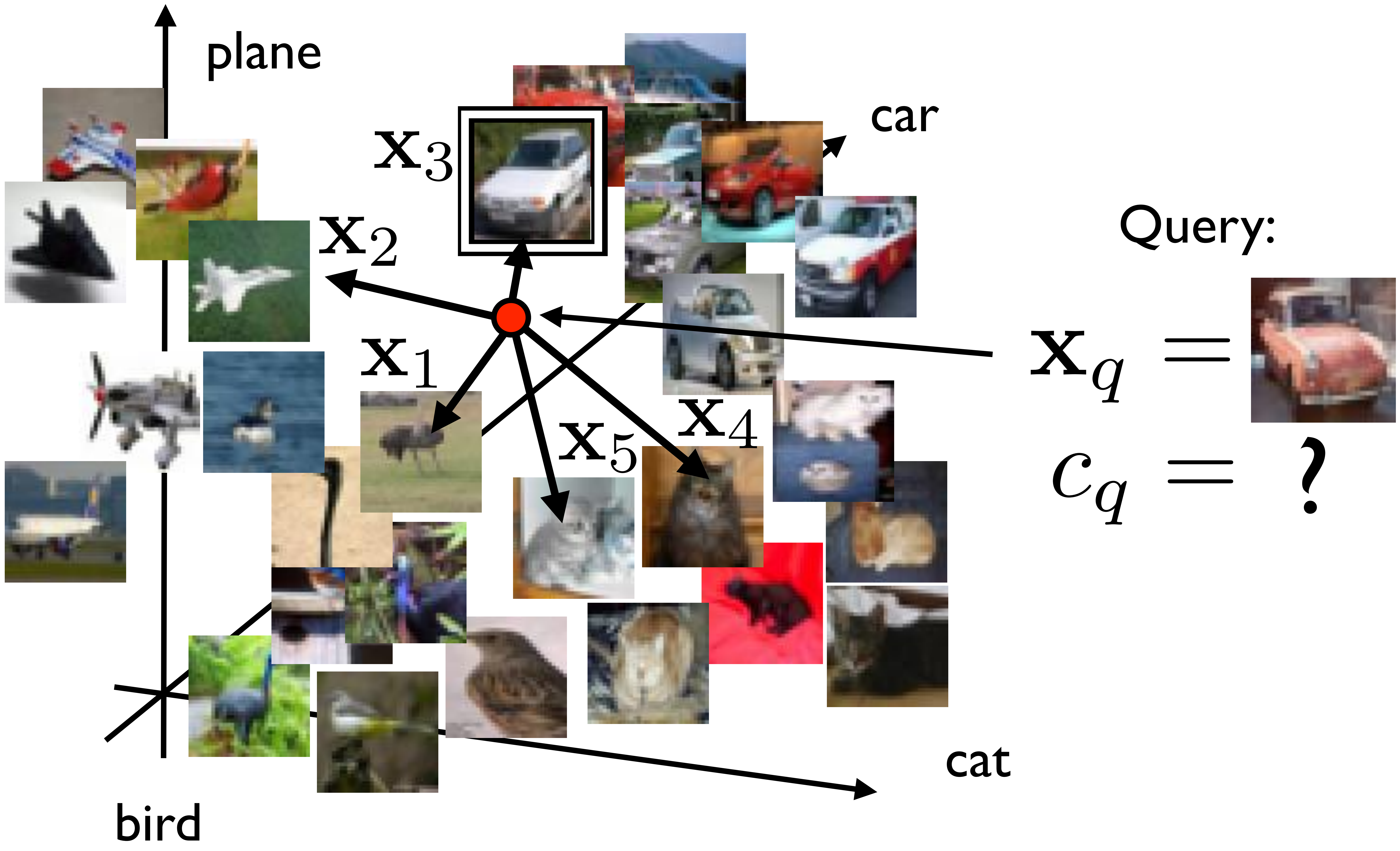
$$\hat{y}(\mathbf{x}_q) = y(\mathbf{x}_{i_{NN}})$$



Calculate $|\mathbf{x}_q - \mathbf{x}_i|$
for all training data

Nearest Neighbor Classifier

We can view each image as a point in a high dimensional space

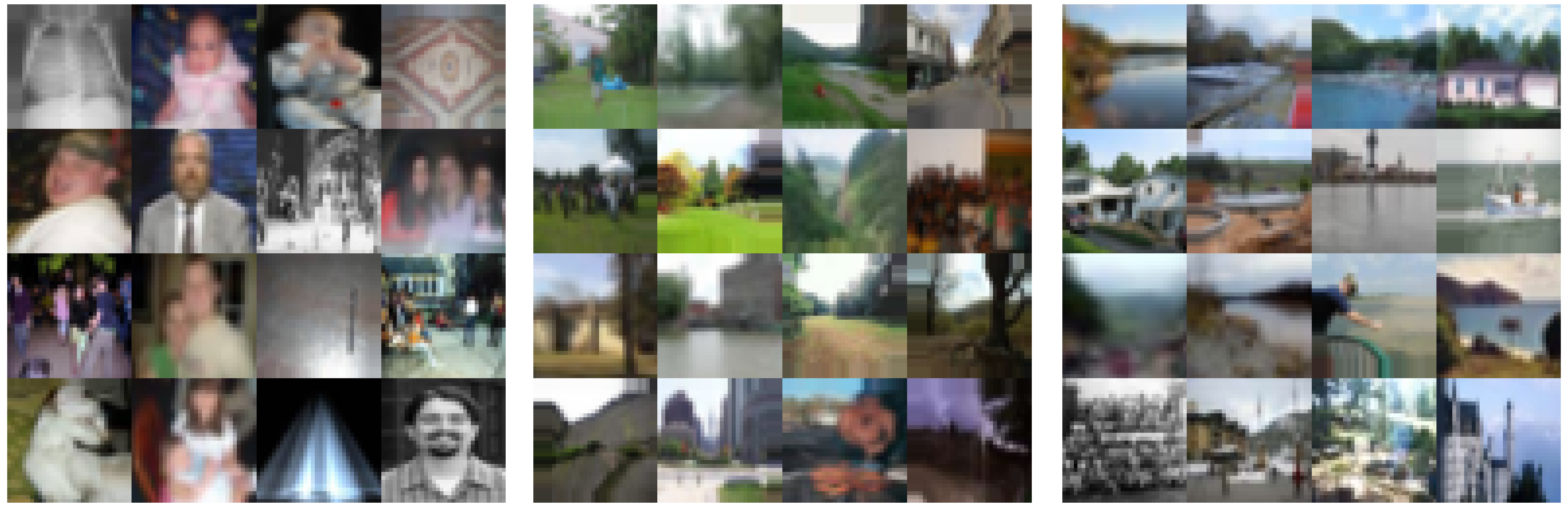


What do nearest neighbours look like with 80 million images?

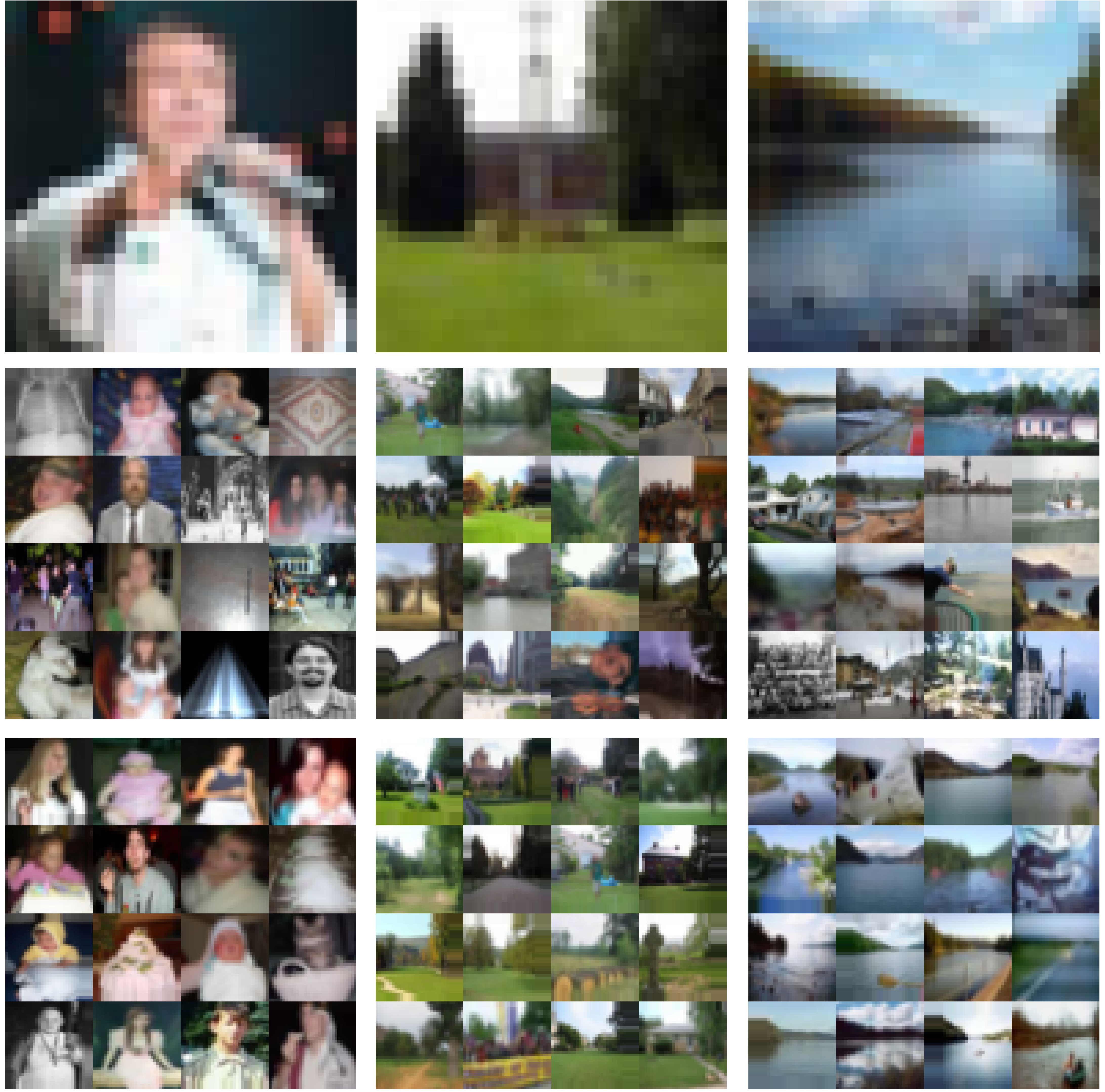
[Torralba, Fergus, Freeman ‘08]

