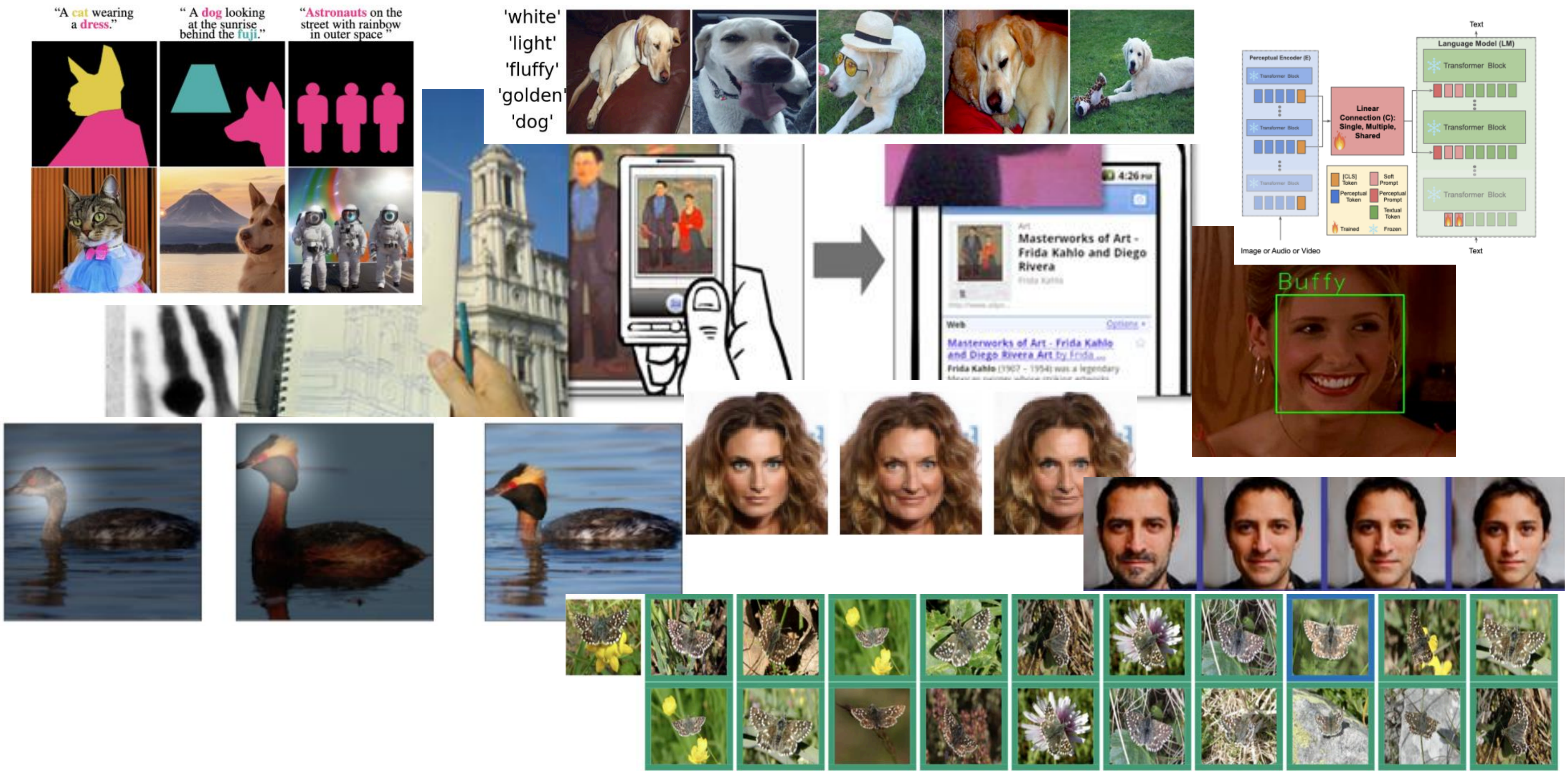


RDFIA: deep learning for Vision

https://cord.isir.upmc.fr/teaching-rdfia/

Matthieu Cord

Sorbonne University

Course Outline

https://cord.isir.upmc.fr/teaching-rdfia/

1. Intro to Computer Vision and Machine Learning
2. Intro to Neural Networks + Machine Learning theory
3. Neural Nets for Image Classification
4. Large ConvNets
5. Vision Transformers
6. Segmentation, Transfer learning and domain adaptation
7. Vision-Language models
8. Explaining VLMS
9. Self Supervised Learning in Vision
10. Generative models with GANs
11. Control Jan 07, 2026
12. Diffusion models
13. Bayesian deep learning

14 Uncertainty, Robustness

Evaluations: Control (30%) + Practicals (3 reports, total=70%)
can be modified by 10% between the 2 evaluations

COMPUTER VISION:

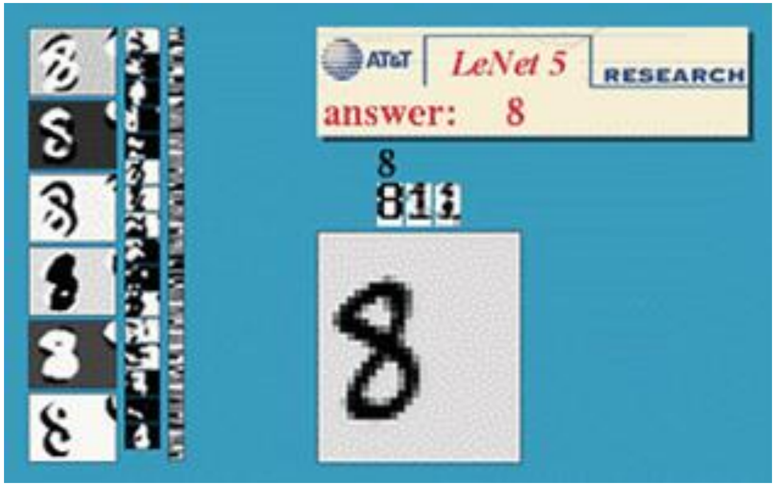
(Processing, analyzing and) **understanding visual data**

=>**WHERE ARE WE NOW?**

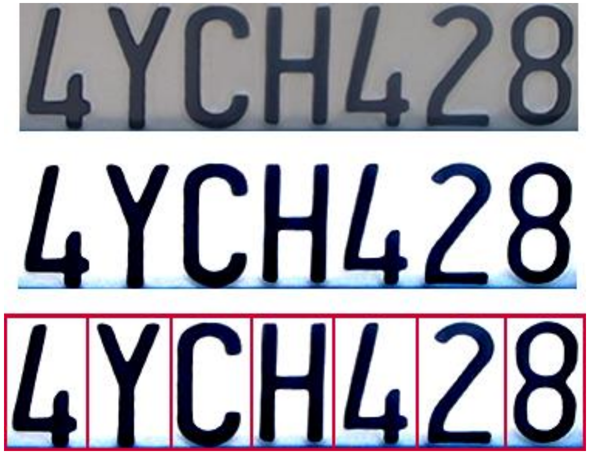
Source (many slides): Cornell CV course

Deployed: Optical character recognition (OCR)

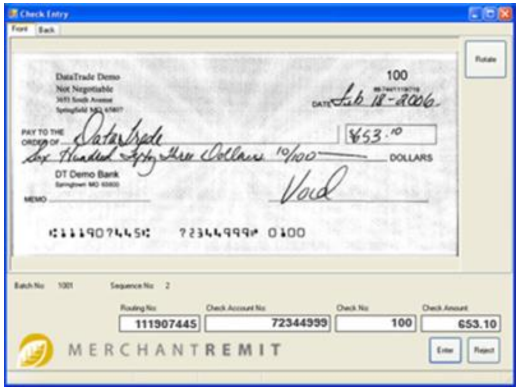
- If you have a scanner, it probably came with OCR software



Digit recognition, AT&T labs
http://www.research.att.com/~yann/



License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition



DataTrade Demo
Not Negotiable
3651 South Avenue
Springfield MO 65807

100

DATE Feb 18-2006

PAY TO THE ORDER OF DataTrade $653.10

Six Hundred Fifty Three Dollars 10/100 DOLLARS

DT Demo Bank
Springtown MO 65800

MEMO

Void

⑆111907445⑆ 72344999⑈ 0100

Batch No 1001 Sequence No 2

Routing No	Check Account No	Check No	Check Amount
111907445	72344999	100	653.10

Automatic check processing

Source: S. Seitz

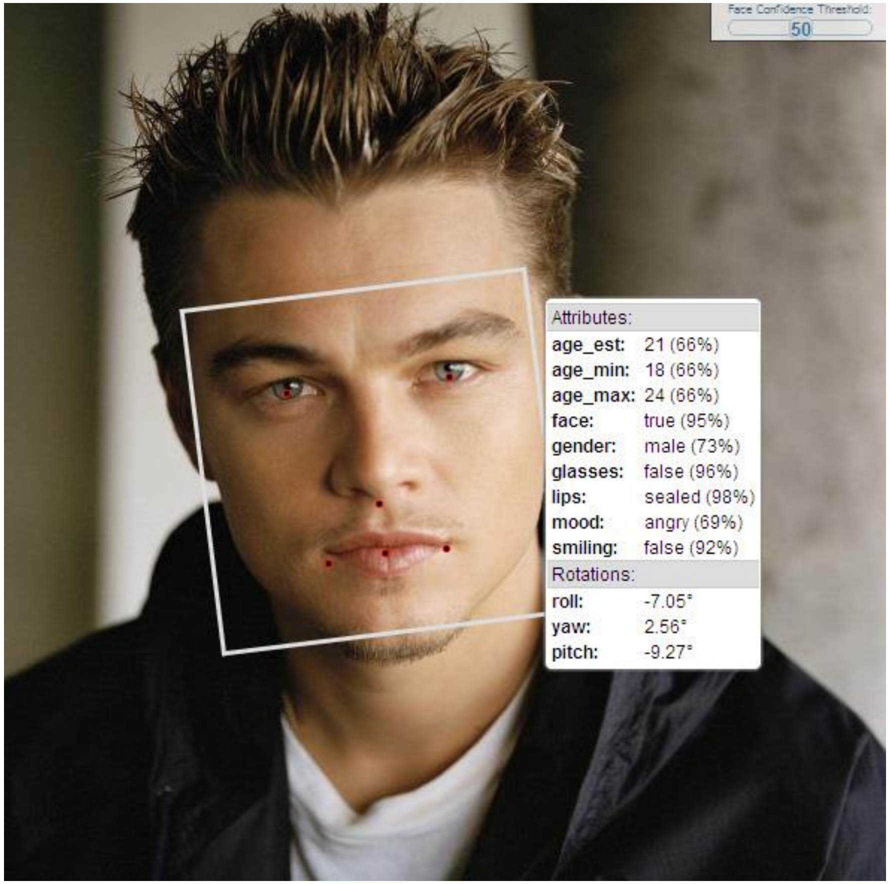
Deployed: Face detection

- Cameras now detect faces
  - – Canon, Sony, Fuji, …

Source: S. Seitz

Deployed&Significant progress: Face Recognition

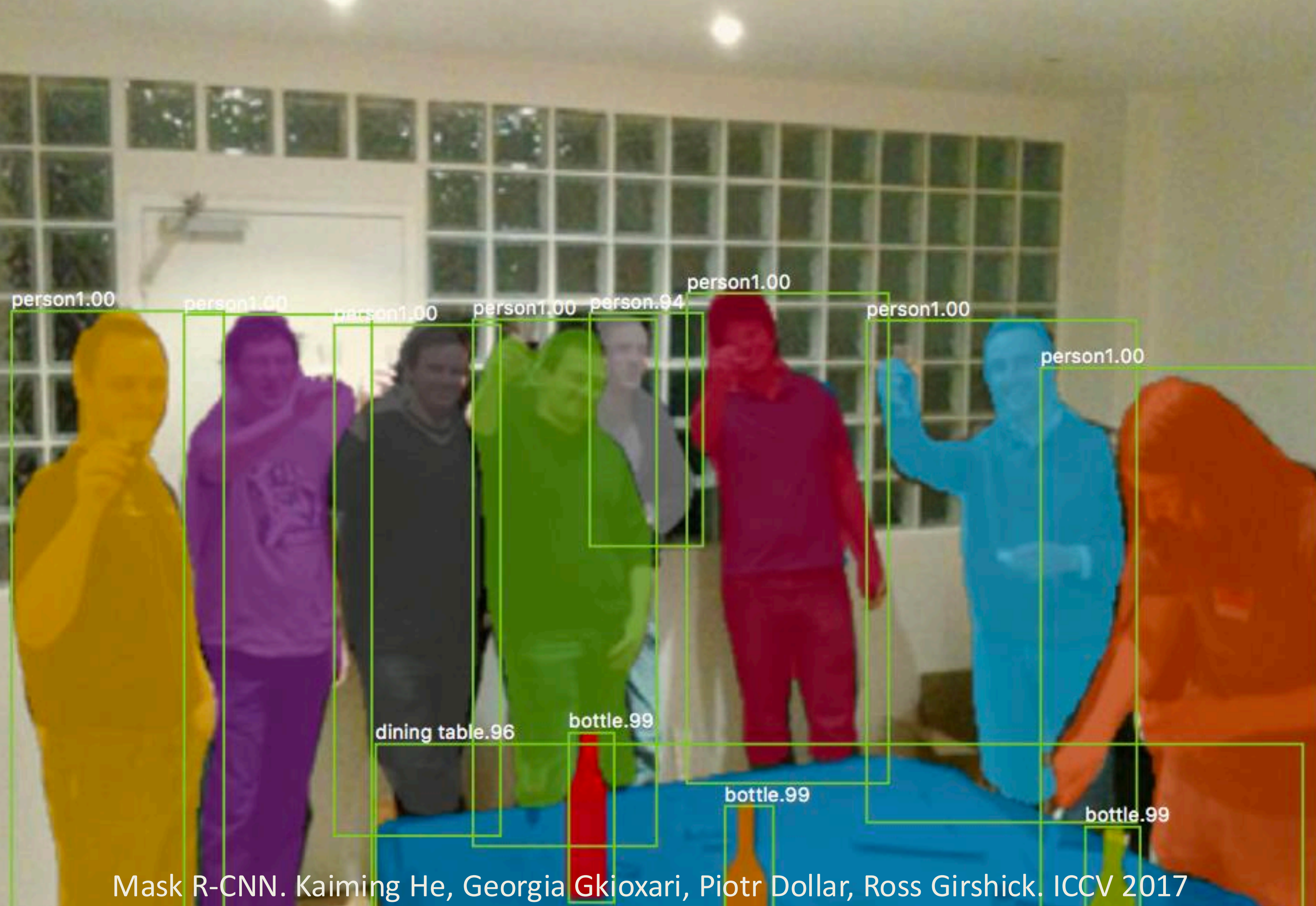
FACE++



Face Confidence Threshold: 50

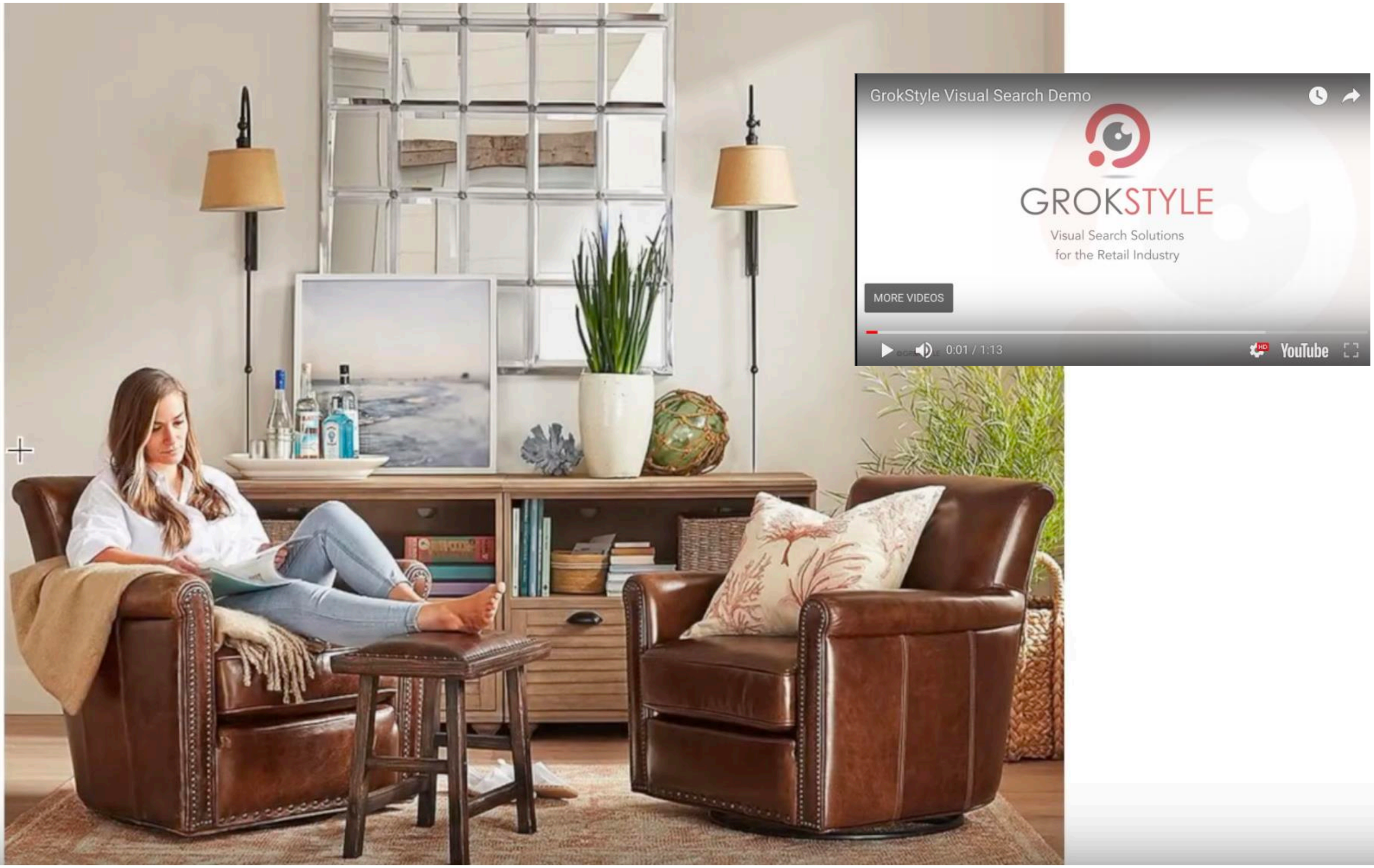
Attributes:	
age_est:	21 (66%)
age_min:	18 (66%)
age_max:	24 (66%)
face:	true (95%)
gender:	male (73%)
glasses:	false (96%)
lips:	sealed (98%)
mood:	angry (69%)
smiling:	false (92%)
Rotations:	
roll:	-7.05°
yaw:	2.56°
pitch:	-9.27°