Query

Query



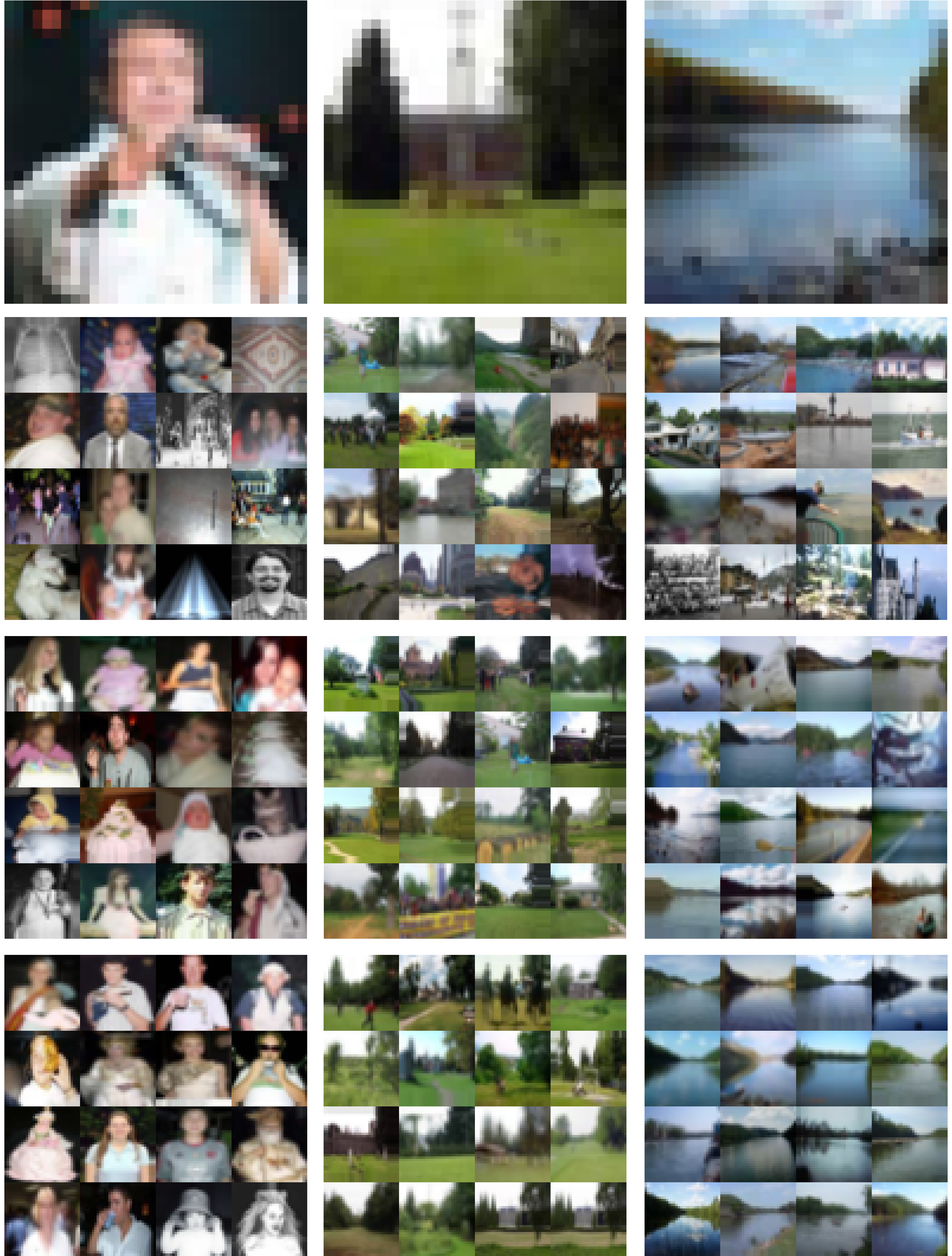
7900



Query

7900

790,000



Query

7900

790,000

79,000,000

k-**Nearest Neighbor** (kNN) Classifier

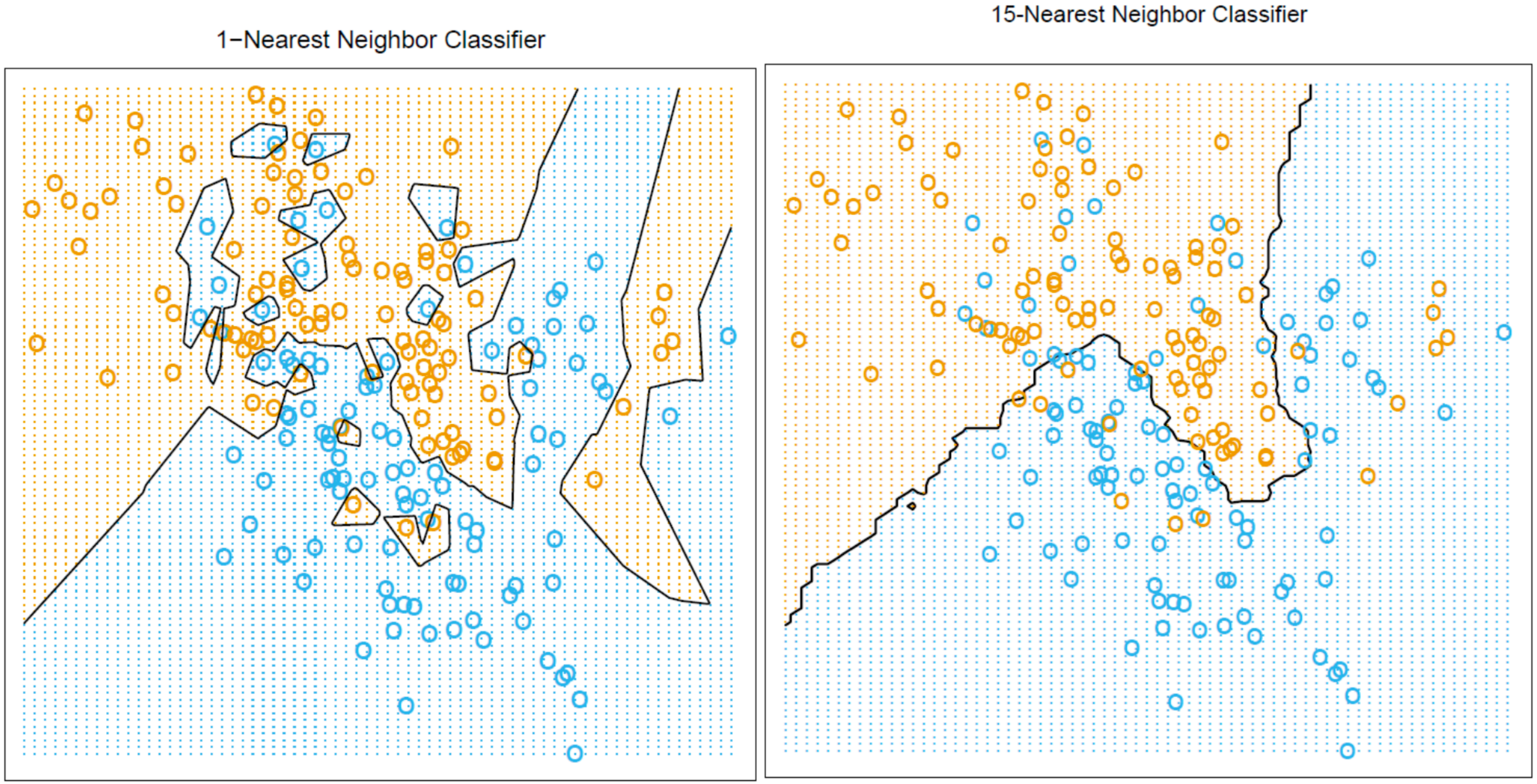
We can gain some robustness to noise by voting over **multiple** neighbours.

Given a **new** data point, find the k nearest training examples. Assign the label by **majority vote**.

Simple method that works well if the **distance measure** correctly weights the various dimensions

For **large data sets**, as k increases kNN approaches optimality in terms of minimizing probability of error

k-**Nearest Neighbor** (kNN) Classifier



kNN decision boundaries respond to local clusters where one class dominates

Figure credit: Hastie, Tibshirani & Friedman (2nd ed.)

Classifier Strategies

Classification strategies fall under two broad types: **parametric** and **non-parametric**.

Parametric classifiers are **model driven**. The parameters of the model are learned from training examples. New data points are classified by the learned model.
— fast, compact
— flexibility and accuracy depend on model assumptions

Non-parametric classifiers are **data driven**. New data points are classified by comparing to the training examples directly. "The data is the model".
— slow
— highly flexible decision boundaries

Support Vector Machines (SVM)

Idea: Try to obtain the decision boundary directly

The decision boundary is parameterized as a **separating hyperplane** in feature space.
— e.g. a separating line in 2D

We choose the hyperplane that is as far as possible from each class - that maximizes the distance to the closest point from either class.

Linear Classifier

Defines a score function:

$$f(\mathbf{x}_i, \mathbf{W}, \mathbf{b}) = \underbrace{\mathbf{W}}_{\text{weights (parameters)}}\overbrace{\mathbf{x}_i}^{\text{image features}} + \underbrace{\mathbf{b}}_{\text{bias vector}}$$

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Linear Classifier

Example with an image with 4 pixels, and 3 classes (cat/dog/ship)

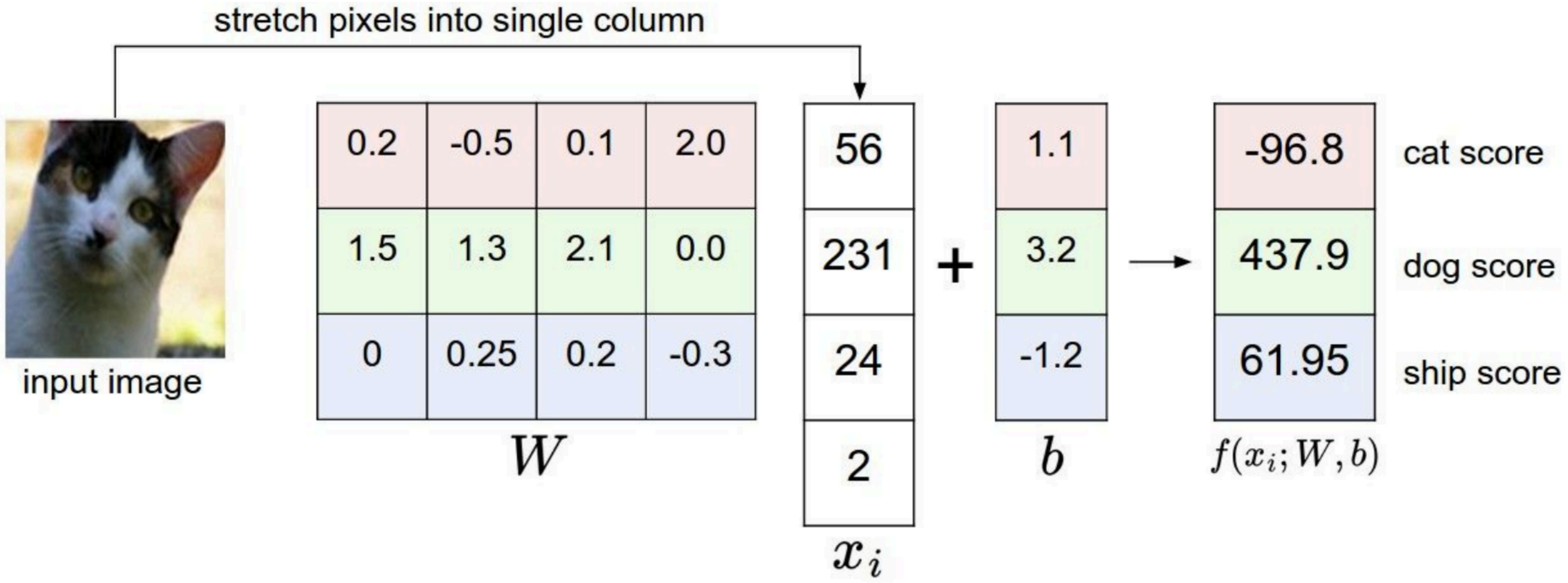


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

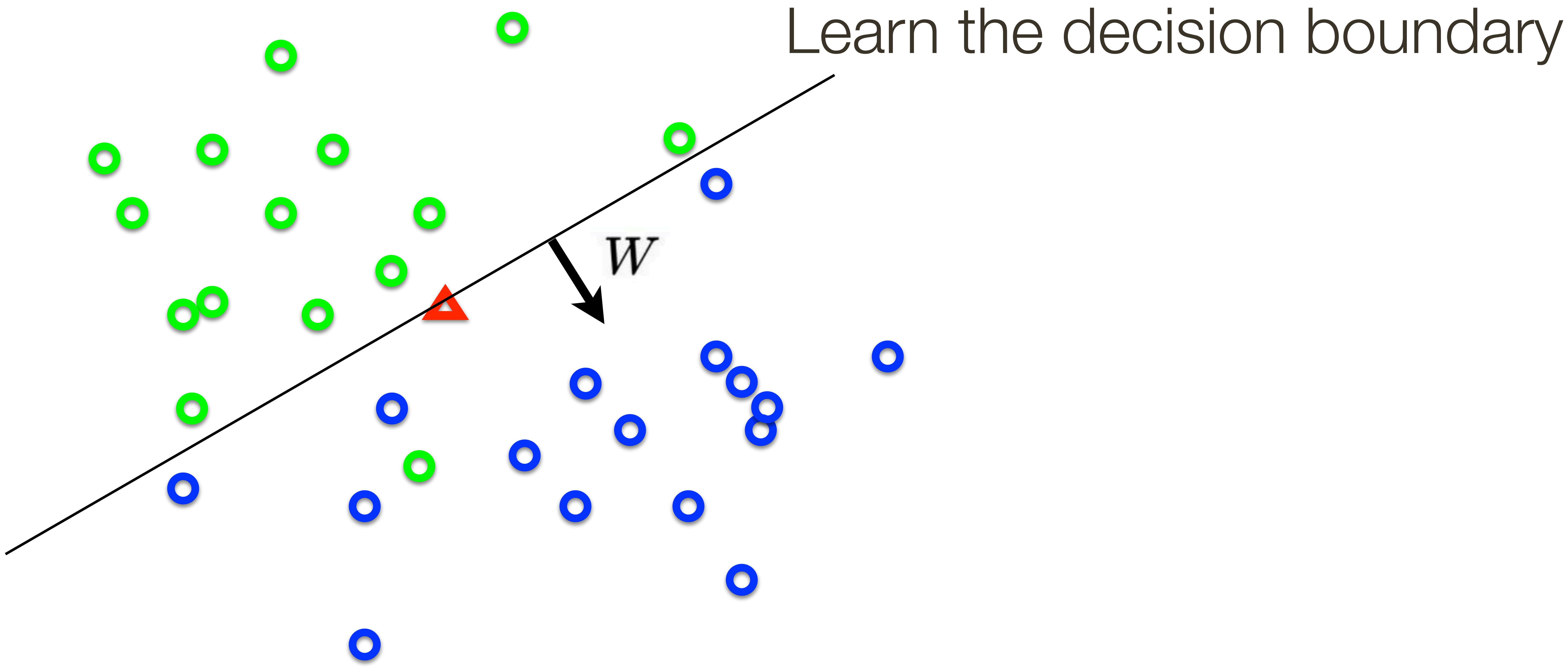


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What's the best **w** ?

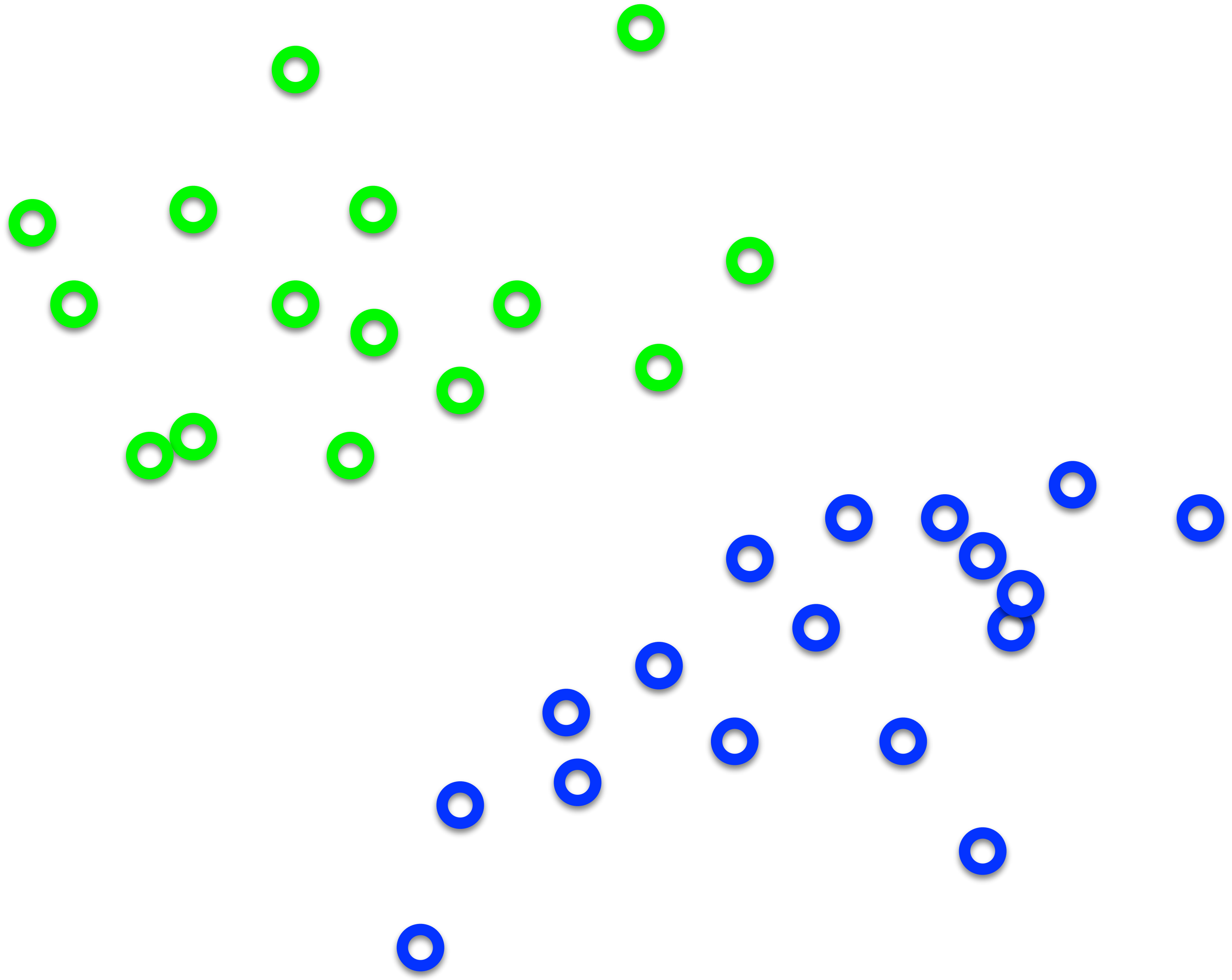


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What's the best **w** ?

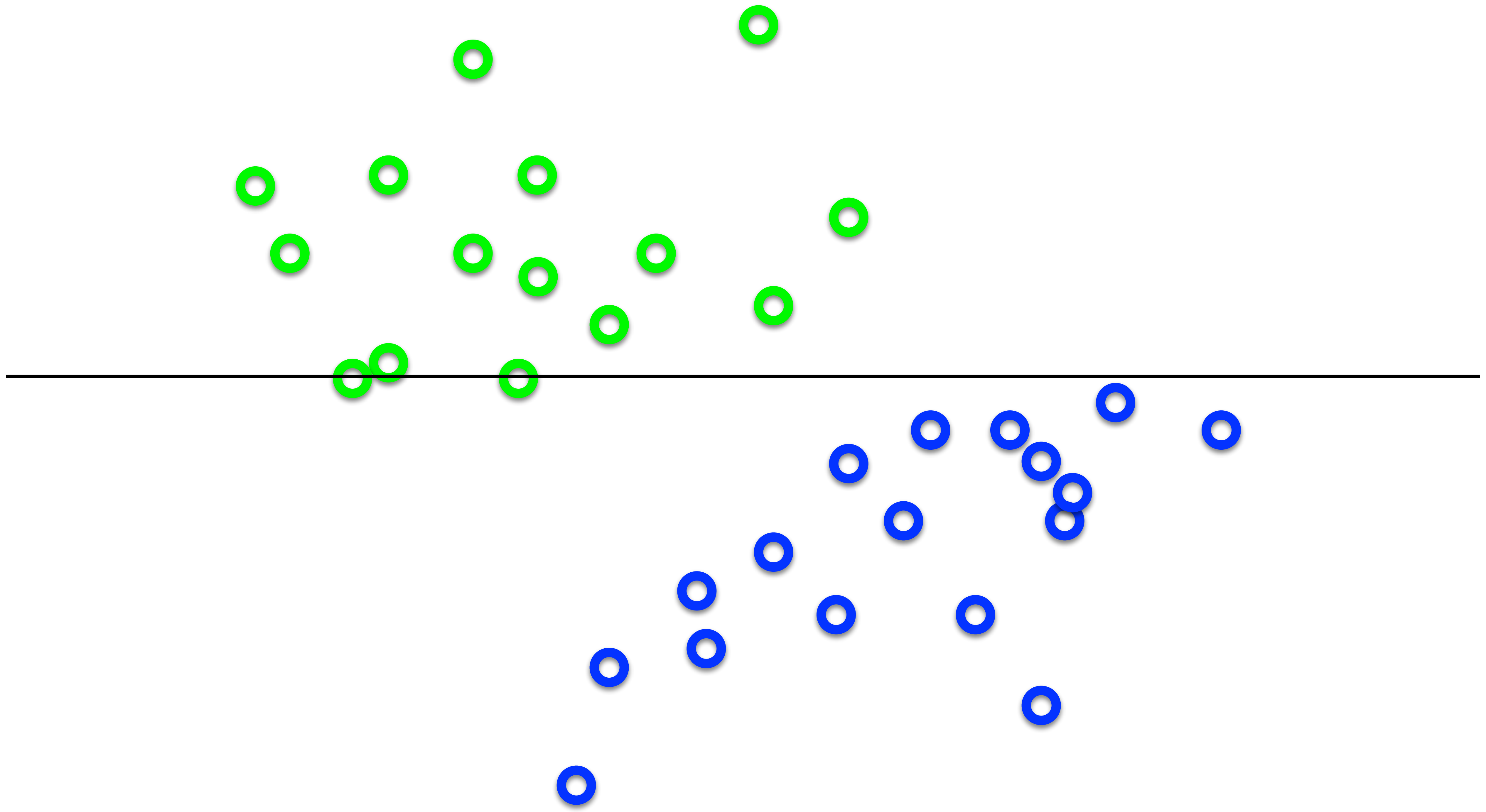


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What's the best **w** ?

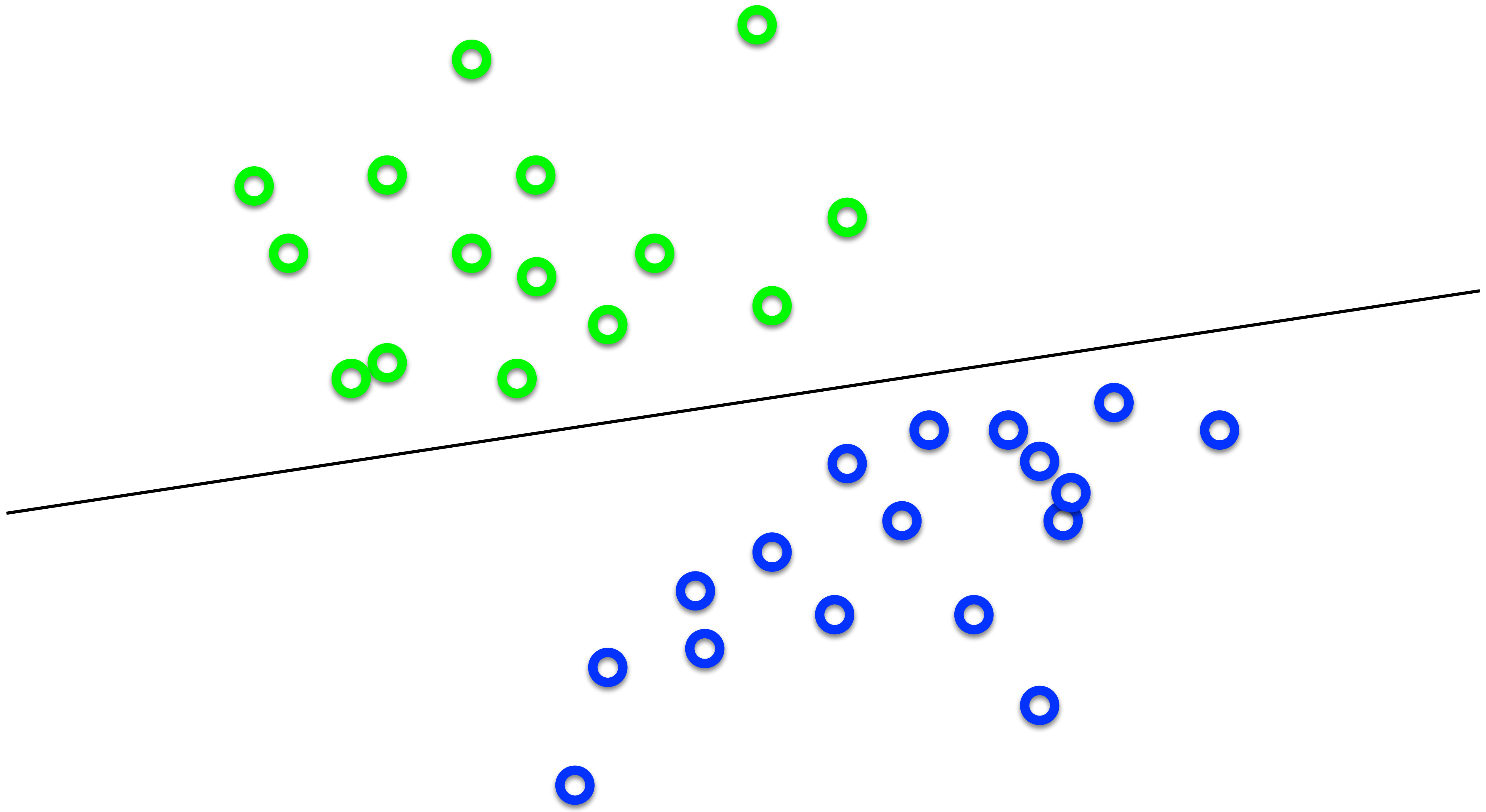


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What's the best **w** ?

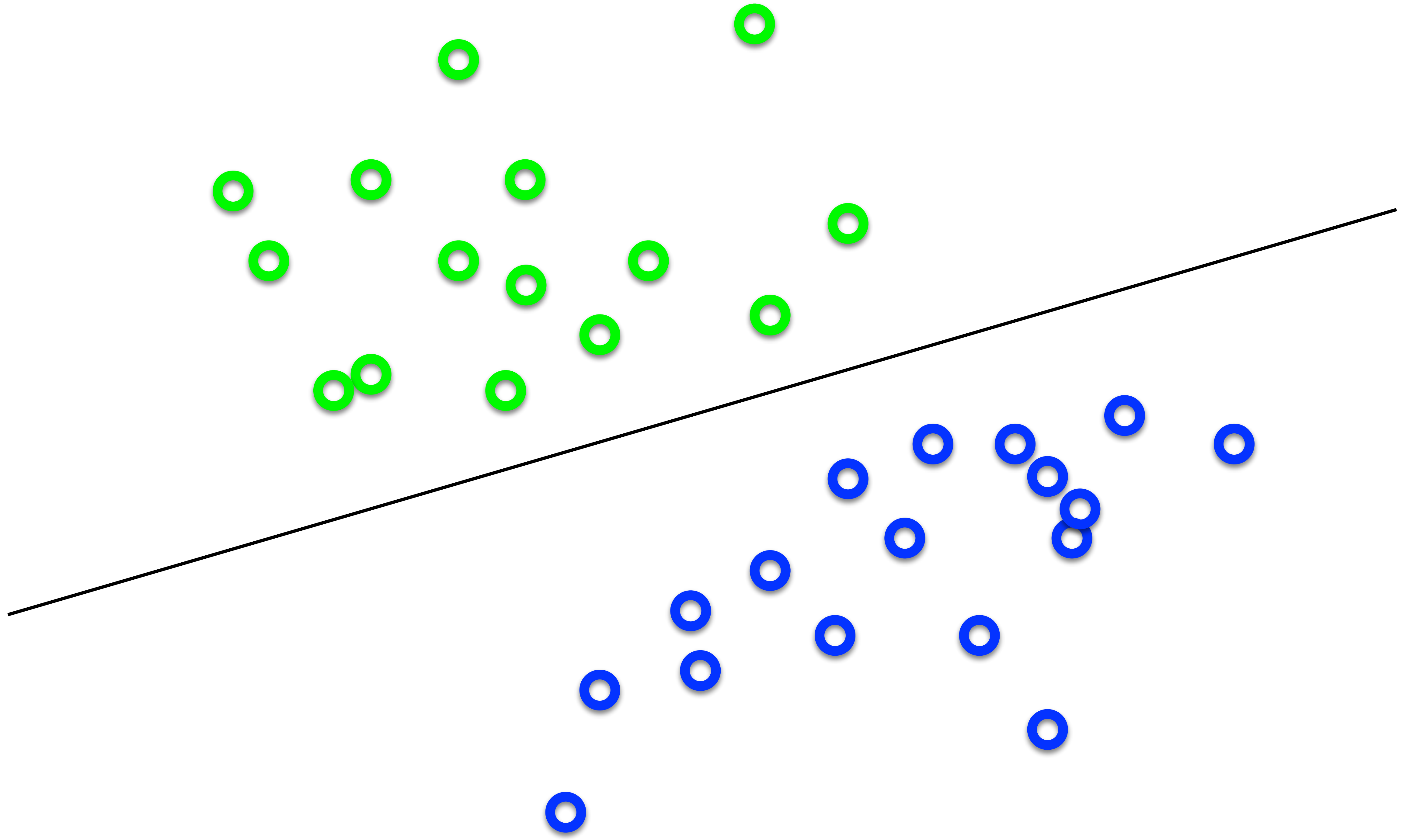


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What's the best **w** ?