Significant progress: Recognizing objects



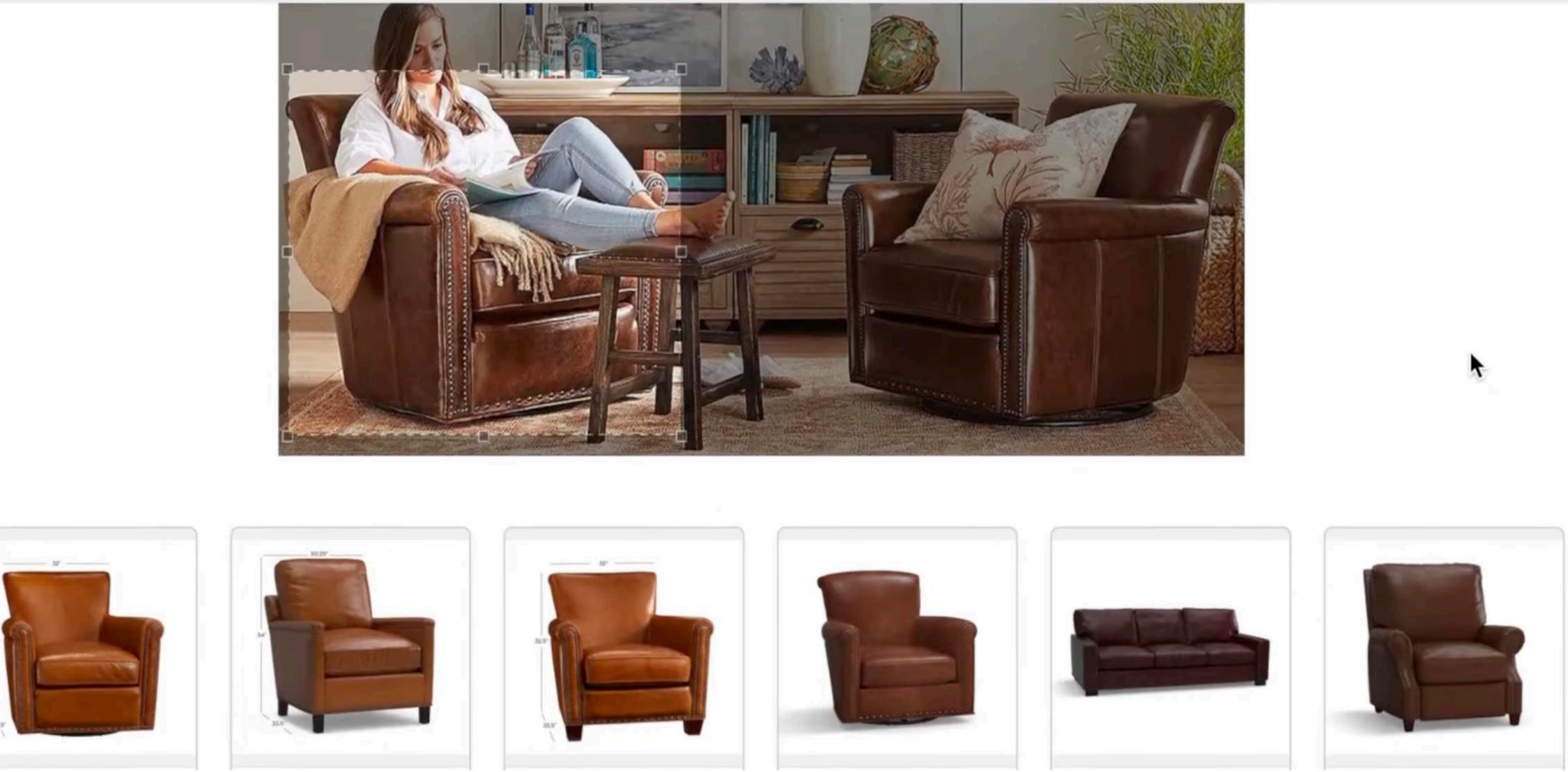
Mask R-CNN. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. ICCV 2017

Ex: Recognition-based product search

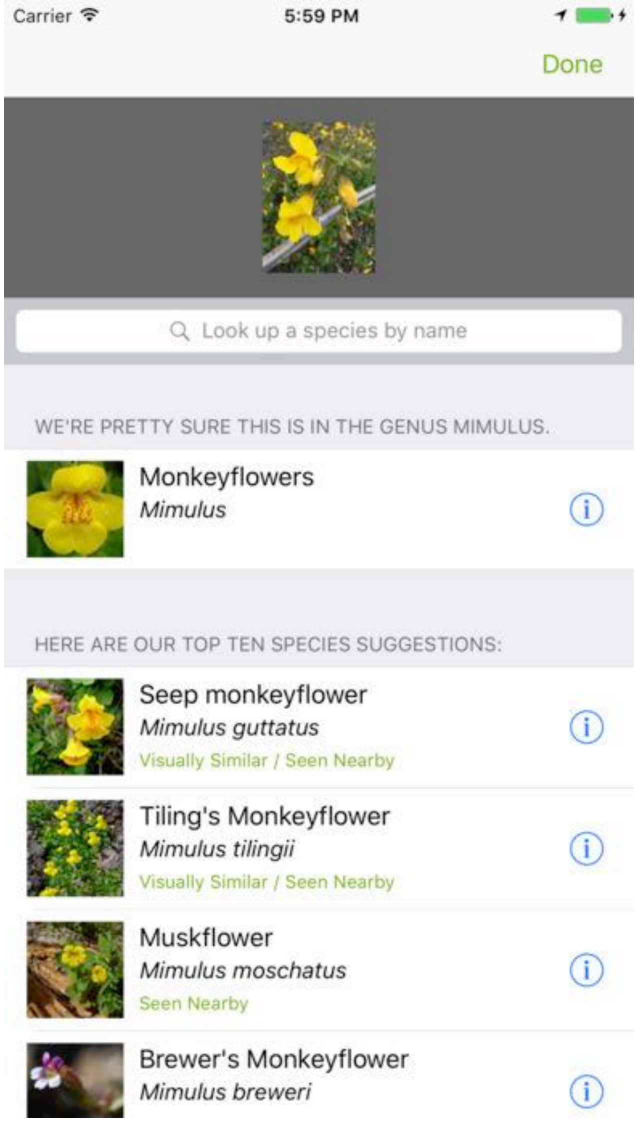


Recognition-based product search

Recognition-based product search



Significant progress: Species recognition



Carrier

5:59 PM

Done

Look up a species by name

WE'RE PRETTY SURE THIS IS IN THE GENUS MIMULUS.

Monkeyflowers
Mimulus

HERE ARE OUR TOP TEN SPECIES SUGGESTIONS:

Seep monkeyflower
Mimulus guttatus
Visually Similar / Seen Nearby

Tiling's Monkeyflower
Mimulus tilingii
Visually Similar / Seen Nearby

Muskflower
Mimulus moschatus
Seen Nearby

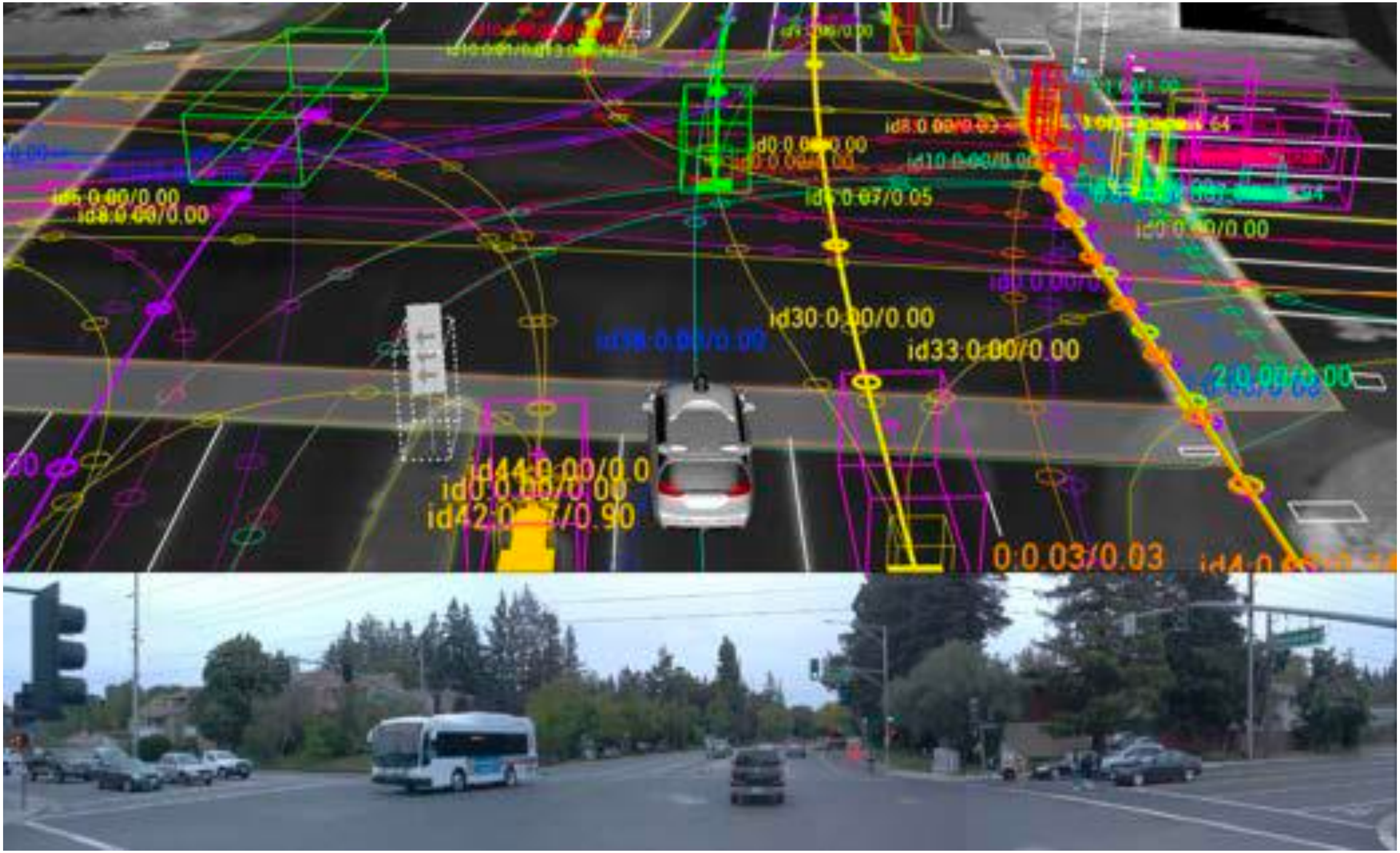
Brewer's Monkeyflower
Mimulus breweri

iNaturalist dataset

Challenges:

- fine-grained recognition
- Detecting rare concepts

Challenges: Fully autonomous driving



Challenges: Medical Imaging, Health

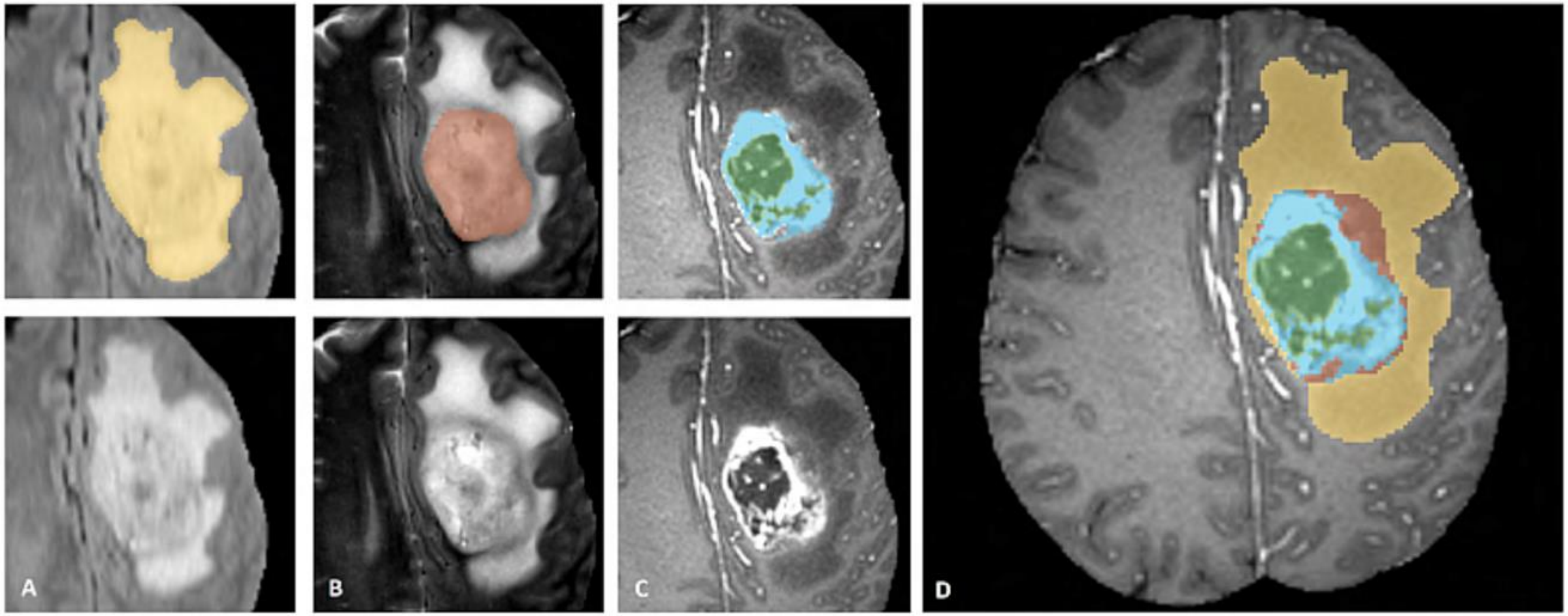
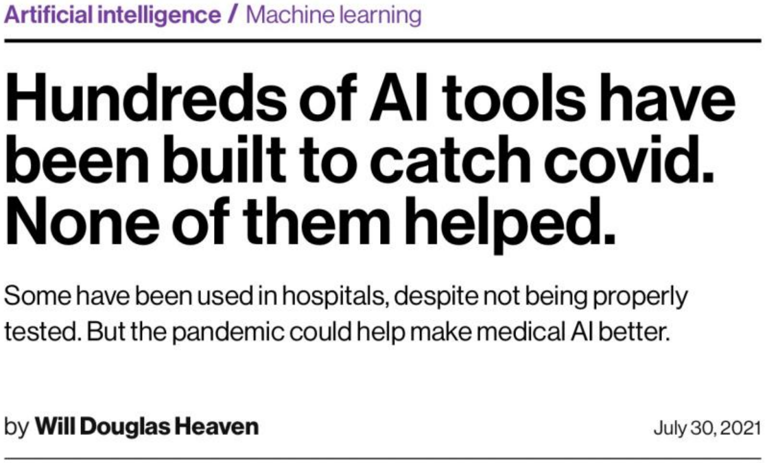


Fig.1: Glioma sub-regions. Shown are image patches with the tumor sub-regions that are annotated in the different modalities (top left) and the final labels for the whole dataset (right). The image patches show from left to right: the whole tumor (yellow) visible in T2-FLAIR (Fig.A), the tumor core (red) visible in T2 (Fig.B), the enhancing tumor structures (light blue) visible in T1Gd, surrounding the cystic/necrotic components of the core (green) (Fig. C). The segmentations are combined to generate the final labels of the tumor sub-regions (Fig.D): edema (yellow), non-enhancing solid core (red), necrotic/cystic core (green), enhancing core (blue). (Figure taken from the BraTS IEEE TMI paper.)

Challenges: Medical Imaging, Health

Building system to detect Covid in chest x rays
What should a metric measure?
Accuracy = P(pred_label == true_label)
Accuracy of candidate system = 95%
Is this good? Did it actually help / work?



Artificial intelligence / Machine learning

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

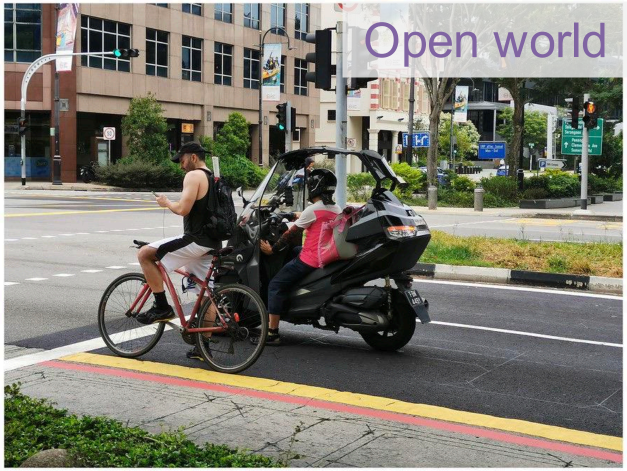
by **Will Douglas Heaven** July 30, 2021

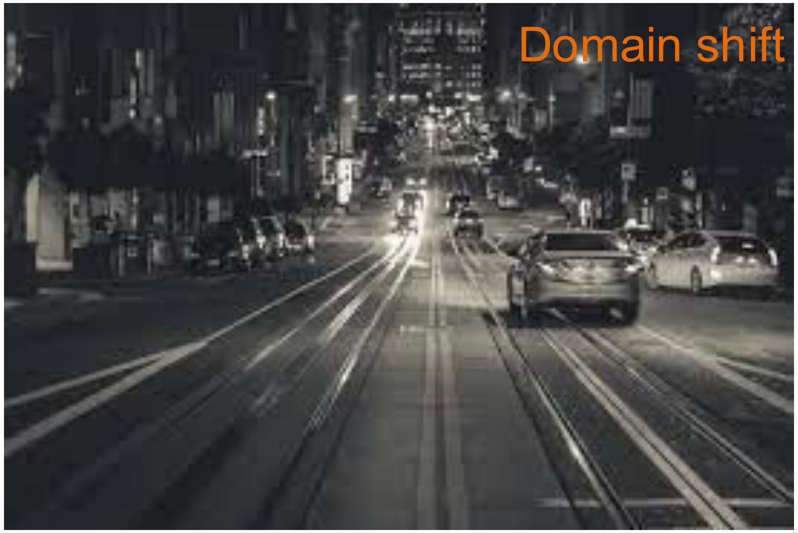
Why?

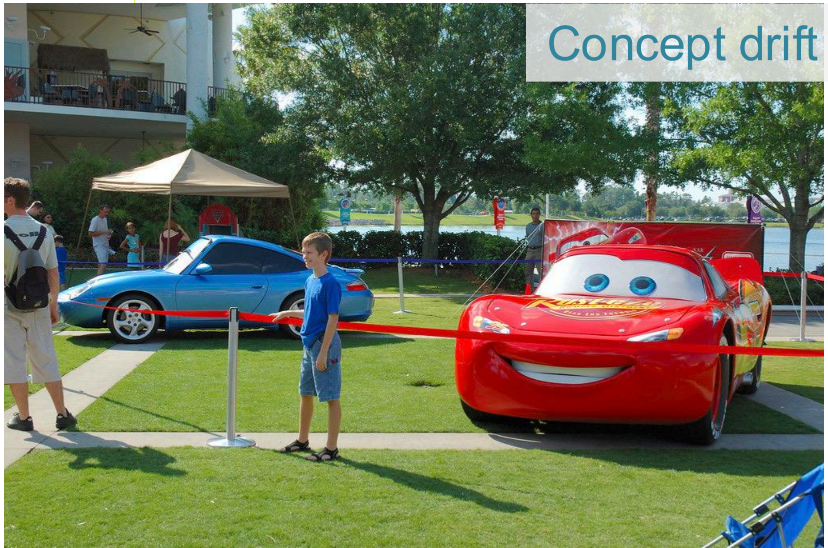
Typical issues that plague deployment

- Images seen during deployment are very different: domain shift
- Meaning of classes etc. change: concept drift
- Unforeseen circumstances, e.g., new classes: open world

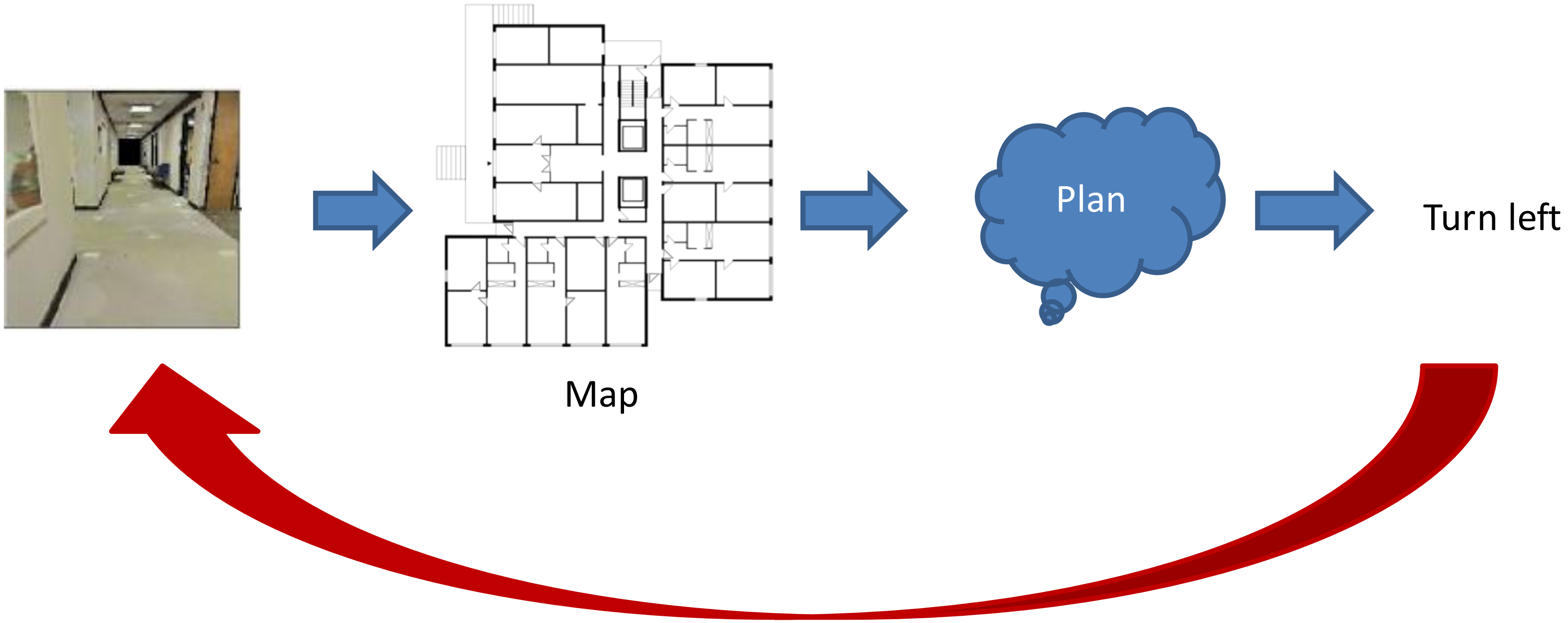
Original data


Open world


Domain shift


Concept drift

Challenges: Integrating Vision and Action, Robotics



Map

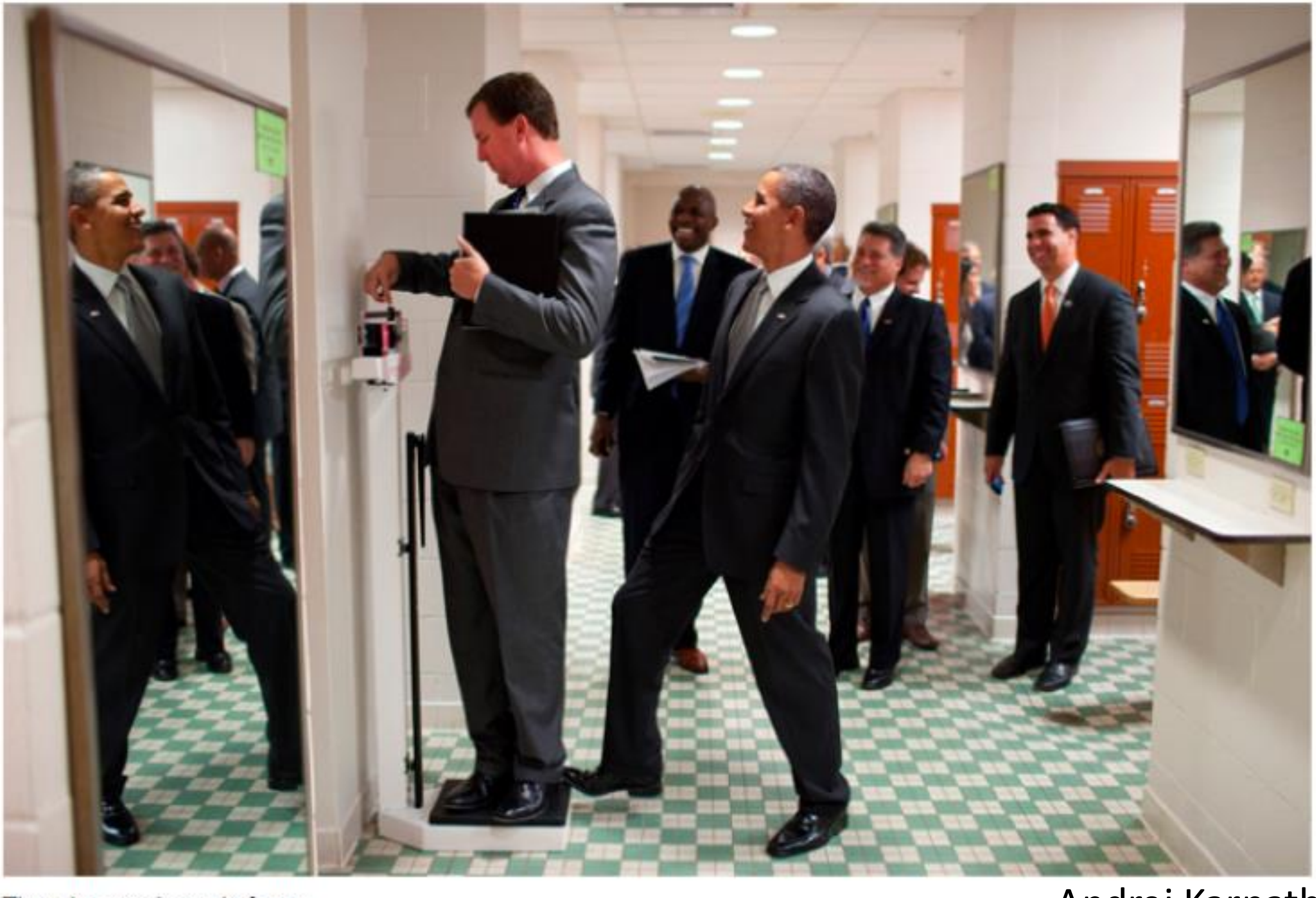
Plan

Turn left

Cognitive Mapping and Planning for Visual Navigation
Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, Jitendra Malik
CVPR 2017

Challenges: Understanding complex situations / Reasoning

Challenges: Visual Reasoning
VQA task: Why is this funny?



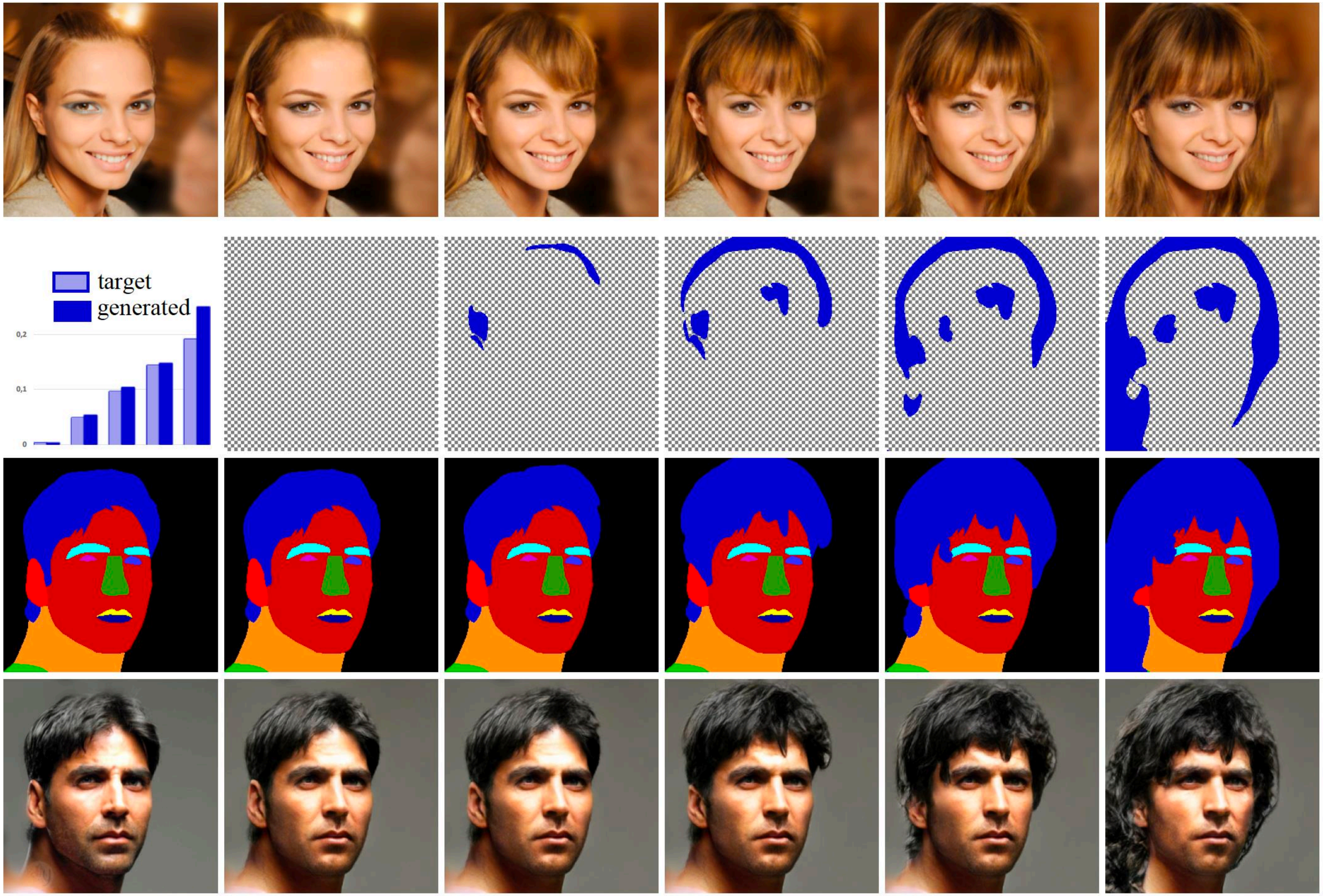
The picture above is funny.

Andrej Karpathy

VQA task: Why is this funny?
ChatGPTo answer :

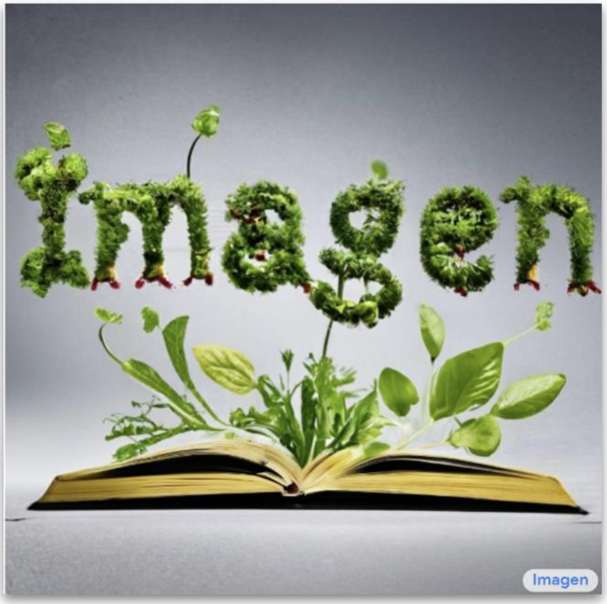
The humor in this image stems from the playful interaction between former President Barack Obama and a much taller man who appears to be getting his height measured. The scene is light-hearted and unexpected, showing a humorous side of a usually serious figure like the President. The contrast in their heights, with Obama leaning back and smiling while the taller man stands on a scale, adds to the comedic effect. The expressions and body language of the other men in the background, who are also smiling and laughing, contribute to the overall jovial atmosphere.