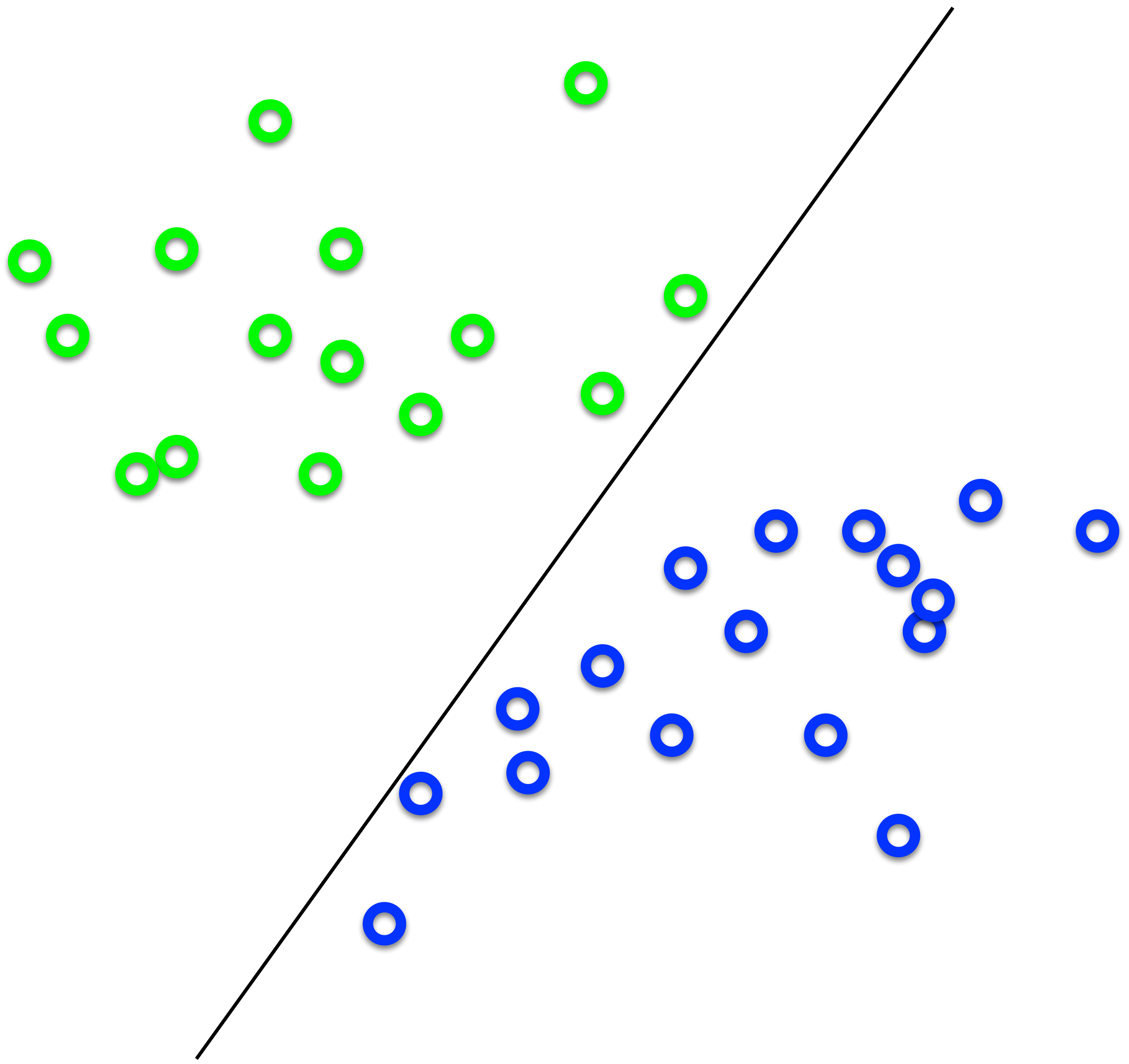
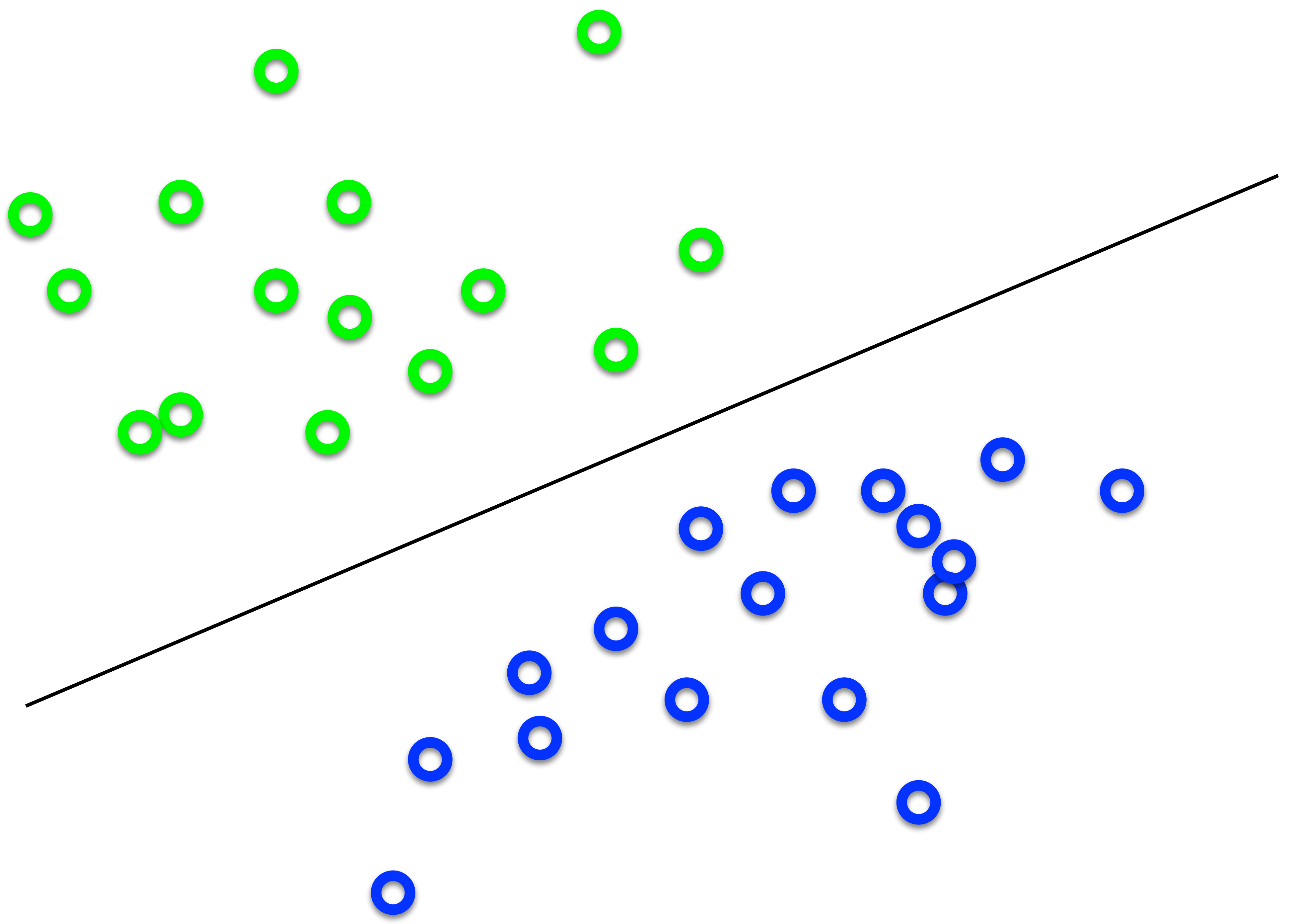


Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What’s the best **w** ?

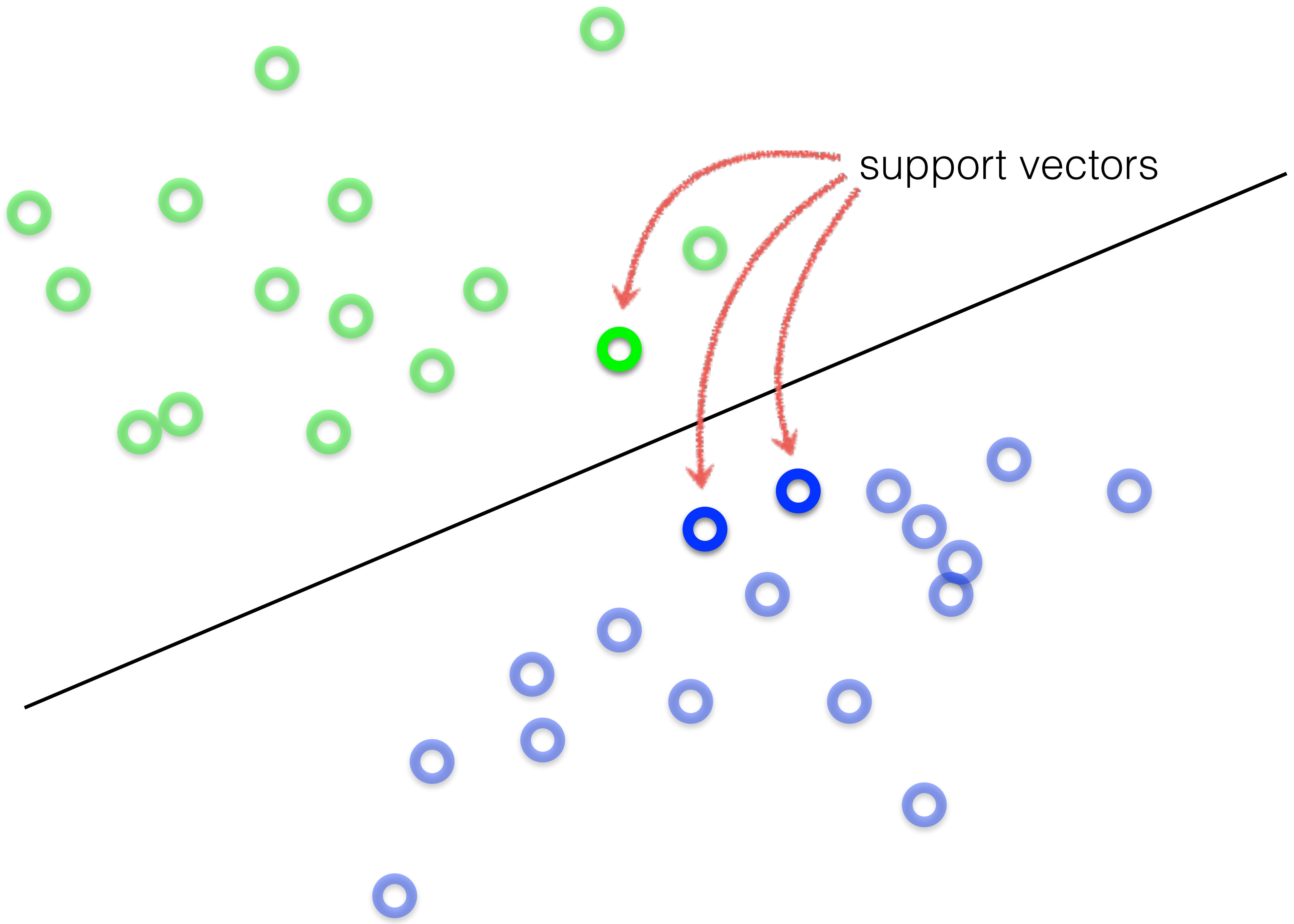


Intuitively, the line that is the farthest from all interior points

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

What's the best **w** ?