Challenges: Generative models for images-editing, manipulation (with GANs)

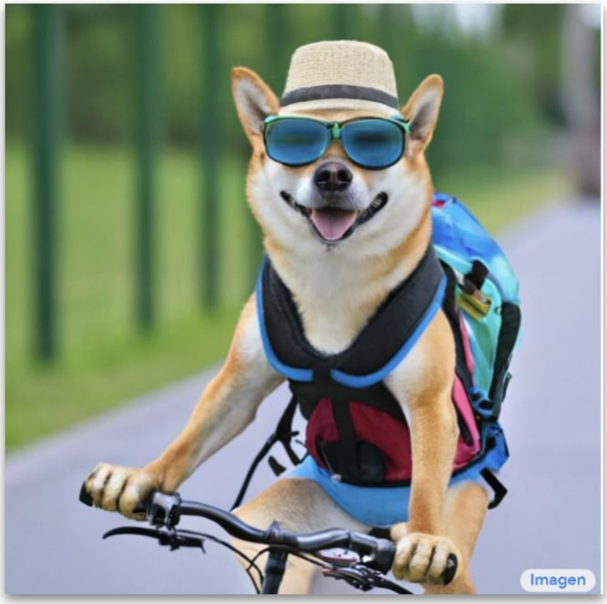


Challenges: Image Generation in 2023 (Diffusion Models) **from Text**

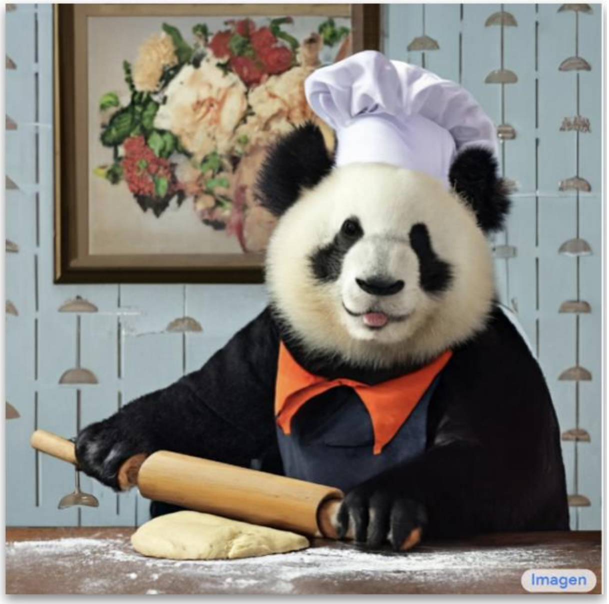
Sprouts in the shape of text 'Imagen' coming out of a fairytale book



A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.



A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.



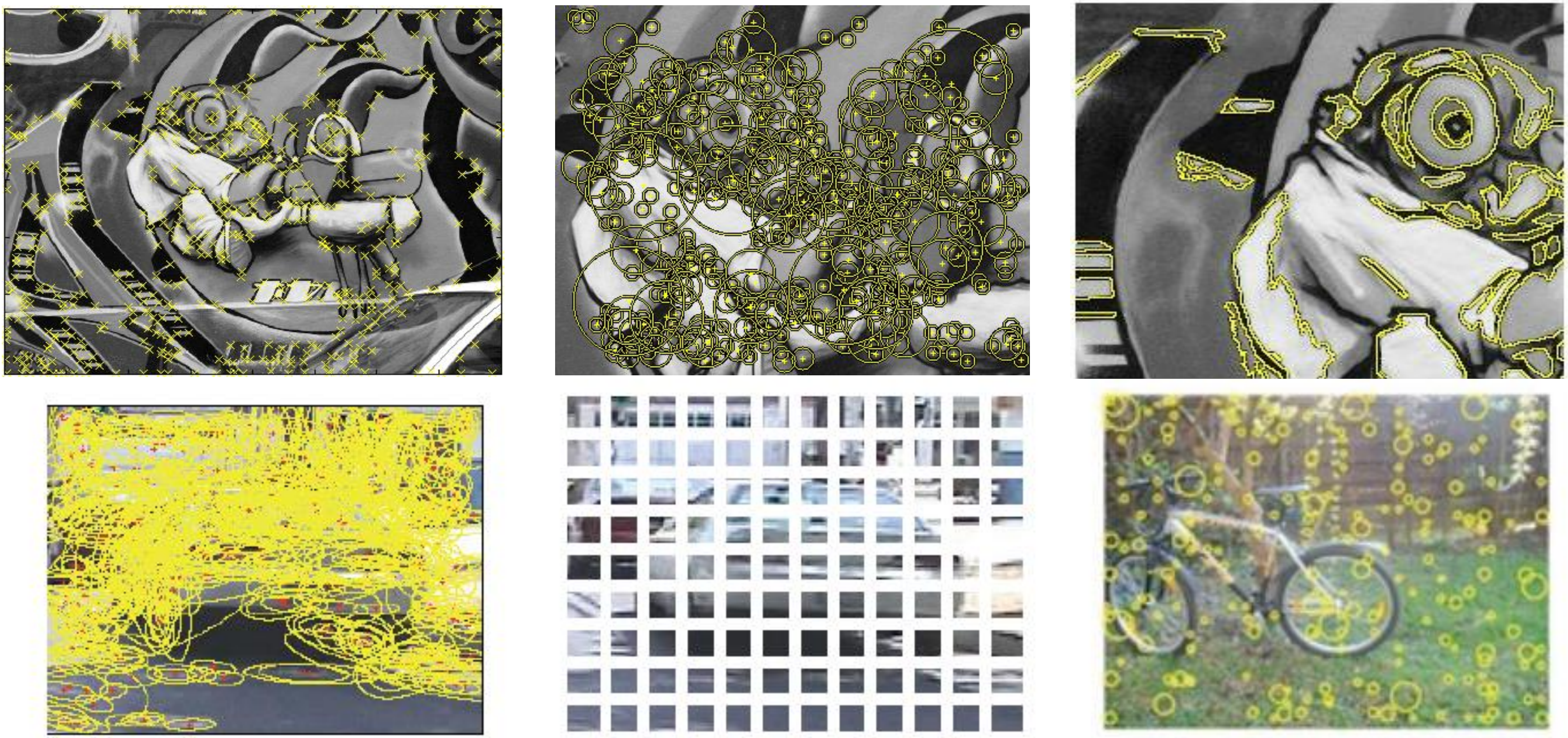
Course Outline

1. Computer Vision and Machine Learning

    Visual (local) feature detection

Local feature detection and description

Points/Regions of Interest detection



Sparse, at interest points

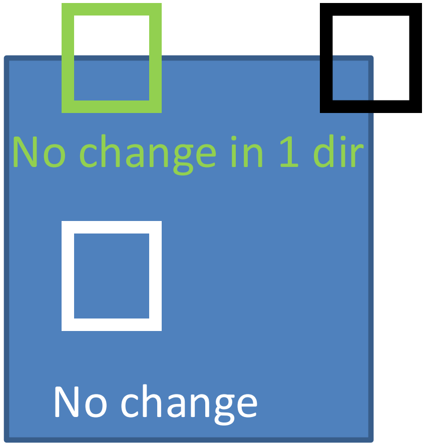
Dense, uniformly

Randomly

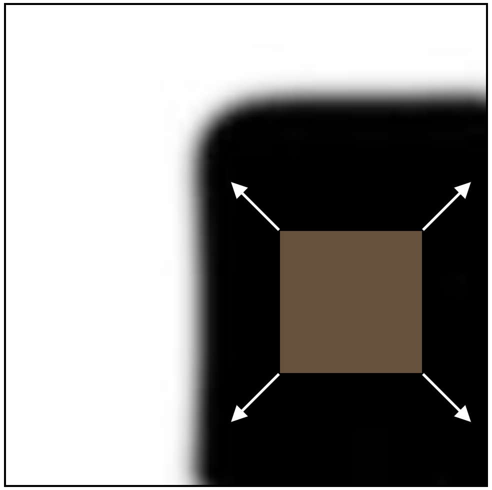
One example: Corner detection (Harris corner detector)

Corner detection

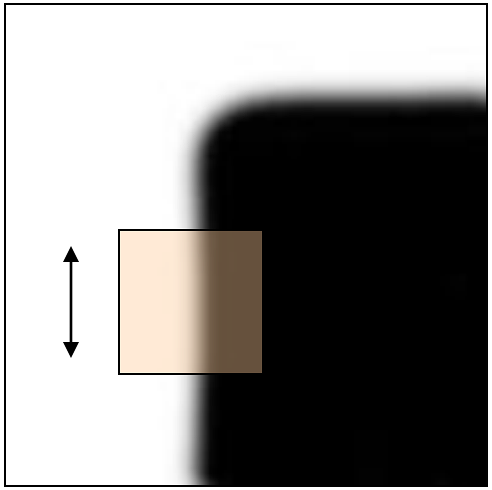
- Corner point: singular point highly informative, rare, …
- Basic idea for Algo: For each pixel (x,y) from image I, *translating* a centered window: Iff (x,y) is a corner, it should cause large differences in patch appearance (whatever the translation)



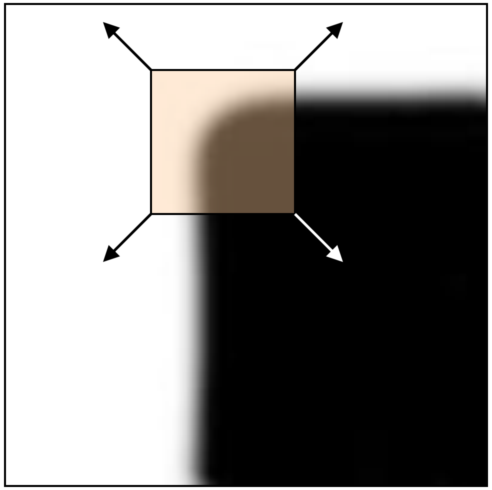
Corner Detection: Basic Idea



“flat” region:
no change in
all directions



“edge”:
no change
along the edge
direction



“corner”:
significant
change in all
directions

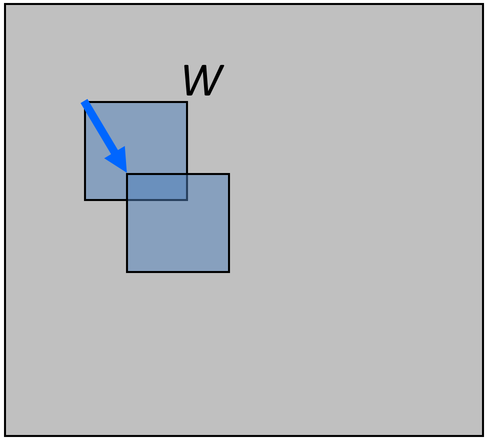
Corner detection op == For all pix, shift a window in *any direction*, keep the ones that give *a large change* in intensity

Corner detection

Consider a small window W, a shifting vector (*u,v*):

- how do the pixels (x,y) in W change?
- compare each pixel before and after by summing up the squared differences (SSD)
- this defines an SSD "error" *E*(*u,v*):

$$E(u,v) = \sum_{(x,y)\in W} \left[I(x+u,y+v) - I(x,y)\right]^2$$



- If E(u,v) high, the center of W is candidate to be a corner

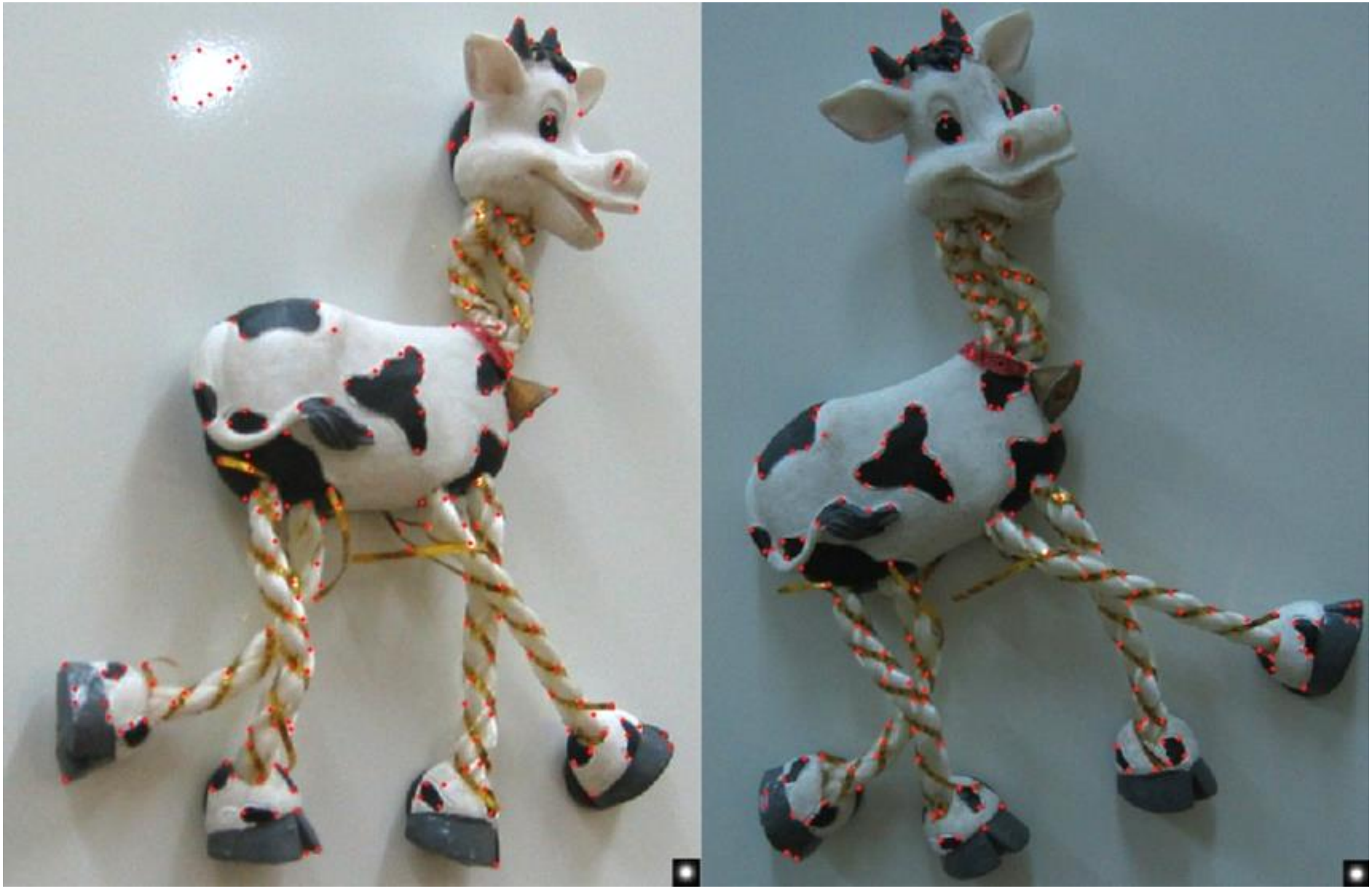
Repeat for for all shifting directions (u,v)

Finally, the center of W is a corner if E(u,v) high for all (u,v)

This ALGO is very computationally expensive => Simplified to get the Harris corner detector

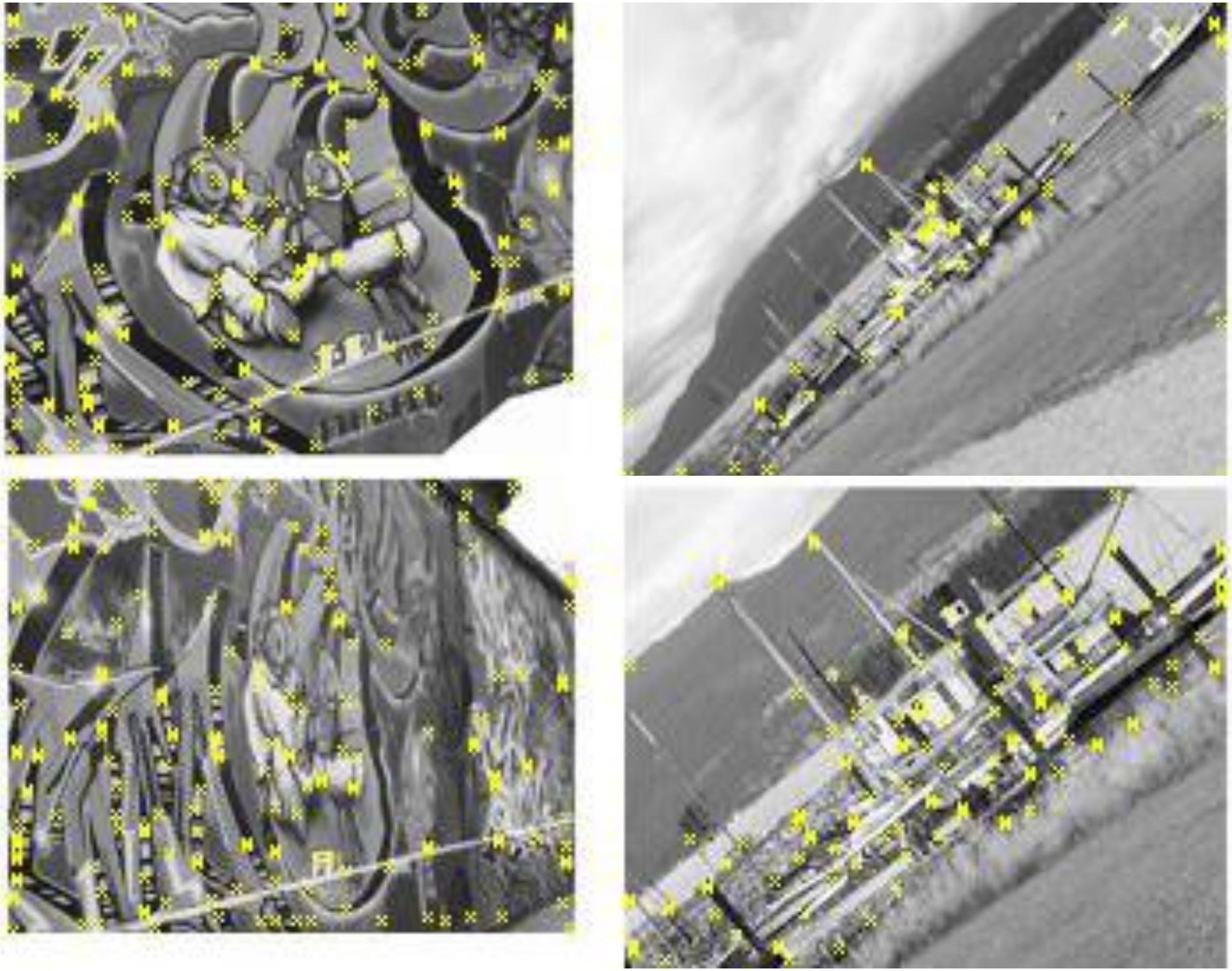
Harris detector example

Harris features (in red)



Local feature detection

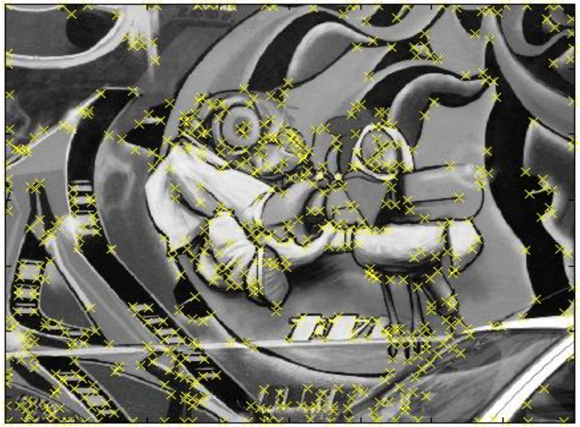
Looking for repeatability

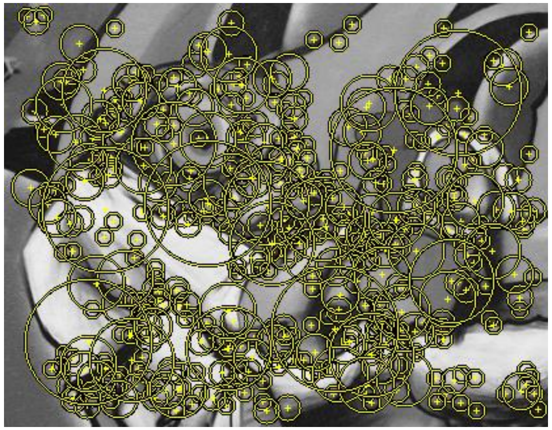


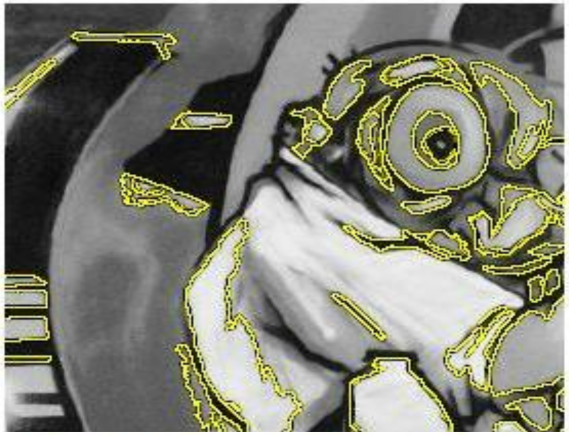
Local feature detection

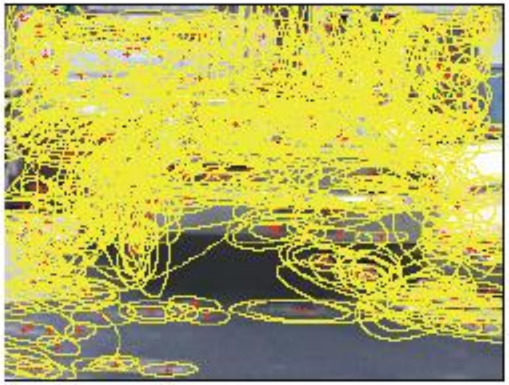
One example: Corner detection (Harris corner detector)

Many other Points/Regions of Interest detectors

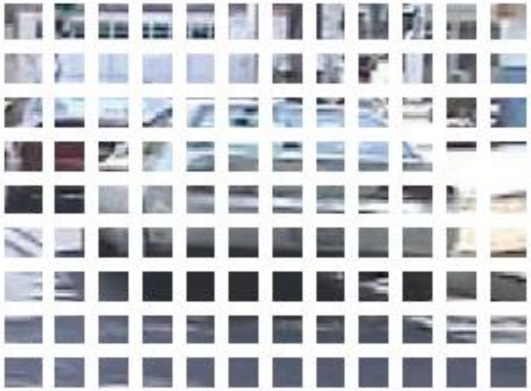




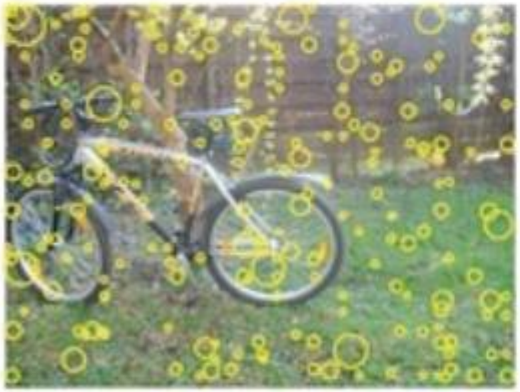




Sparse, at interest points



Dense, uniformly



Randomly

Course Outline

1. Computer Vision and Machine Learning basics

   Visual (local) feature detection

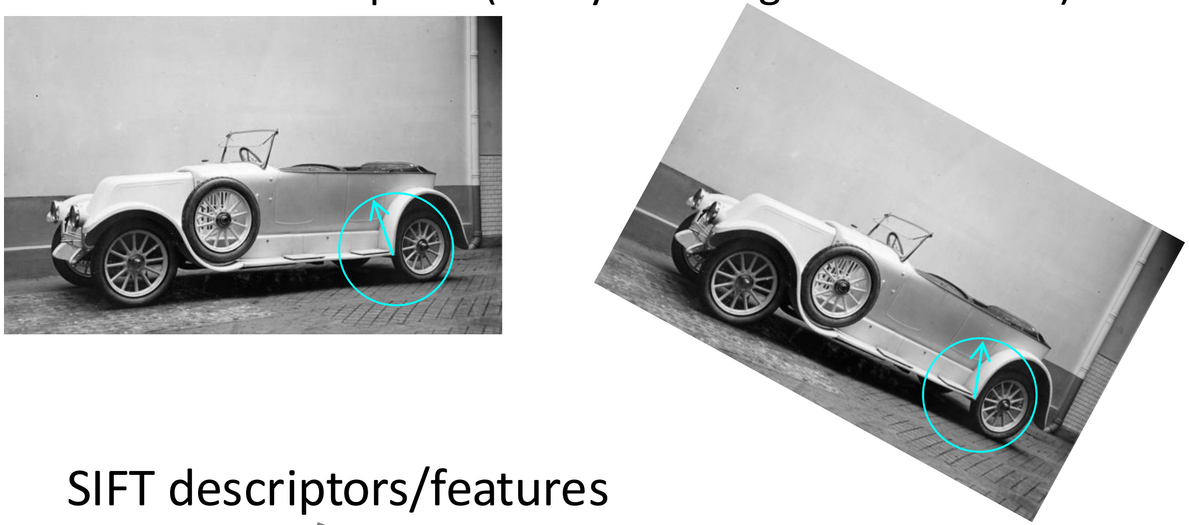
   Visual (local) feature description

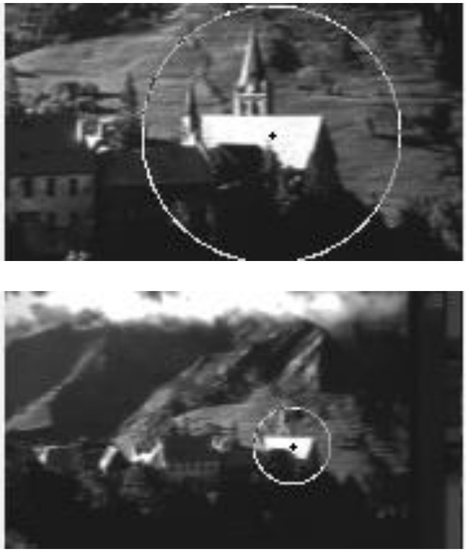
Local feature description

Many Points/Regions of Interest descriptors

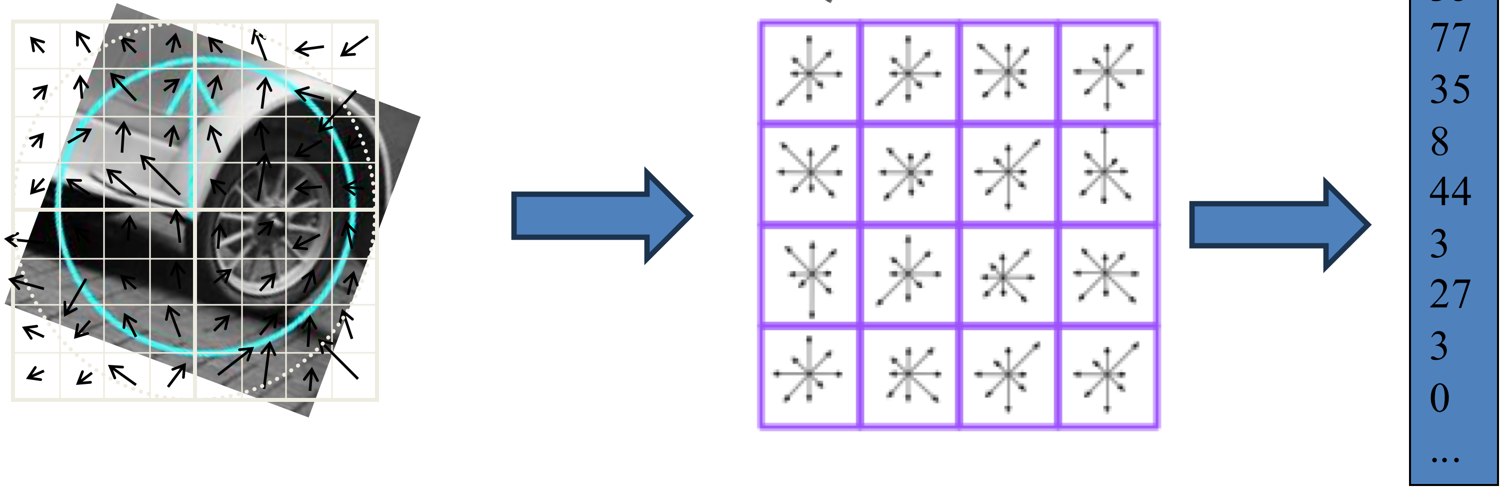
One example: SIFT descriptor

Local description (always looking for invariance)



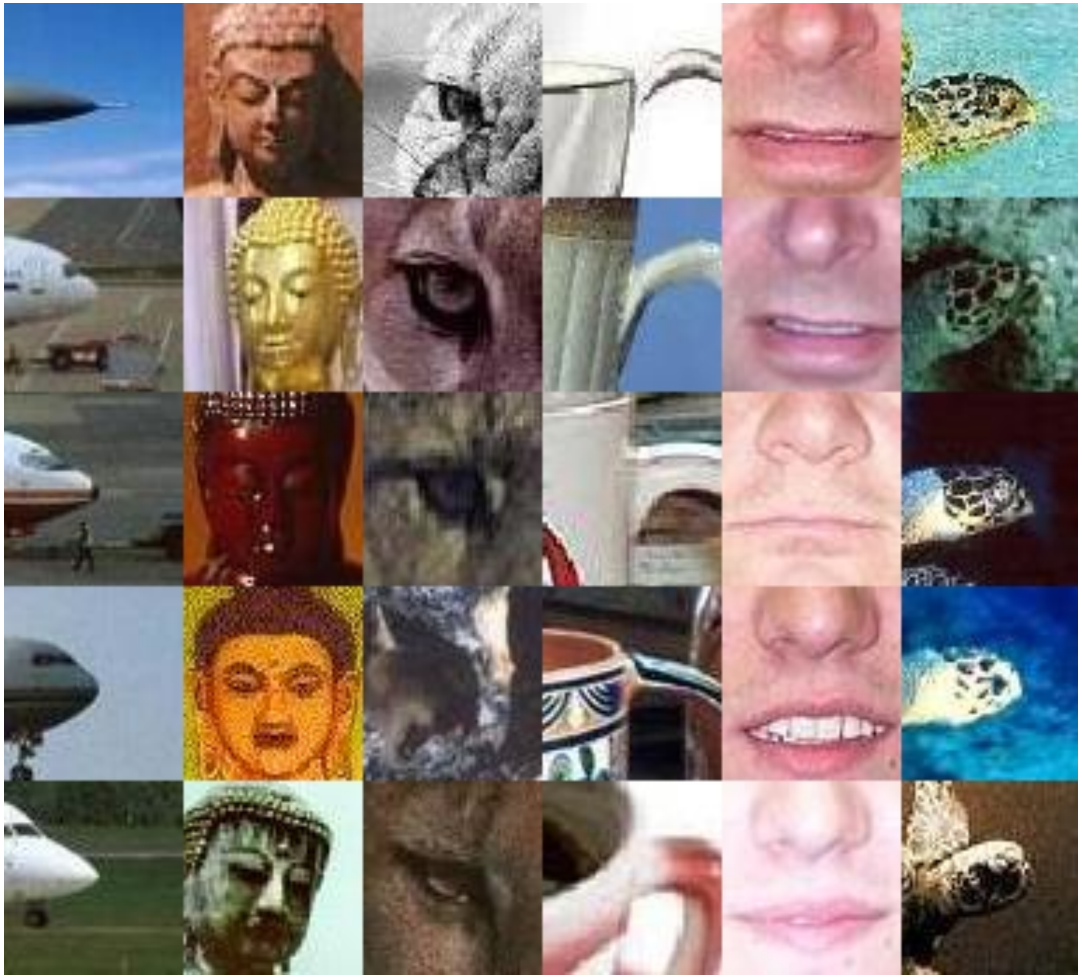


SIFT descriptors/features

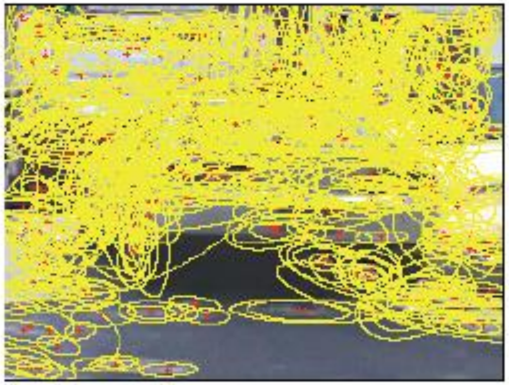


Feature descriptors

- Expected properties?
  - Similar patches => close descriptors
  - Invariance (robustness) to geom. transformation : rotation, scale, view point, luminance, semantics ? …