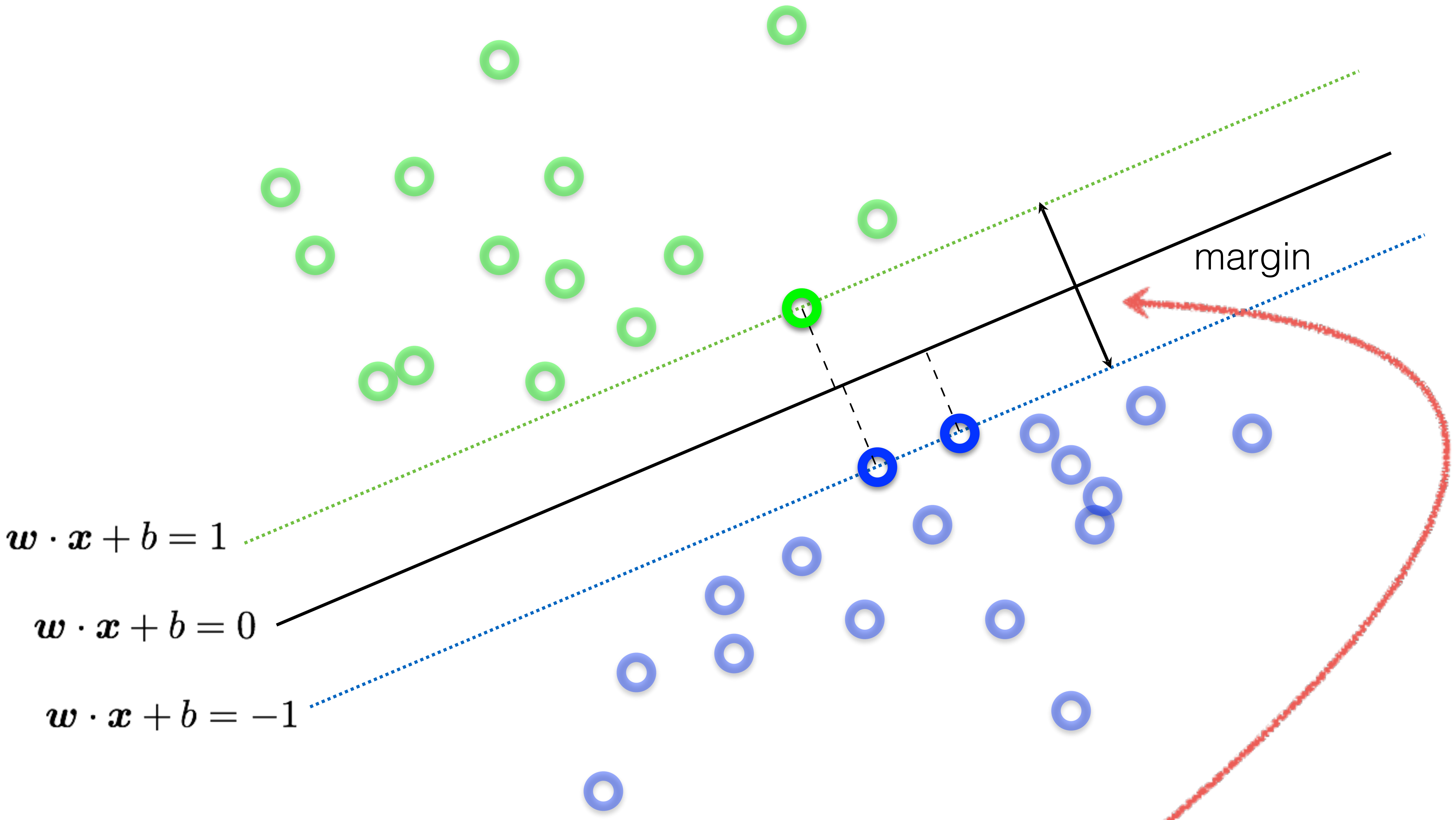


Want a hyperplane that is far away from 'inner points'

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Support Vector Machines (SVM)

Find hyperplane **w** such that …



the gap between parallel hyperplanes $\frac{2}{\|\boldsymbol{w}\|}$ is maximized

Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

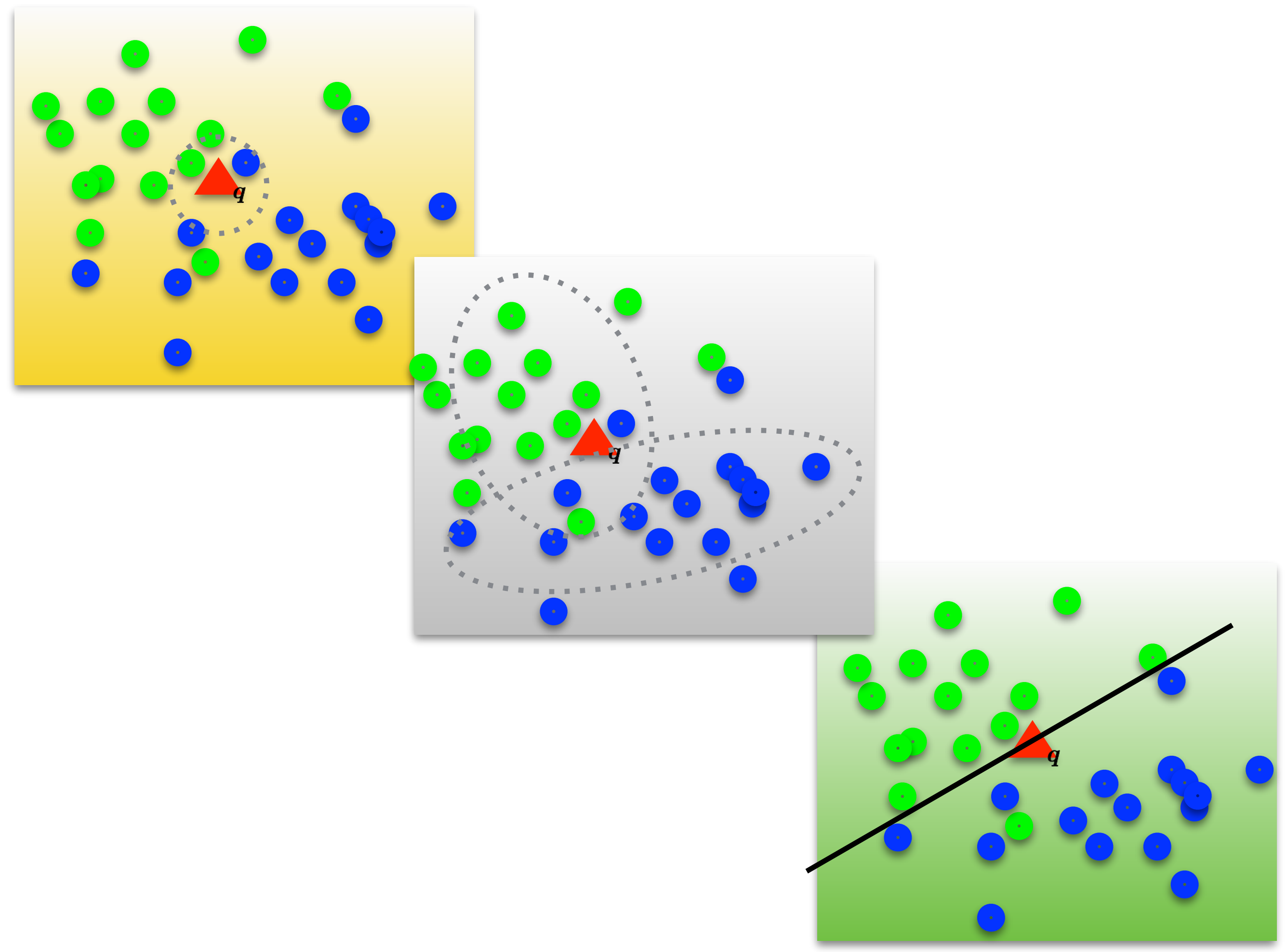
Image Classification

Classification Algorithms

— Bayes' Classifier

— Nearest Neighbor Classifier

— SVM Classifier

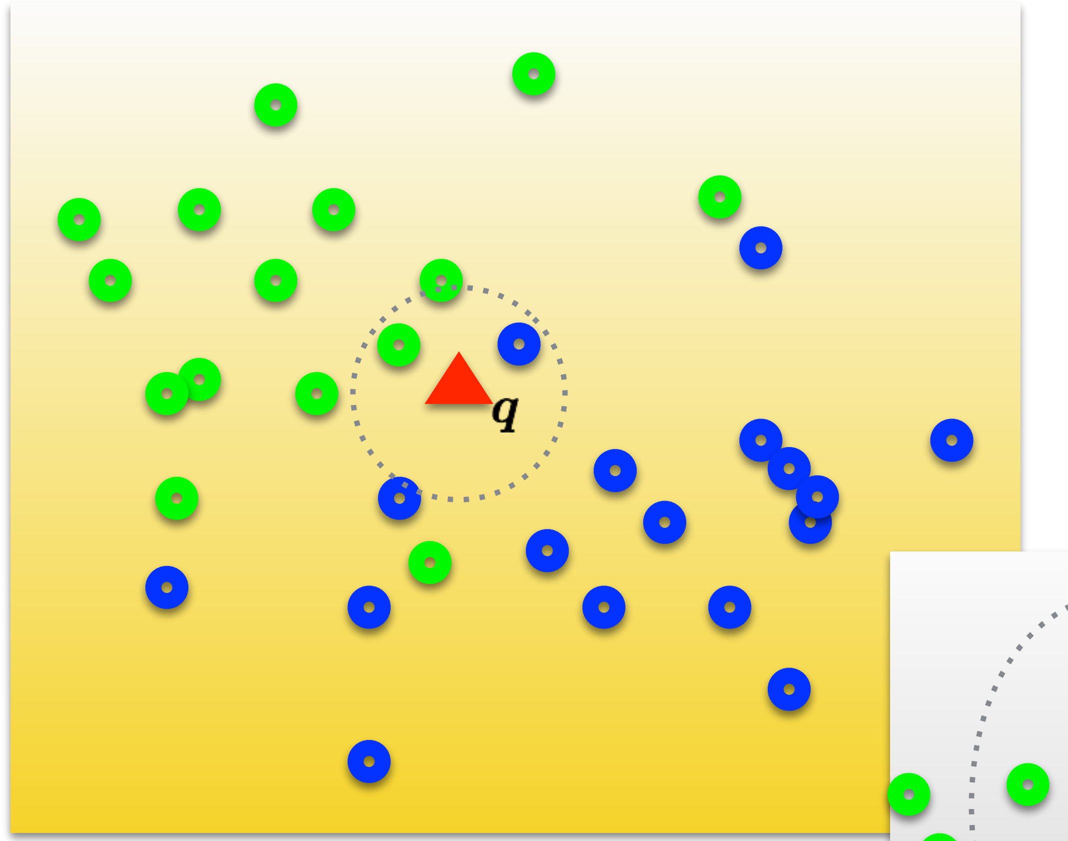


Representation of Images

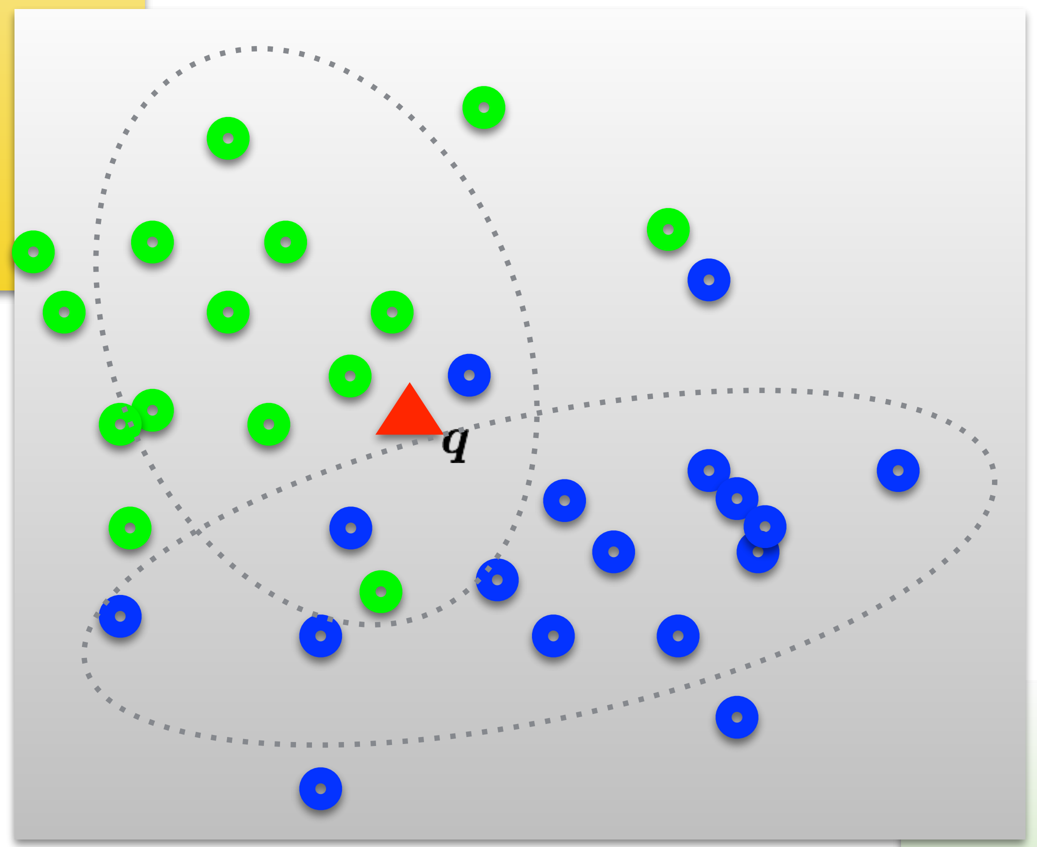
— Image pixels directly

— Bag of Words

3. Classify: Train and text classifier using BOWs

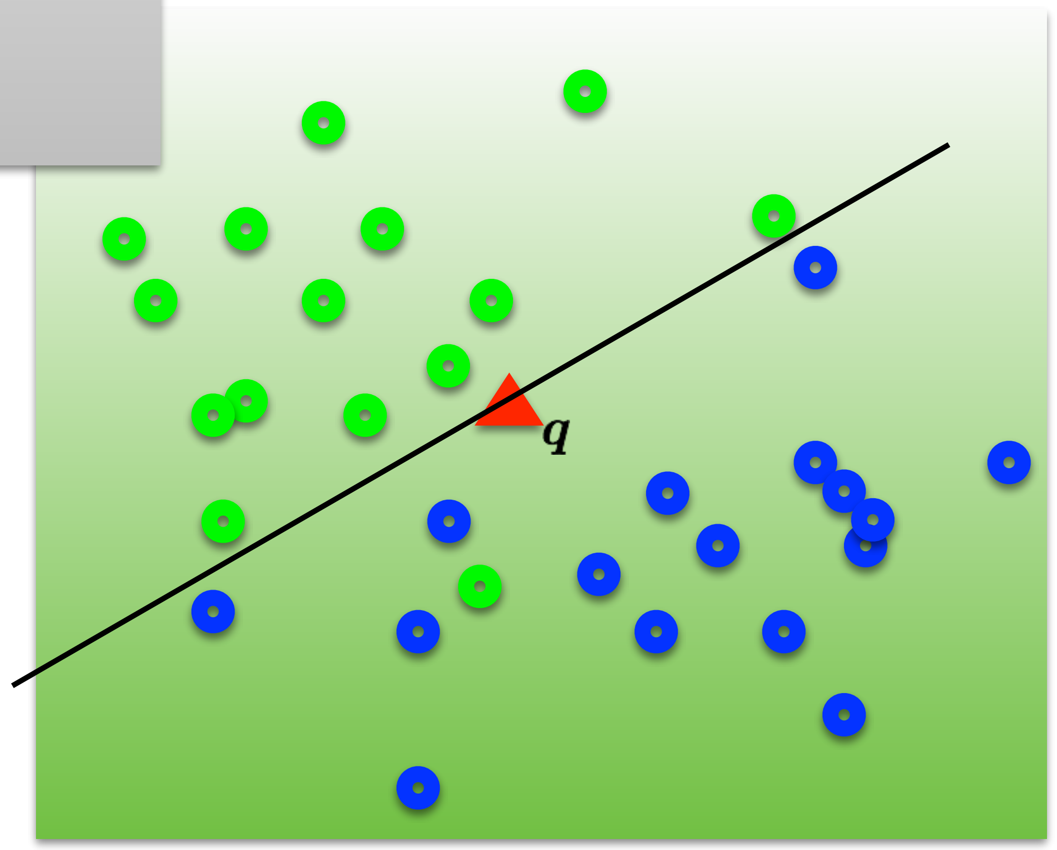


K nearest neighbors



Naïve Bayes

Support Vector Machine



Slide Credit: Ioannis (Yannis) Gkioulekas (CMU)

Bag-of-Words Representation

Algorithm:

Initialize an empty K-bin histogram, where K is the number of codewords
Extract local descriptors (e.g. SIFT) from the image
For each local descriptor **x**

  Map (Quantize) **x** to its closest codeword → **c**(**x**)

  Increment the histogram bin for **c**(**x**)

Return histogram

We can then classify the histogram using a trained classifier, e.g. a support vector machine or k-Nearest Neighbor classifier

Spatial Pyramid

The bag of words representation does not preserve any spatial information

The **spatial pyramid** is one way to incorporate spatial information into the image descriptor.