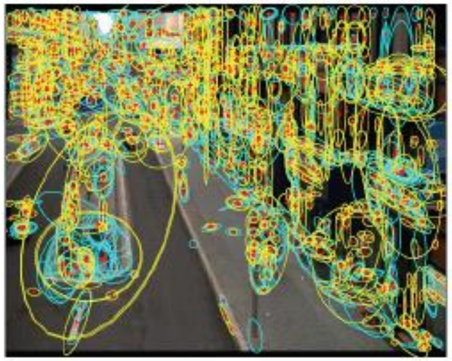
BoF: (First) Image representation



Sparse, at interest points

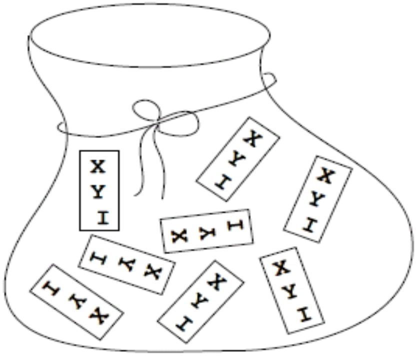
Dense, uniformly

Randomly



Multiple interest operators

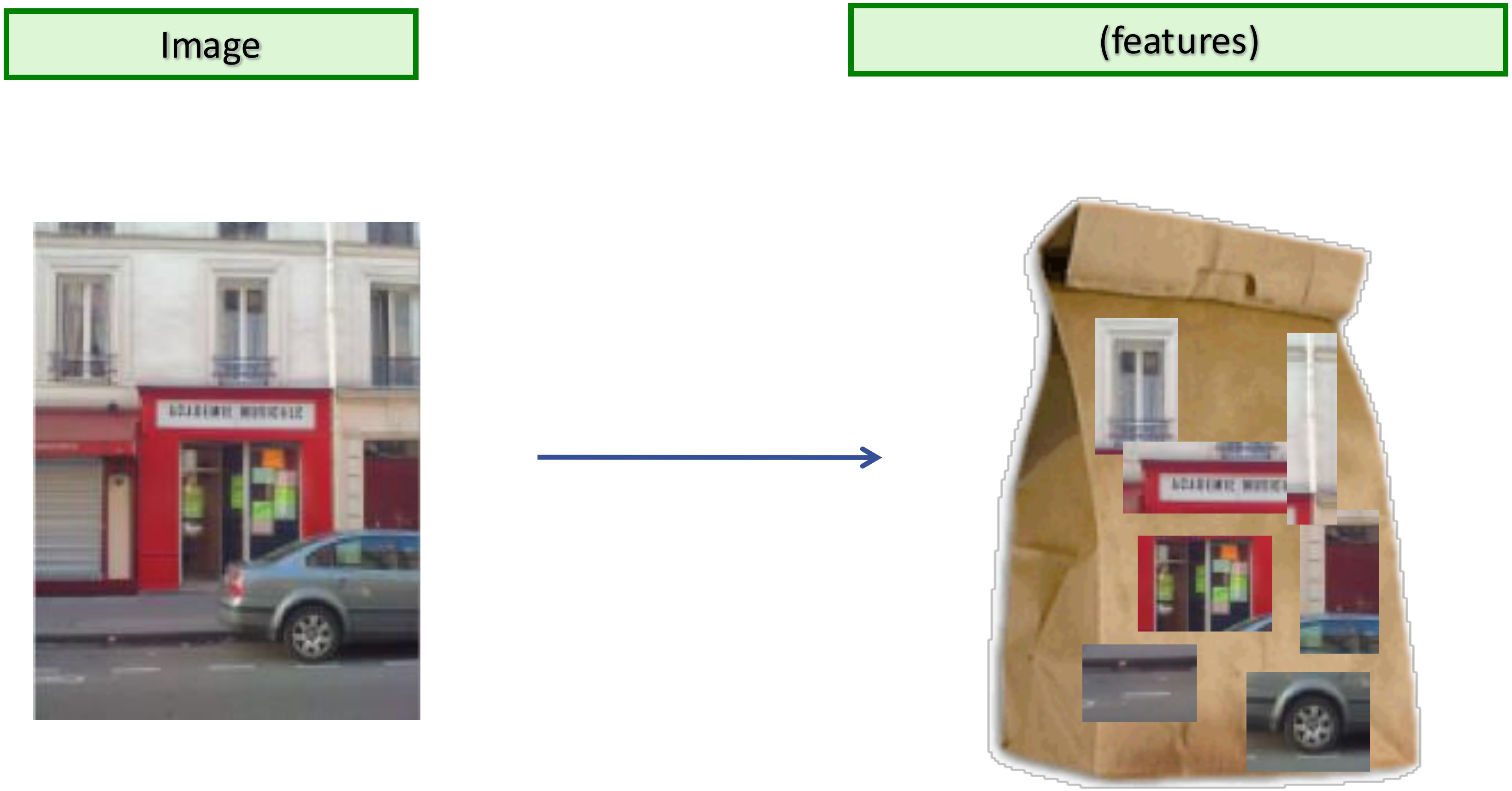
Feature extraction



A bag of features
BoF

© F-F. Li, E. Nowak, J. Sivic

Bag of Feature (BoF) Model



Image

(features)

Image repsentation

- BoF (Bag of features)
  - Local signatures: not a scalable representation
  - Not a *semantic* representation

- The missing bits: **the visual word**
- From BoF to Bag of (Visual) words

Course Outline

1. Computer Vision and Machine Learning basics

   Visual (local) feature detection

   Visual (local) feature description

   Bag of Word Image representation

Course Outline

1. Computer Vision and Machine Learning basics

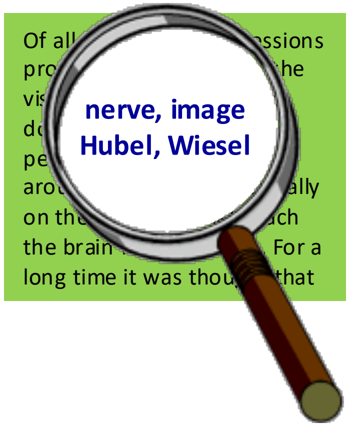
   Visual (local) feature detection

   Visual (local) feature description

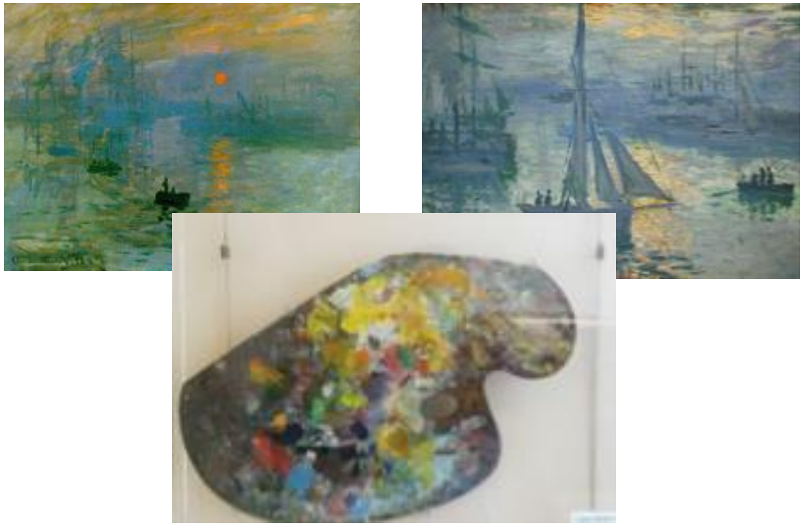
   Bag of Word Image representation

   1. Introduction to Bag of Words
   2. Visual Dictionary
   3. Image signature
   4. Whole recognition pipeline

Bag of Words (BoW) model: basic explication with textual representation and color indexing



China is forecasting a trade
surpl bn) to
$10 efold
inc .
Th aid
the
create 30%
jump to rise further value.
China, trade,
surplus



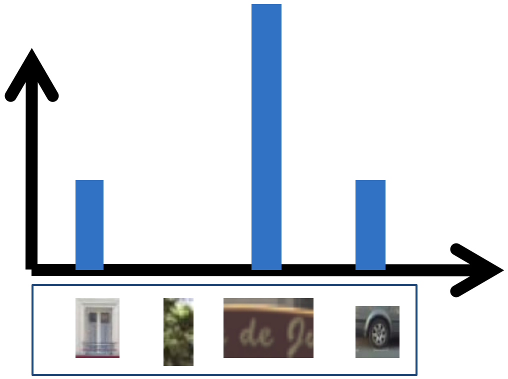
Comparing 2 docs using visual/color/word occurrences

Slide credit L. Fei-Fei

Bag of Visual Words (BoW)

(features)

BoW : histogram on visual dictionary



Questions:

1. Which dictionary ?
2. How to project the BoF onto the dico
3. How to compute the histogram?

Course Outline

1. Computer Vision Introduction:

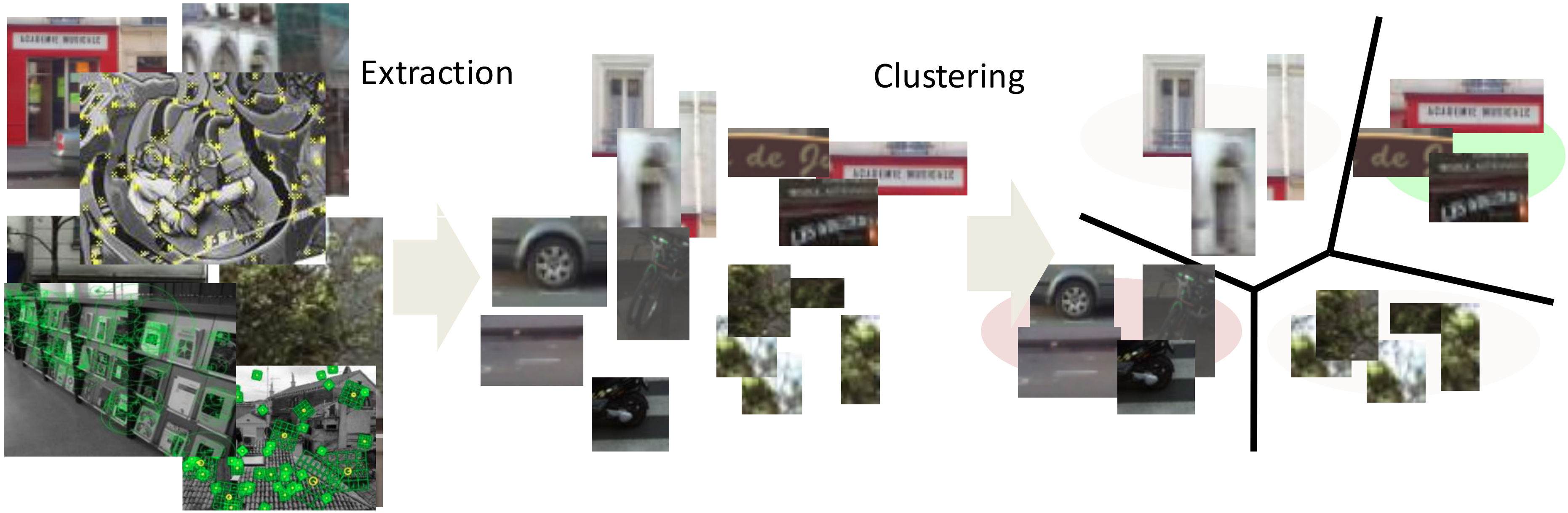
   Visual (local) feature detection and description,

   Bag of Word Image representation

   1. Introduction to Bag of Words
   2. **Visual Dictionary**
   3. Image signature
   4. Whole recognition pipeline

Visual space clustering

1. Extraction of local features (pattern/visual words) in images
   - Training dataset in classification
   - Image dataset in retrieval
2. Clustering of feature space



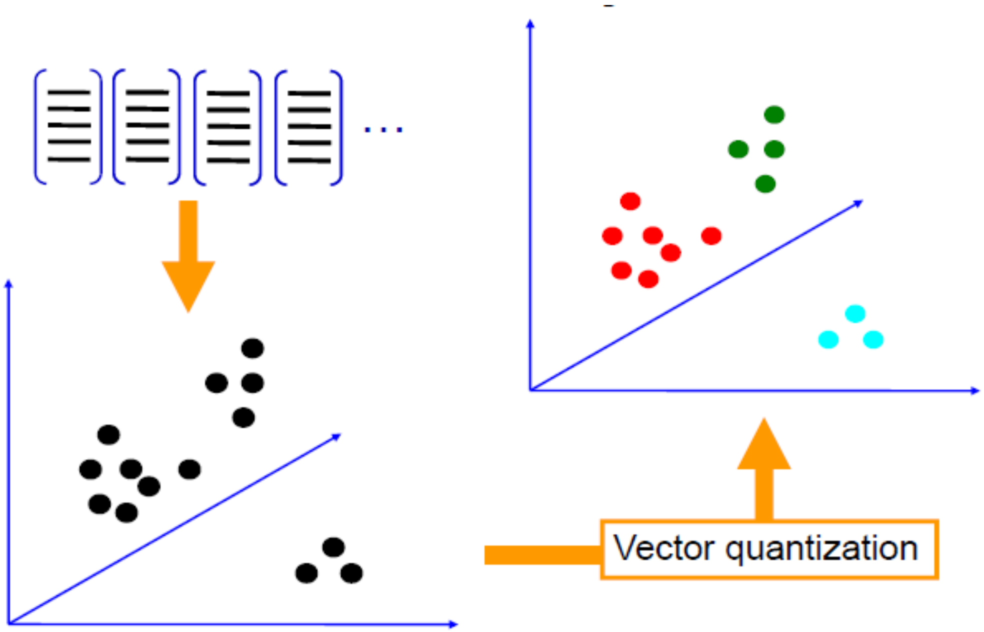
Extraction

Clustering

Training set but no labels => UNSUPERVISED Learning

Visual space clustering

- Many algorithms for clustering :
  - K-Means
  - Vectorial Quantization
  - Gaussian Mixture Models
  - …



Clustering with K clusters

Input: Set of n points $\{x_i\}_n$ in $\mathbb{R}^d$

Goal: Find a set of K ($K \ll n$) points $w = \{w_k\}_K$ that gives an approximation of the n input points, i.e., minimizing the error criterion $C(w)$:

$$C(w) = \sum_{i=1}^{n} \min_{k} \|x_i - w_k\|^2$$

At fixed K, complexity is $O(n^{(Kd+1)} \log(n))$

There are many strategies to approximate the global optimization problem.

Clustering with K-means algo

$$C(w) = \sum_{i=1}^{n} \min_{k} \|x_i - w_k\|^2$$

Init K centers $\{c_k\}$ by sampling K points w_k in $\mathbb{R}^d$

$$\min_{k} \|x_i - w_k\|^2$$

1. (Re)assign each point x_i to the cluster s_i with the center w_{s_i} so that dist(x_i, w_{s_i}) is less than dist from x_i to any other current cluster center.

$$\sum_{i=1}^{n} \|x_i - w_{s_i}\|^2$$

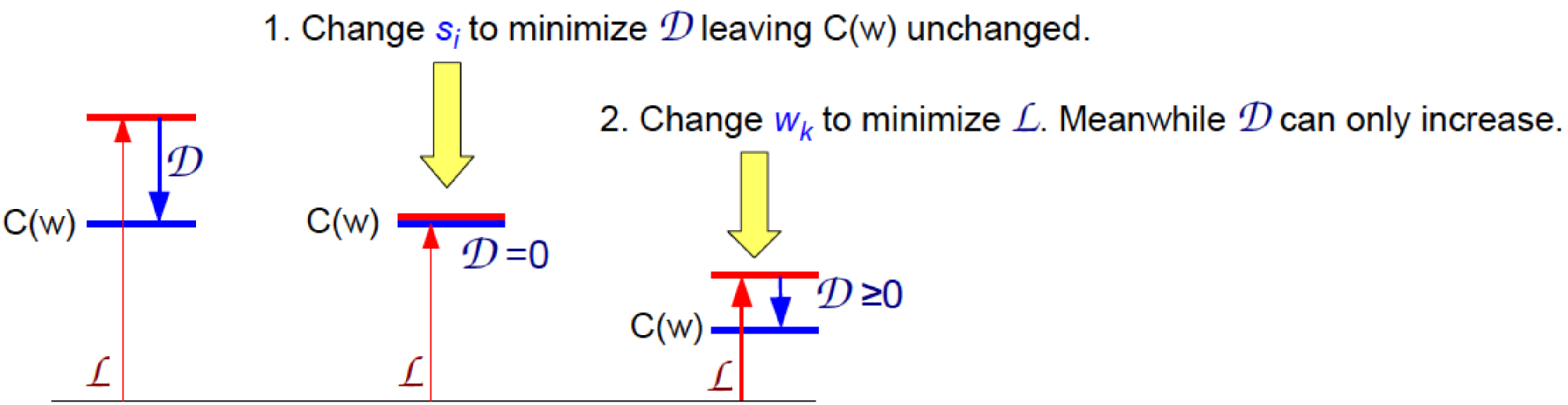
2. Move all w_k inside each cluster as the new barycenter from all the points assigned to the cluster k (equivalent to minimizing the corresponding mean square error).
3. Go to step 1 if some points changed clusters during the last iteration.

Output: The set of the final K cluster centers $\{c_k = w_k\}$

K-means : why it is successful ?

Consider an arbitrary cluster assignment s_i.

$$C(w) = \sum_{i=1}^{n} \min_k \|x_i - w_k\|^2 = \underbrace{\sum_{i=1}^{n} \|x_i - w_{s_i}\|^2}_{\mathcal{L}(s,w)} - \underbrace{\sum_{i=1}^{n} \|x_i - w_{s_i}\|^2 - \min_k \|x_i - w_k\|^2}_{\mathcal{D}(s,w) \geq 0}$$



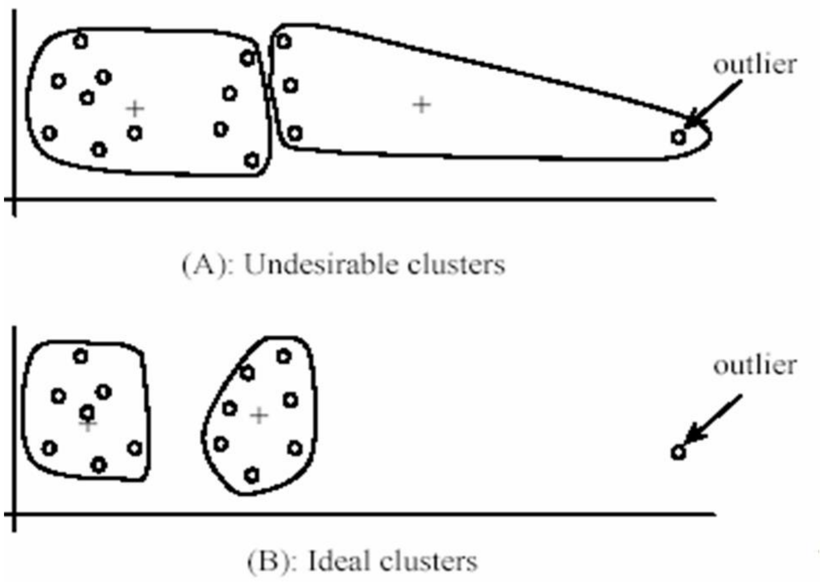
1. Change s_i to minimize $\mathcal{D}$ leaving C(w) unchanged.

2. Change w_k to minimize $\mathcal{L}$. Meanwhile $\mathcal{D}$ can only increase.

© L. Botou

Clustering

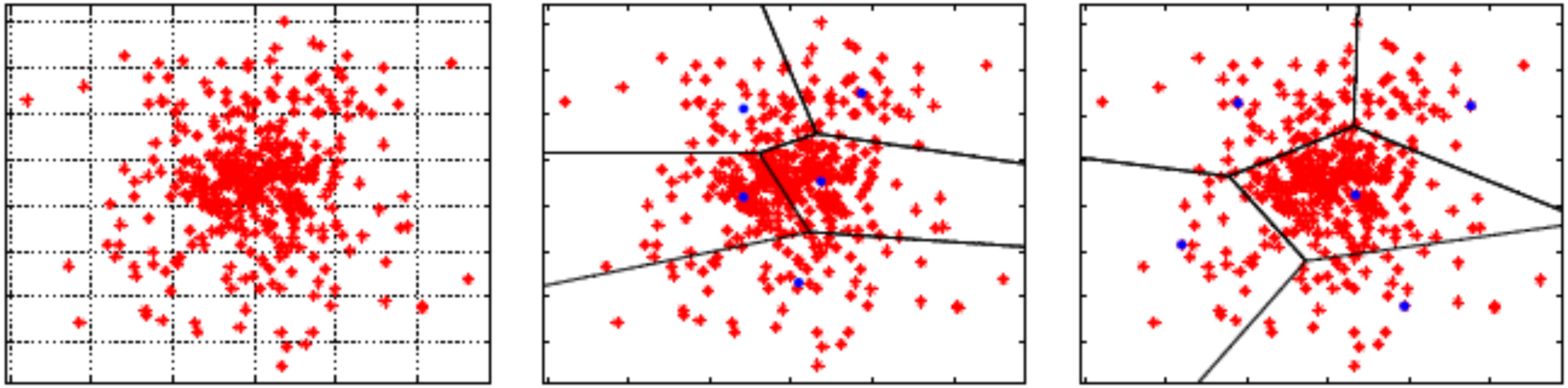
- K-means :
  - Pros
    - Simplicity
    - Convergence (local min)
  - Cons
    - Memory-intensive
    - Depending on K
    - Sensitive to initialization
    - Sensitive to artifacts
    - Limited to spherical clusters
    - Concentration of clusters to areas with high densities of points (Alternatives : radial based methods)



- K-Means deeply used in practice

Clustering

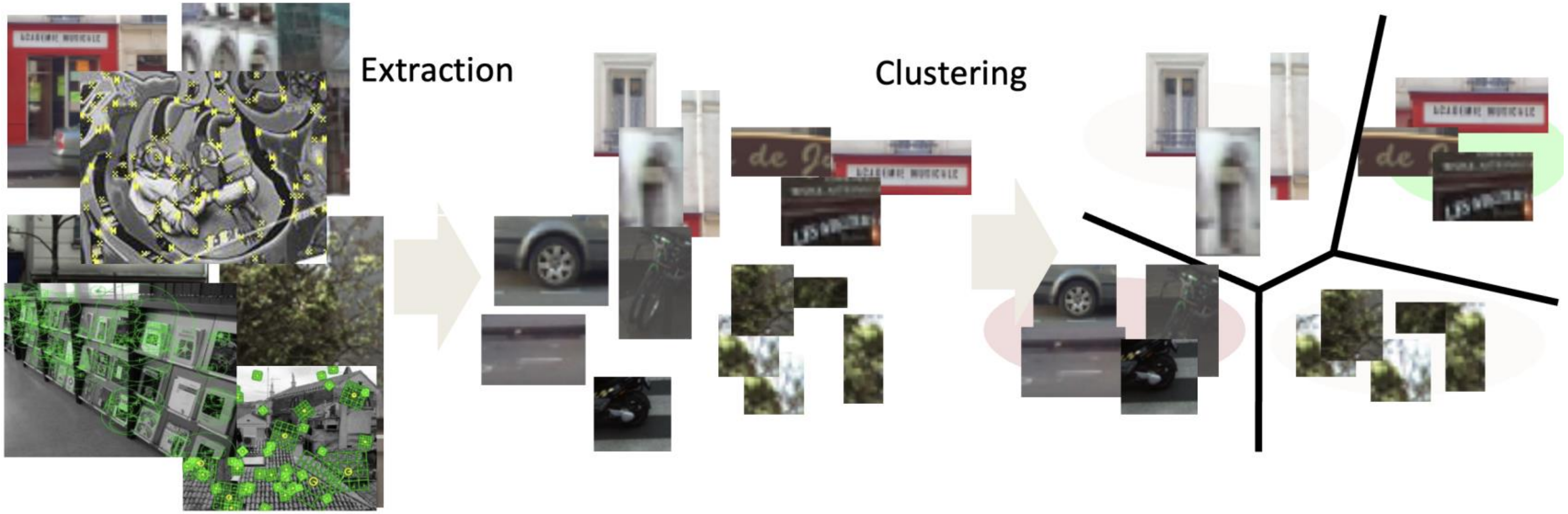
- Uniform / K-means / radius-based :



(a) Histogram (b) *K*-means (c) Radius-based

- *Radius-based clustering assigns all features within a fixed radius of similarity r to one cluster.*

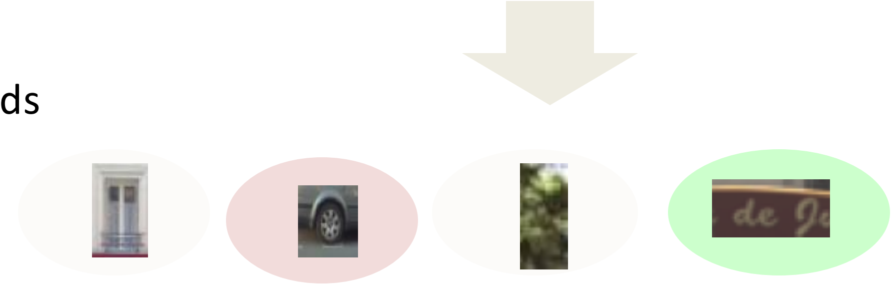
Dictionary = K Visual words



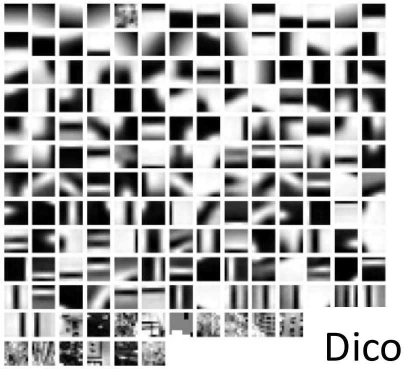
Extraction

Clustering

Dico extraction



Centers = dico. Visual words



Dico examples

Course Outline

1. Computer Vision Introduction:

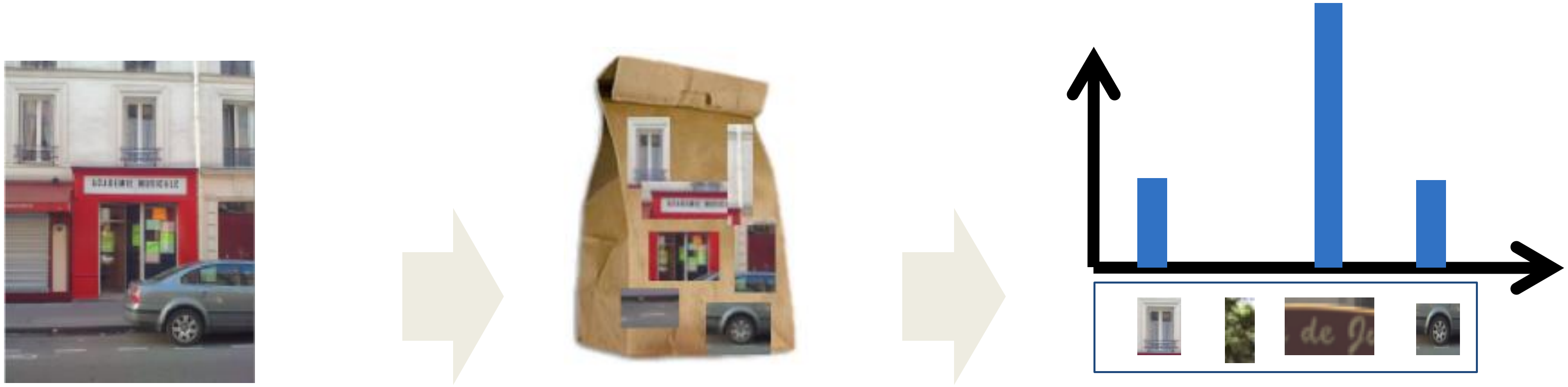
   Visual (local) feature detection and description,

   Bag of Word Image representation

   1. Introduction to Bag of Words
   2. Visual Dictionary
   3. **Image signature**
   4. Whole recognition pipeline

Bag-of-Words (BoW) image signature

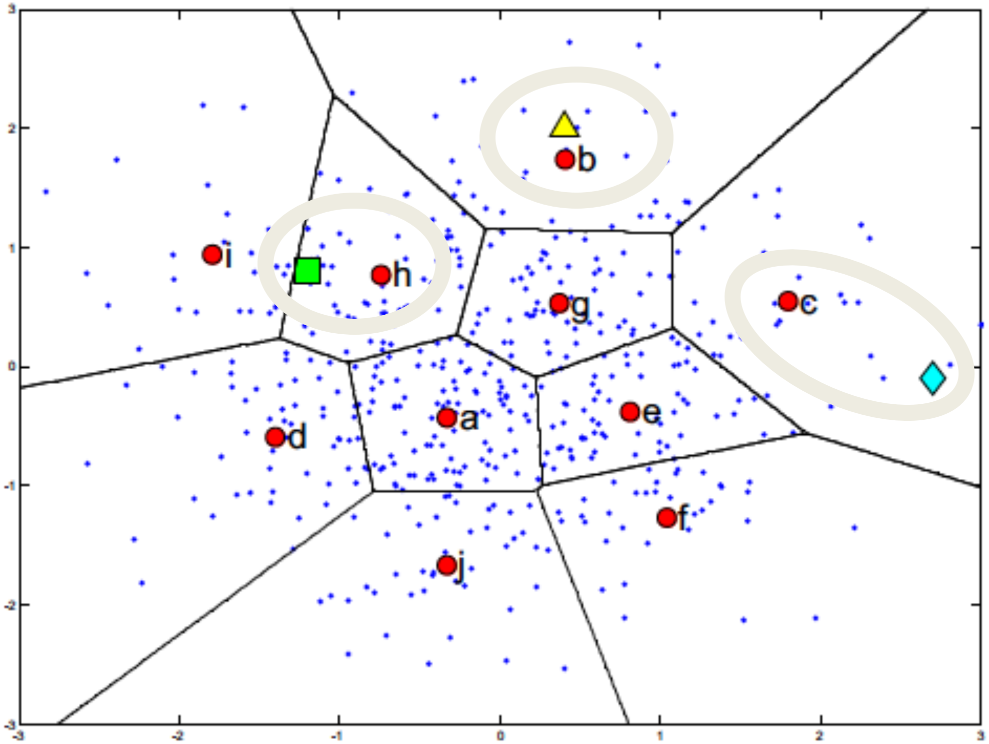
- For each image:
  - For each local feature: find the closest visual word
  - Increase the corresponding bin in histogram of visual dico

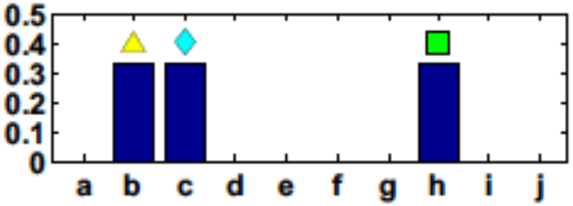


- Image signature (global Index):
  - Vector (histogram of M bins)
  - M= dimension K = dico size
  - Each term represents a Likelihood to get this visual word

Bag-of-Words (BoW) image signature

- Original BoW strategy: **hard assignement/coding**
  - Find the closest cluster for each feature
  - Assign a fix weight (*e.g.* 1)





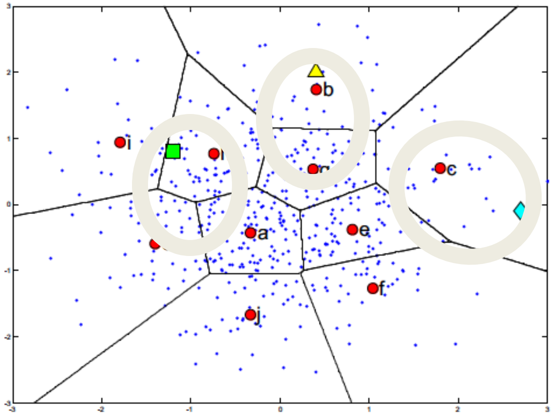
Traditional Codebook

Bag-of-Words (BoW) image signature

Sum pooling : initial BoW strategy (just counting occurrences of words in the document)