A spatial pyramid partitions the image and counts codewords within each grid box; this is performed at multiple levels

Spatial Pyramid

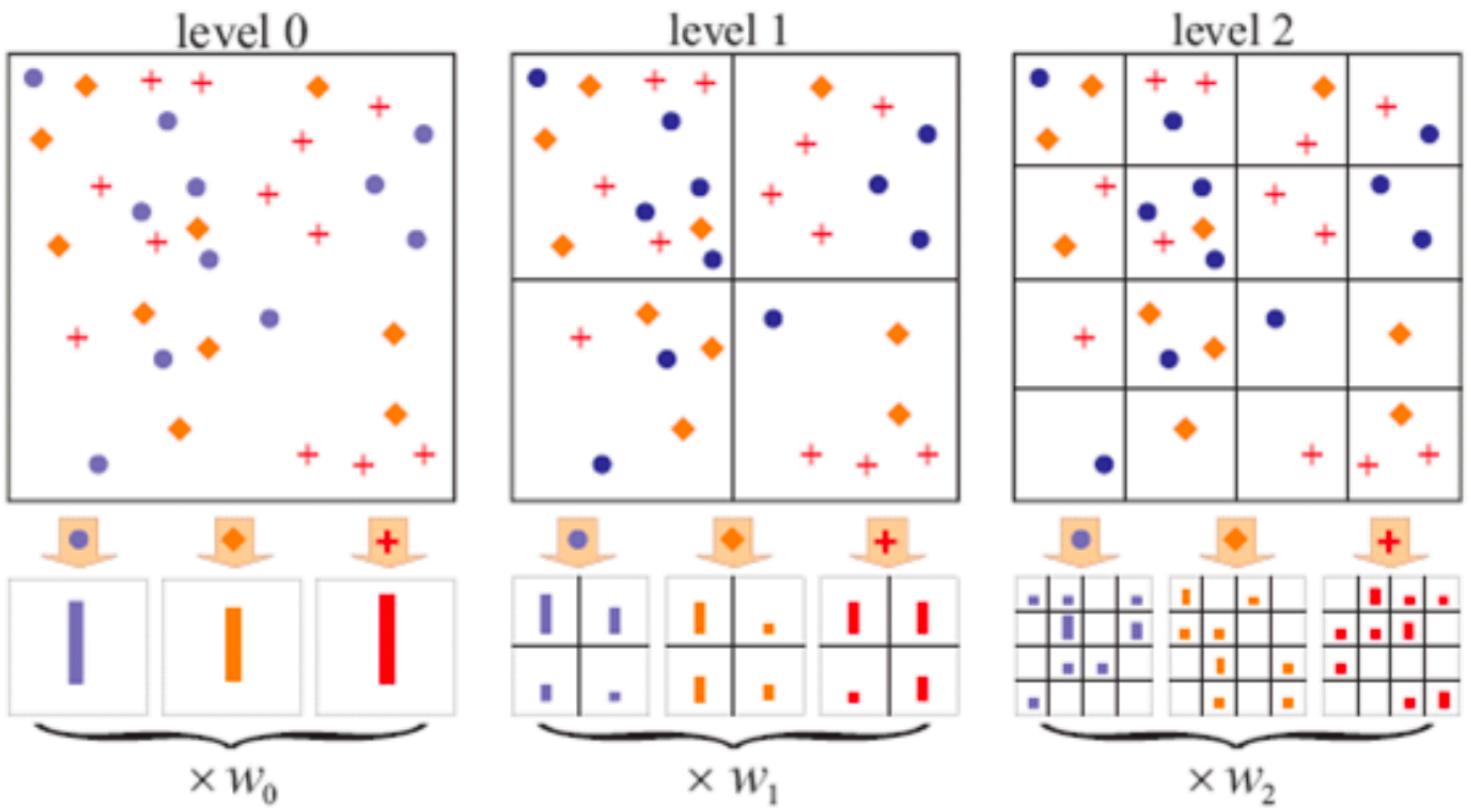


Fig. 16.8 in Forsyth & Ponce (2nd ed.).
Original credit: Lazebnik et al., 2006

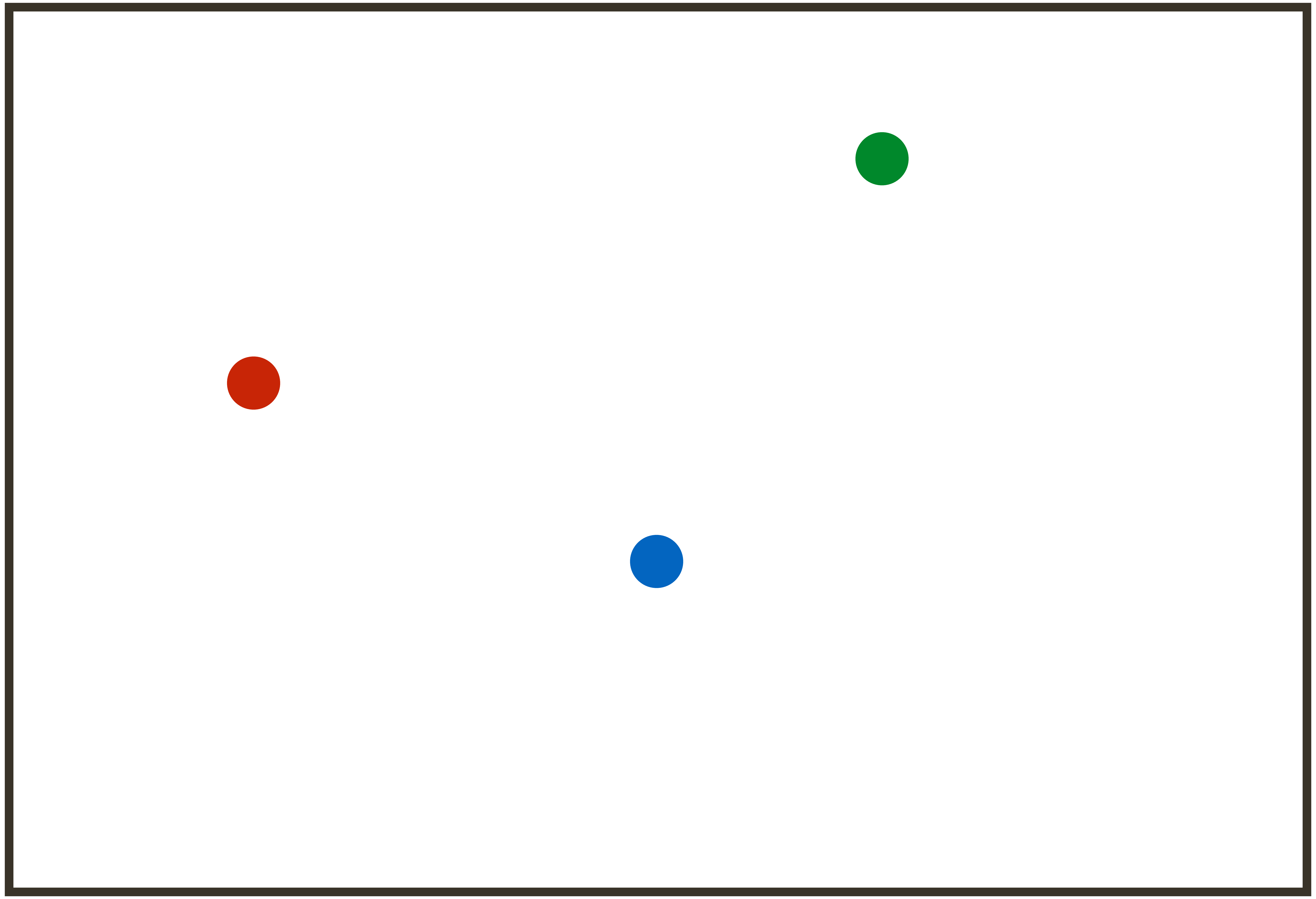
VLAD (Vector of Locally Aggregated Descriptors)

There are more advanced ways to 'count' visual words than incrementing its histogram bin

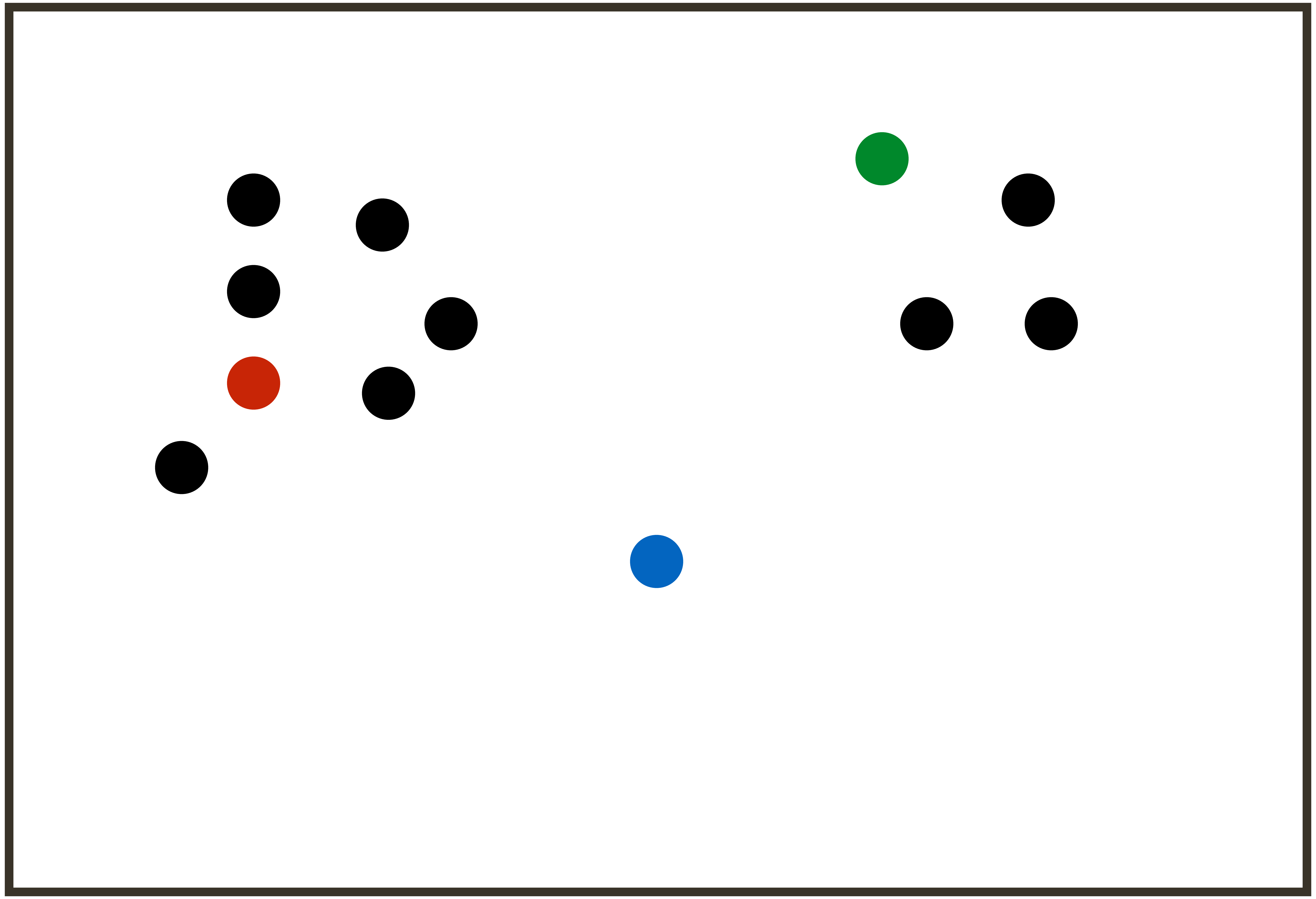
For example, it might be useful to describe how local descriptors are quantized to their visual words

In the VLAD representation, instead of incrementing the histogram bin by one, we increment it by the **residual** vector $\boldsymbol{x} - \boldsymbol{c}(\boldsymbol{x})$

Example: VLAD

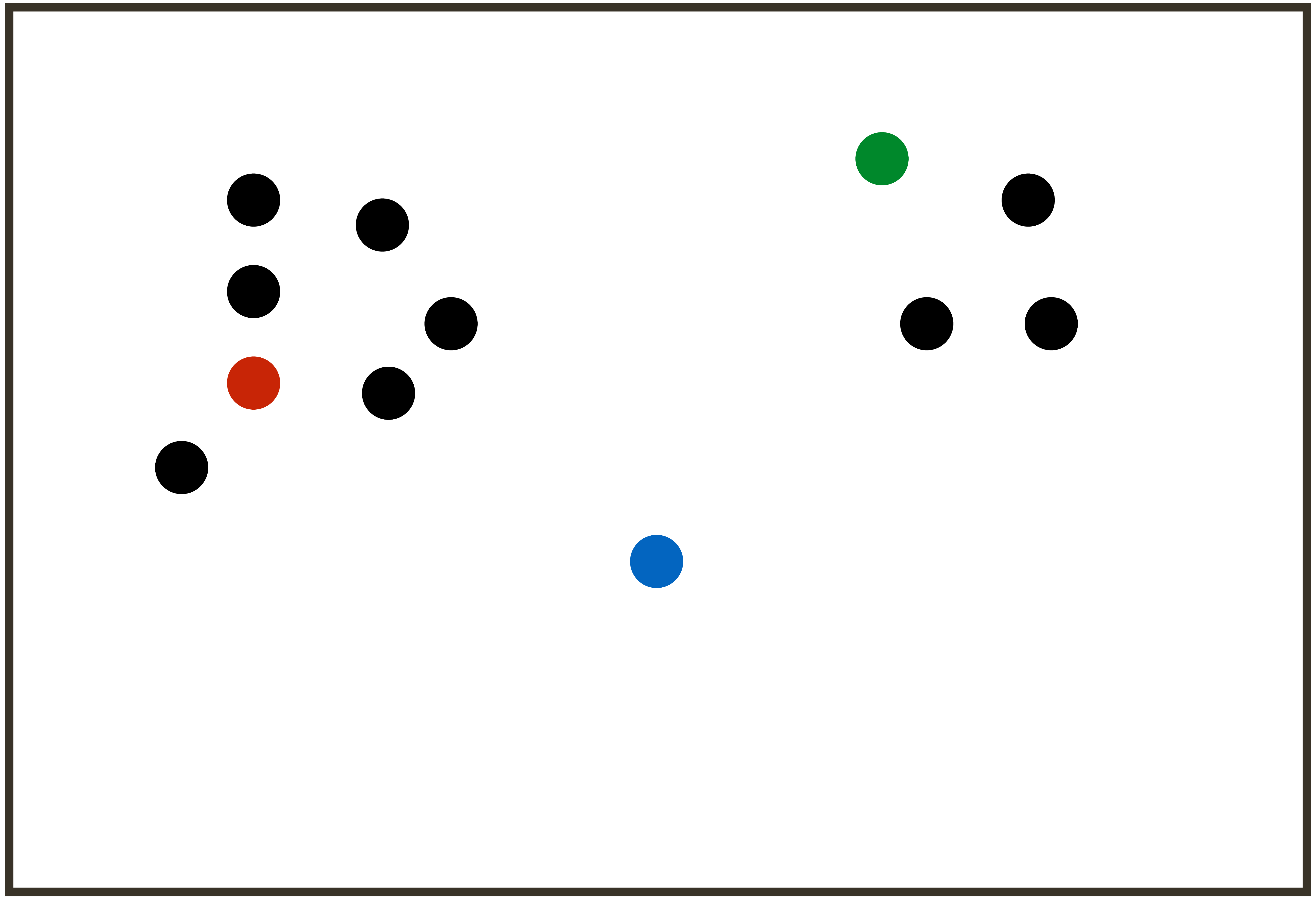


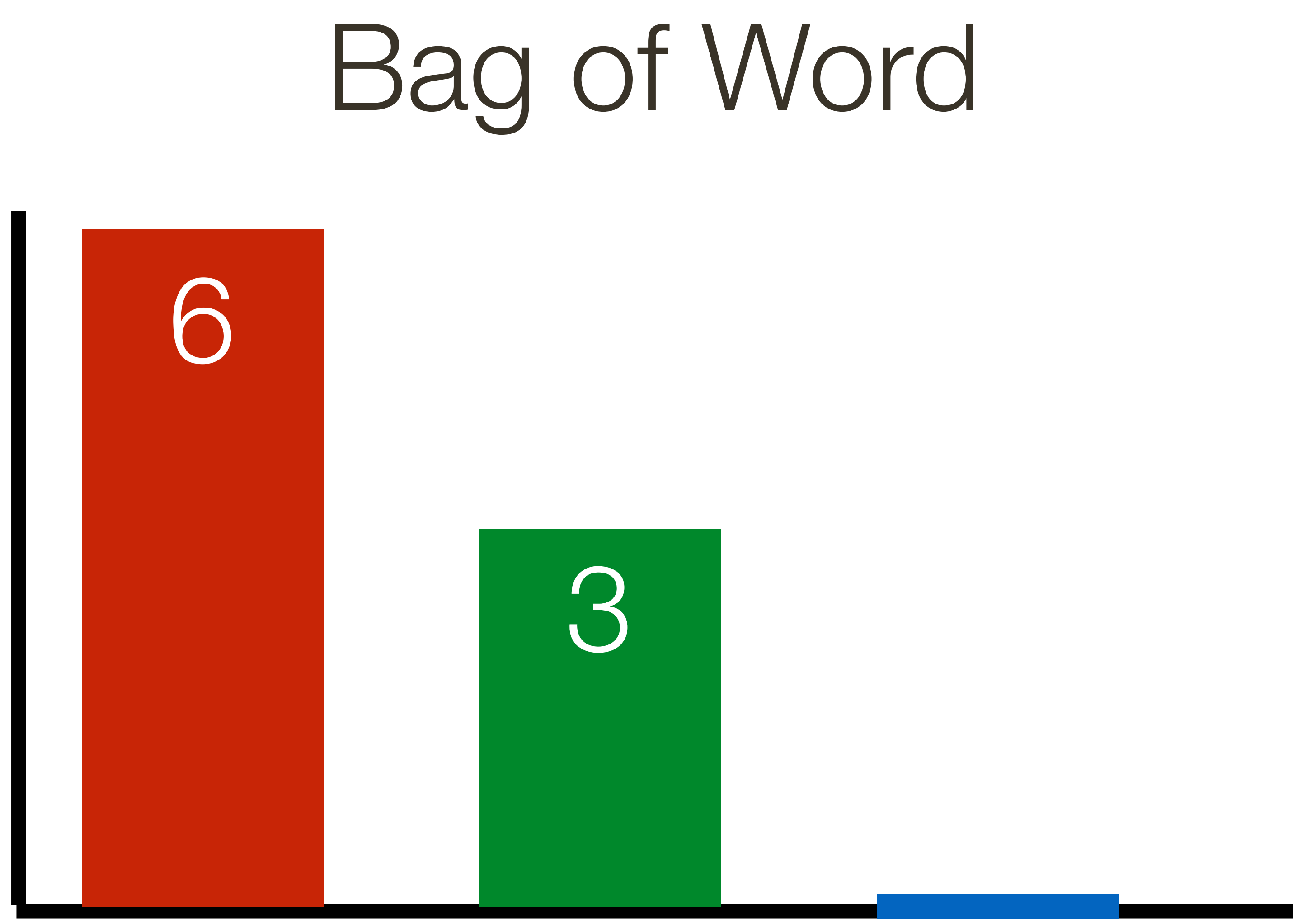
Example: VLAD



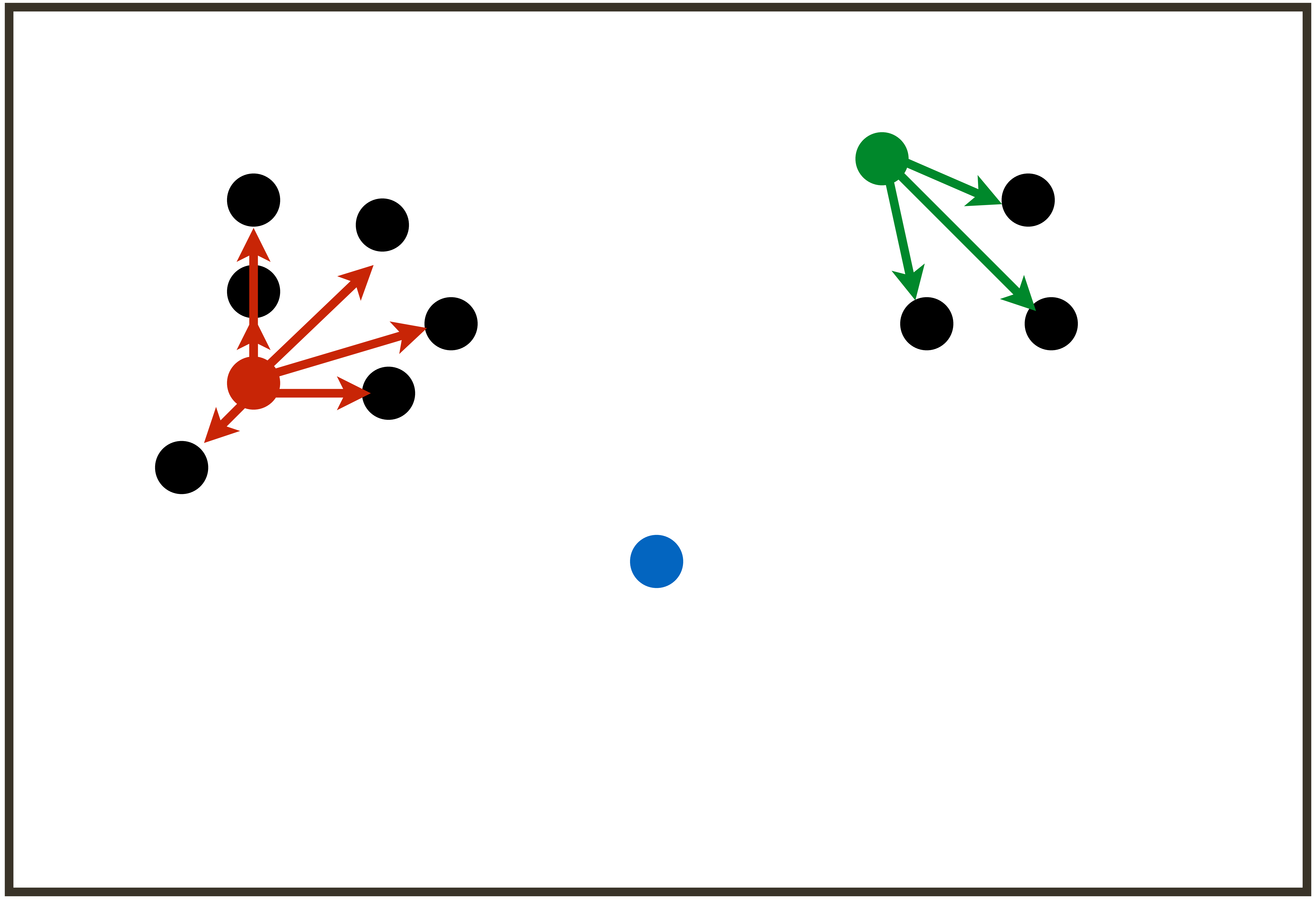
Bag of Word

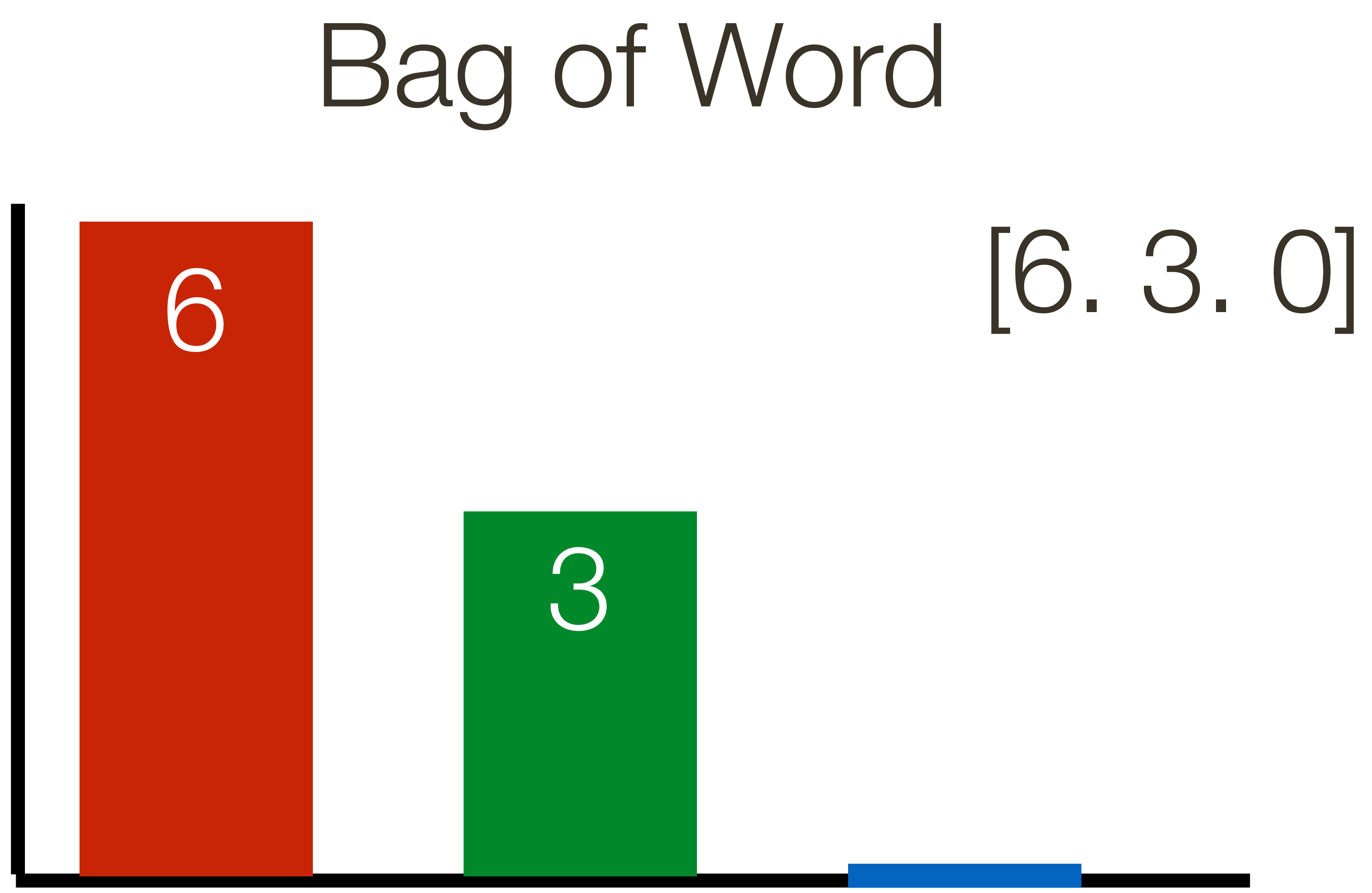
Example: VLAD

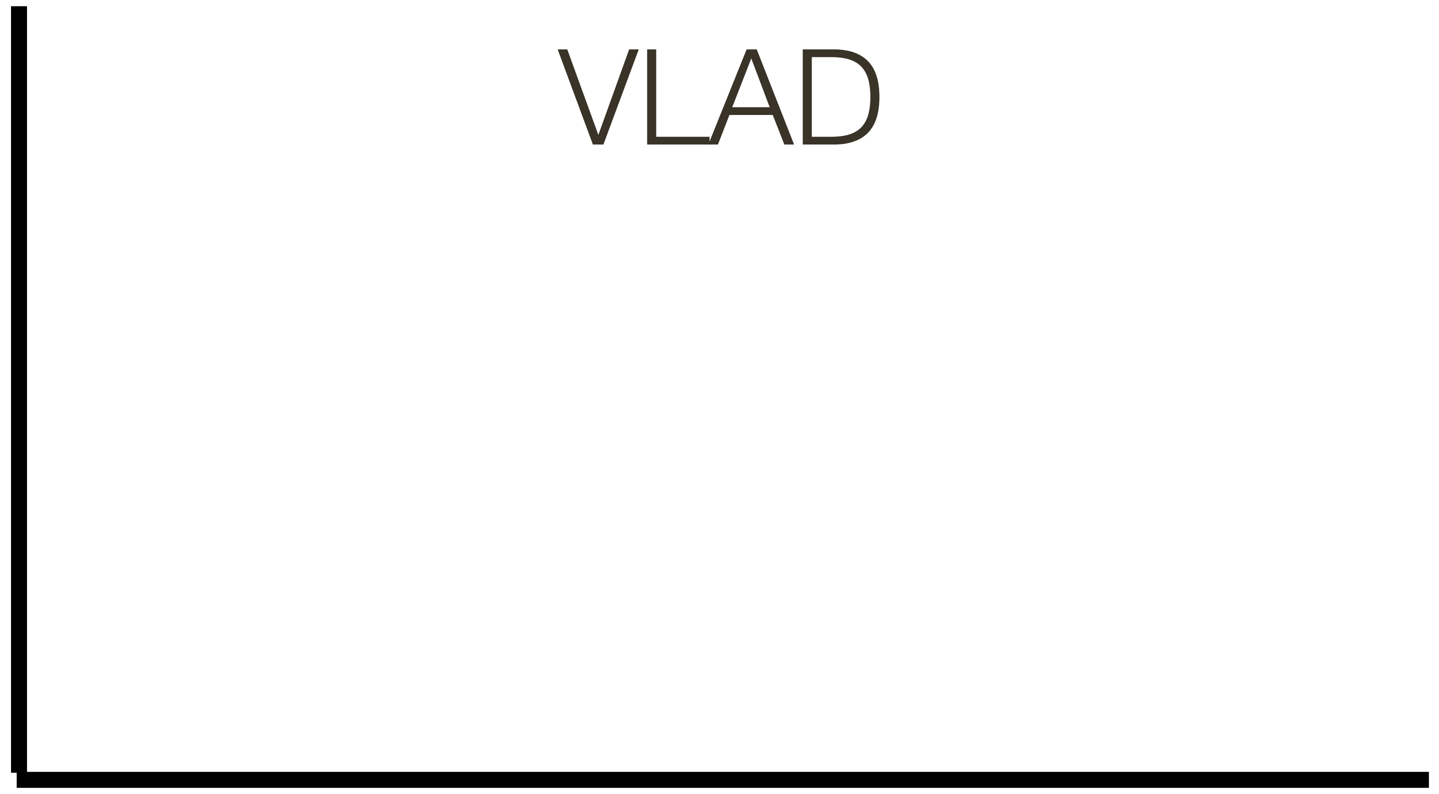




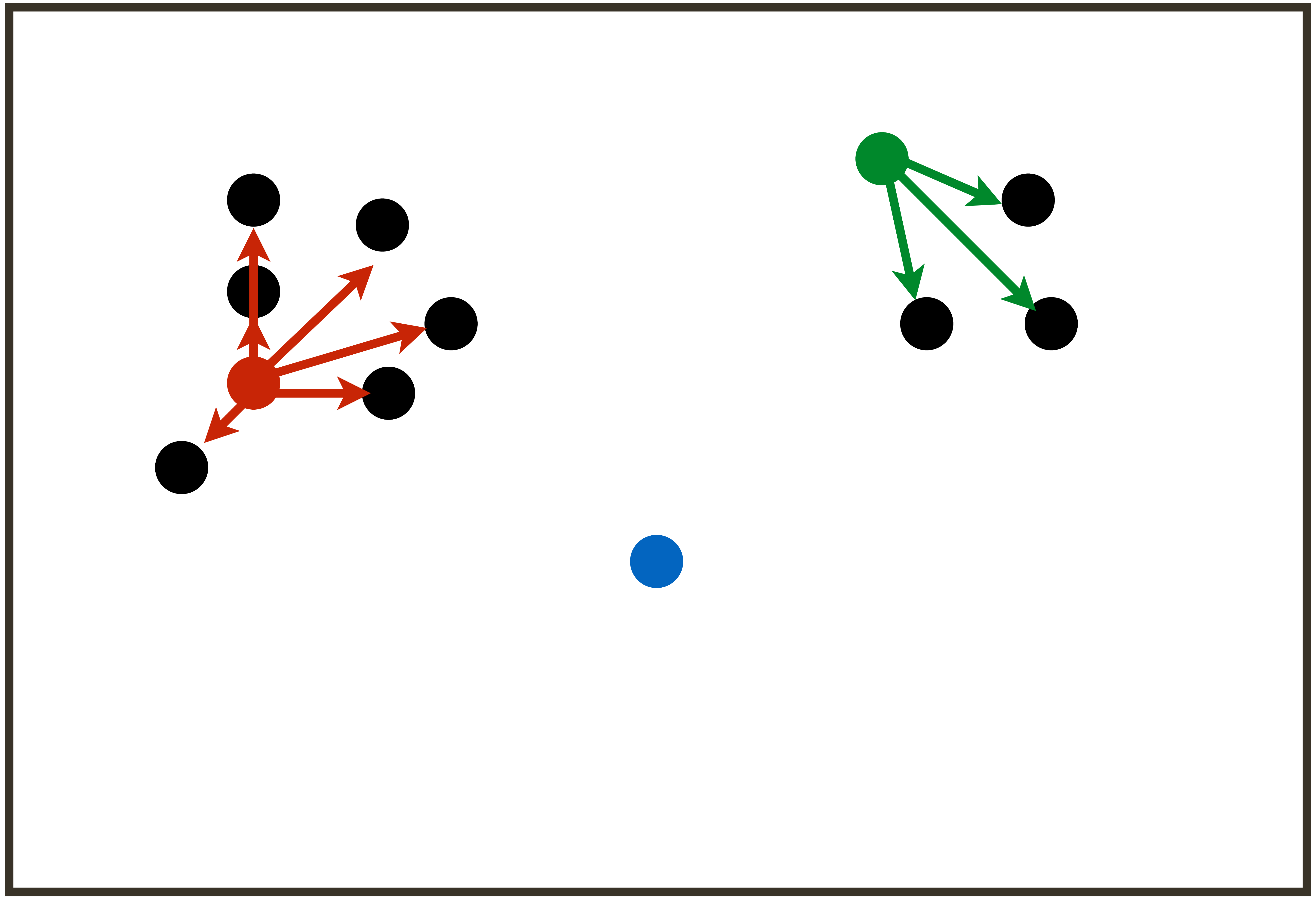
Example: VLAD

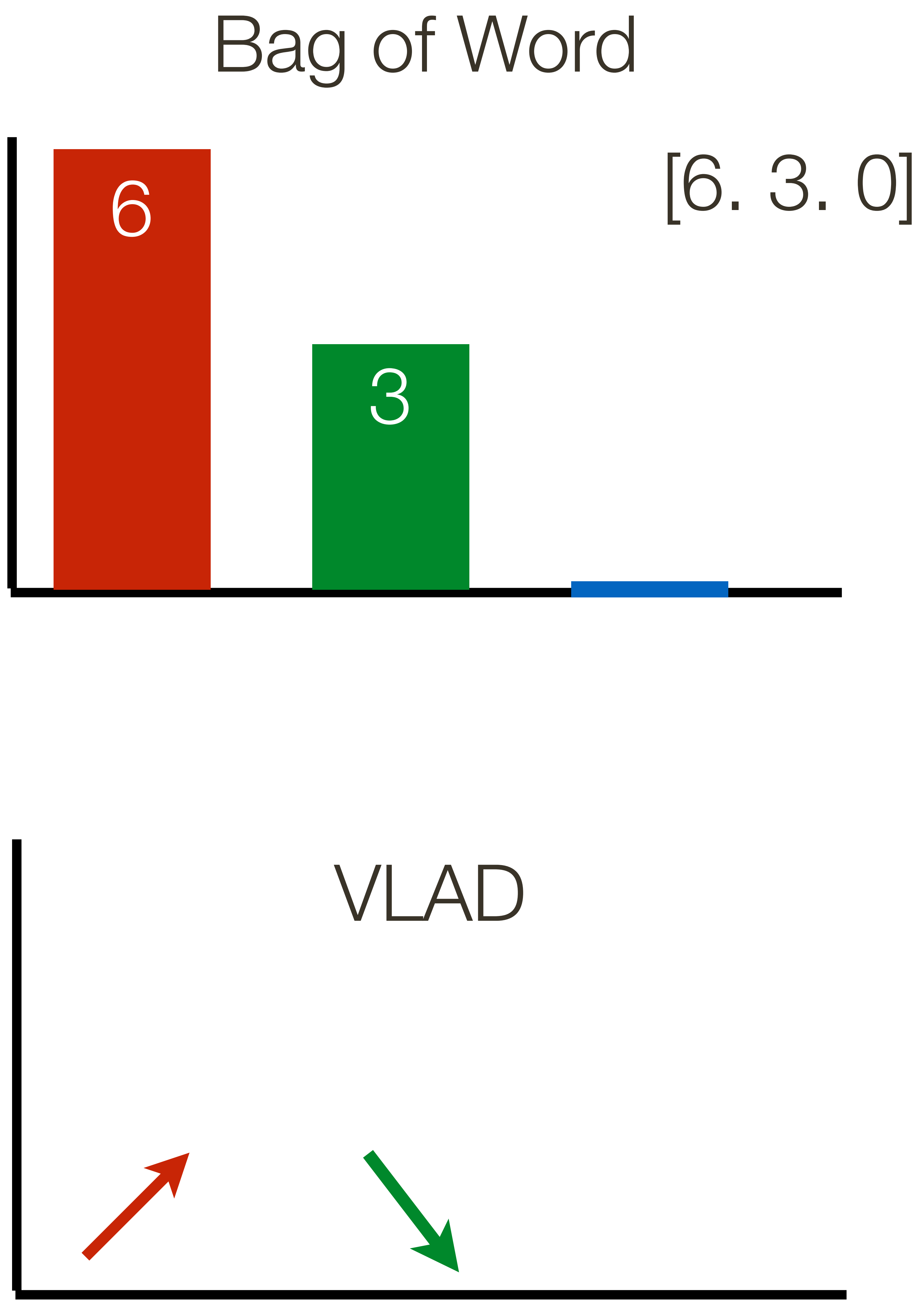






Example: VLAD





VLAD (Vector of Locally Aggregated Descriptors)

The dimensionality of a **VLAD** descriptor is Kd

— K : number of codewords

— d : dimensionality of the local descriptor

VLAD characterizes the distribution of local descriptors with respect to the codewords

Summary

Factors that make image classification hard
— intra-class variation, viewpoint, illumination, clutter, occlusion...

A codebook of **visual words** contains representative local patch descriptors
— can be constructed by clustering local descriptors (e.g. SIFT) in training images

The **bag of words** model accumulates a histogram of occurrences of each visual word

The **spatial pyramid** partitions the image and counts visual words within each grid box; this is repeated at multiple levels