Classical BoW = **hard coding + sum pooling**

1. Find the closest cluster for each feature
2. Assign a fix weight (*e.g.* 1) to this cluster



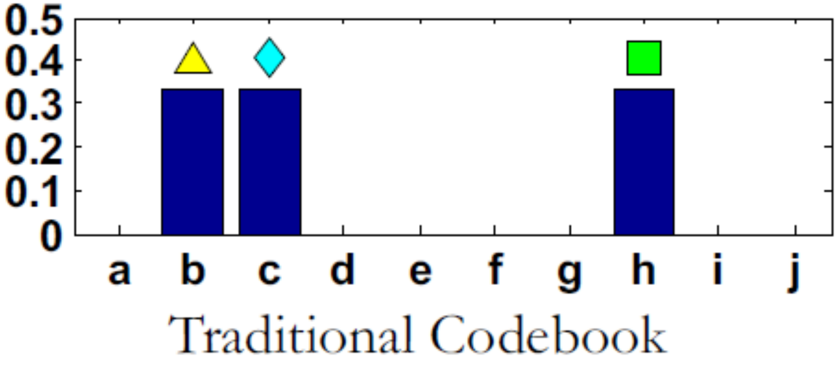
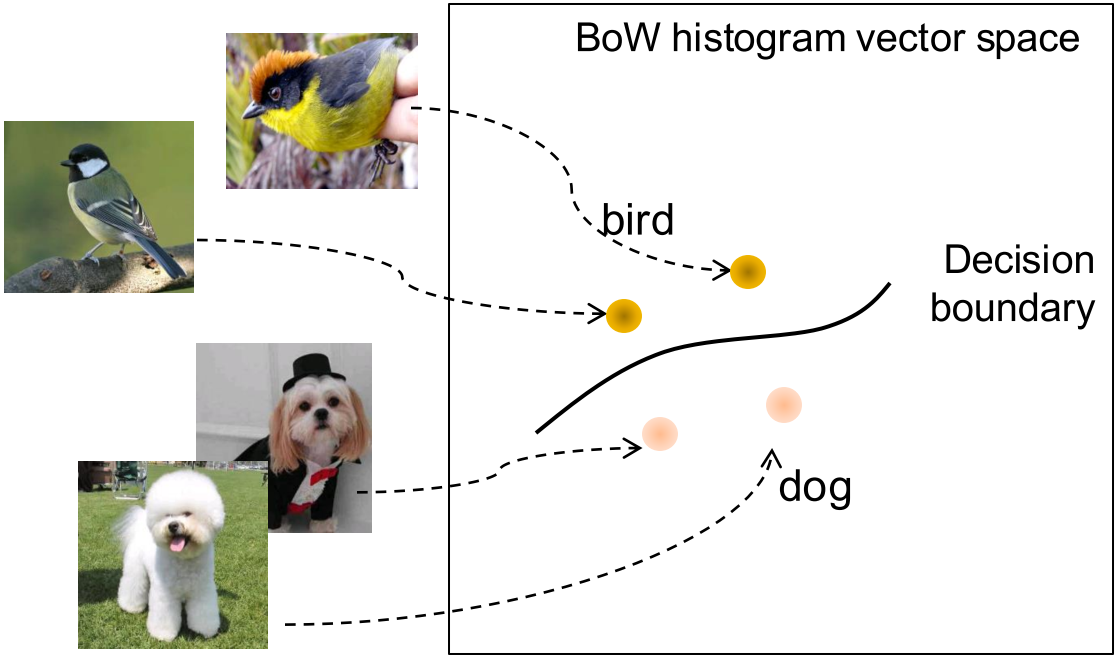
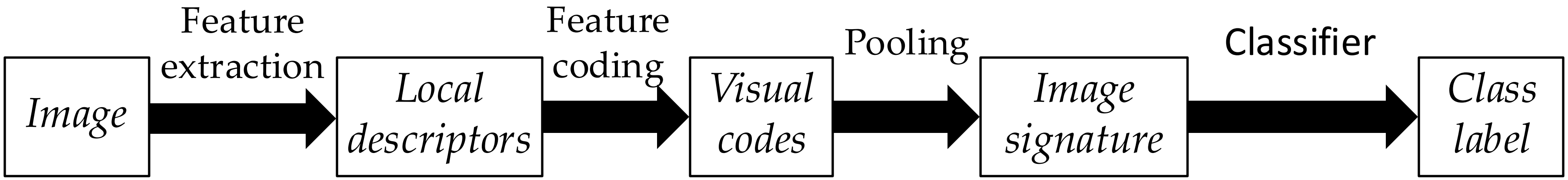


Image classification based on BoW



Learn a classification model to determine the decision boundary

Classification model to determine the decision boundary



SVM

- Image/Patterns $\mathbf{x} \in \mathbf{X}$
- Φ: function transforming the patterns into feature vectors $\Phi(\mathbf{x})$ (like BoW or on top of it)
- $\langle \cdot, \cdot \rangle$: dot product in the feature space endowed by $\Phi(\cdot)$
- SVM: binary classifier. Classes $y = \pm 1$

- Taking the sign of a linear discriminant function:

$$f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b$$

SVM

Question: How to find/estimate f?

- Feature function Φ usually hand-chosen for each problem
- Several Φ for image processing, like BoW
- w and b: parameters to be determined

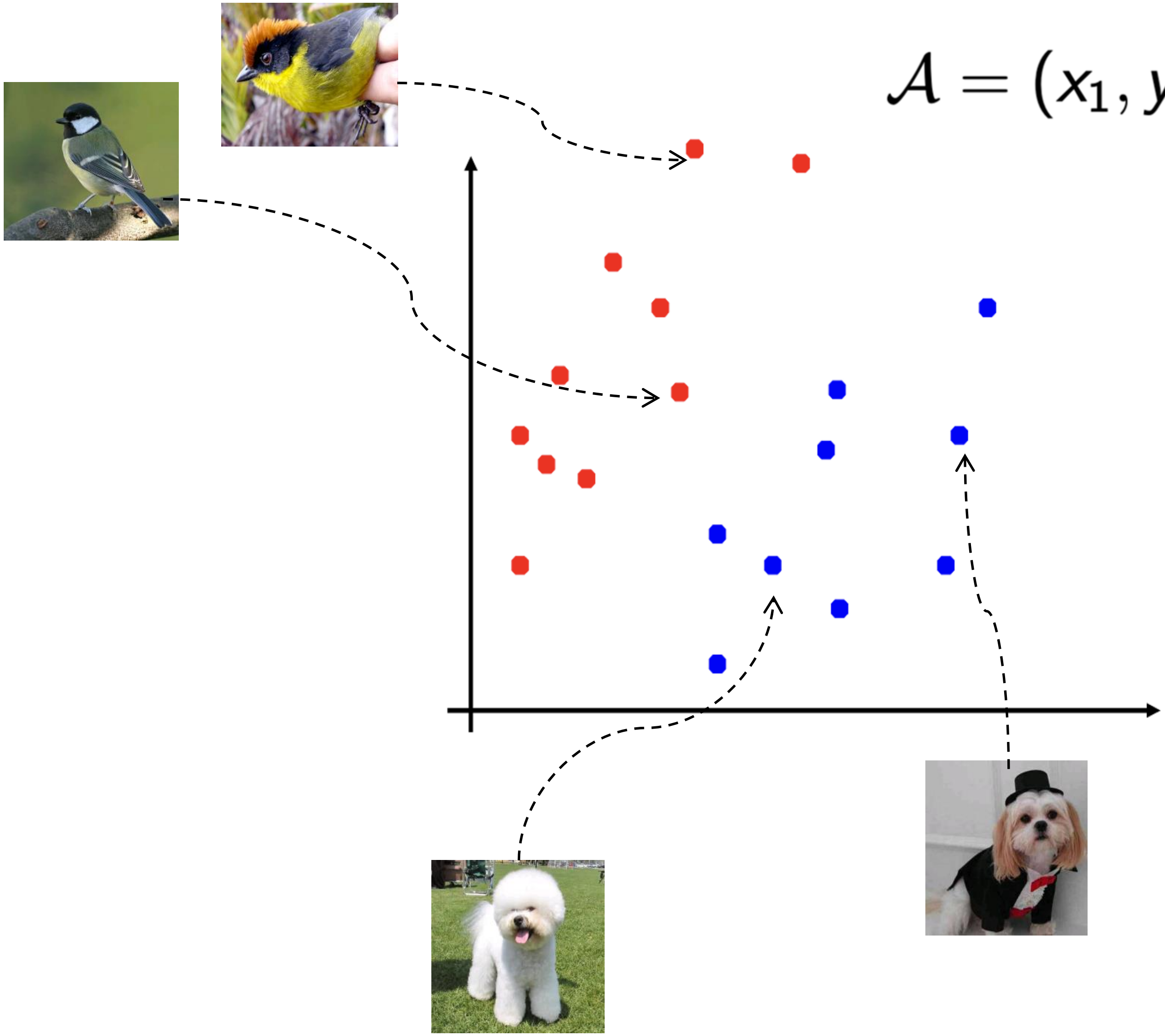
$$f(x) = \langle w, \Phi(x) \rangle + b$$

Learning algorithm on a set of training examples:

$$\mathcal{A} = (x_1, y_1) \cdots (x_n, y_n)$$

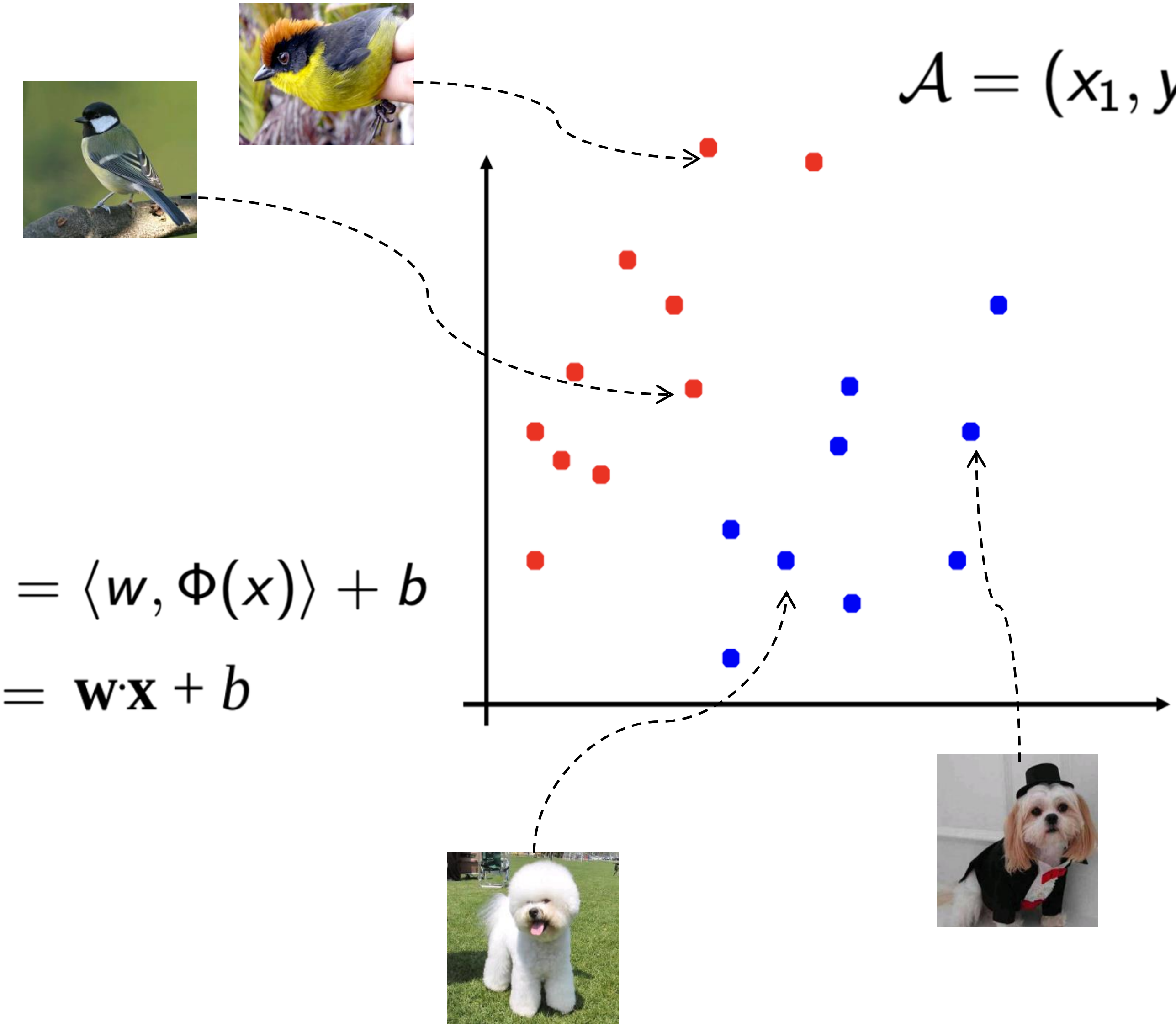
Which hyperplane w? and b?

$$\mathcal{A} = (x_1, y_1) \cdots (x_n, y_n)$$



Which hyperplane w? and b?

$$\mathcal{A} = (x_1, y_1) \cdots (x_n, y_n)$$

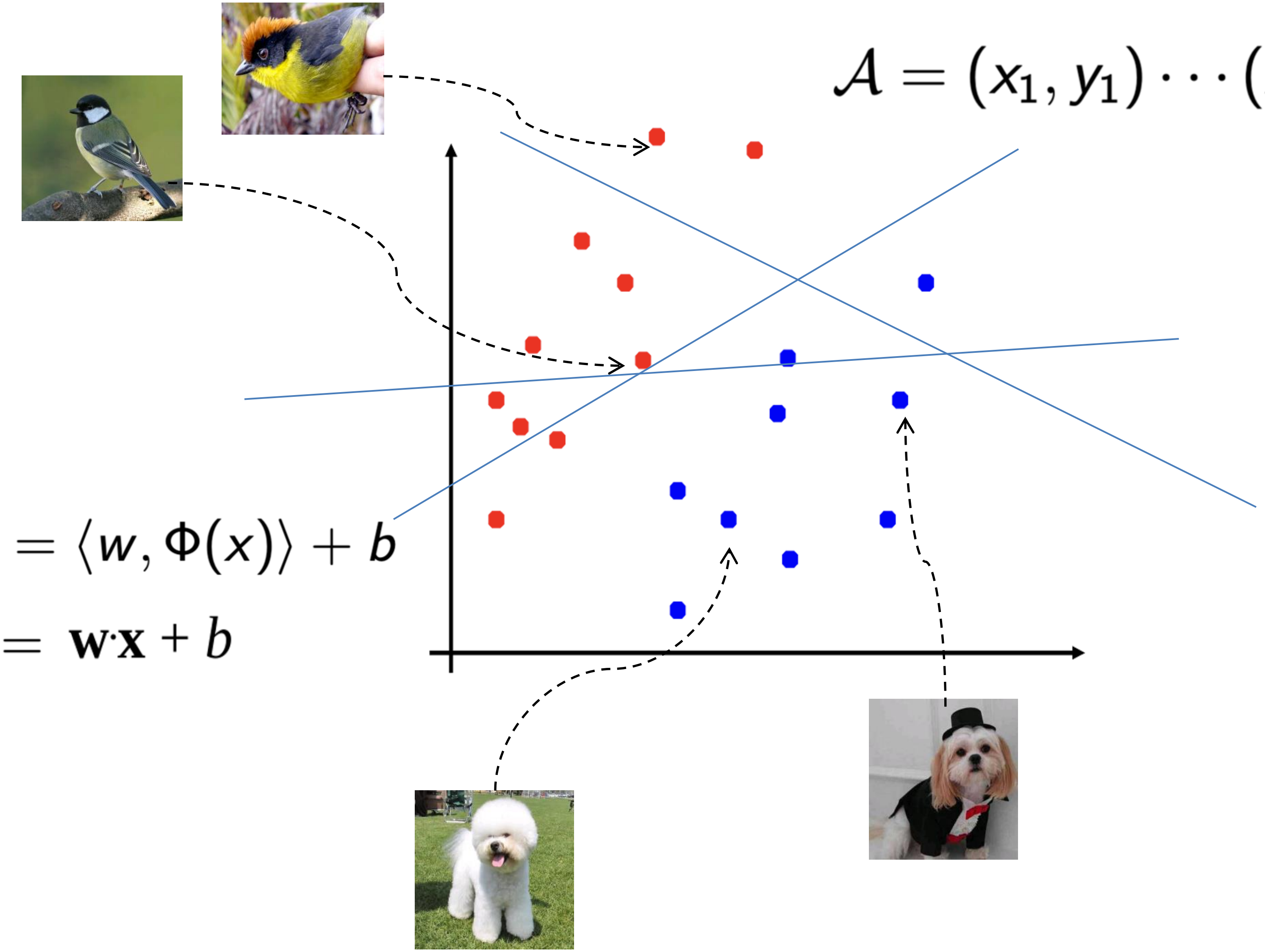


$$f(x) = \langle w, \Phi(x) \rangle + b$$

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b$$

Which hyperplane w? and b?

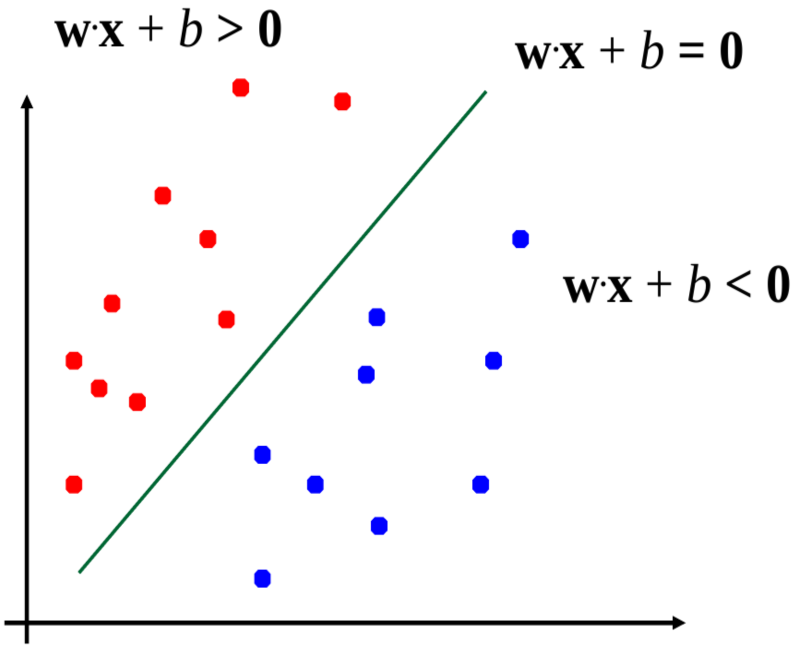
$$\mathcal{A} = (x_1, y_1) \cdots (x_n, y_n)$$



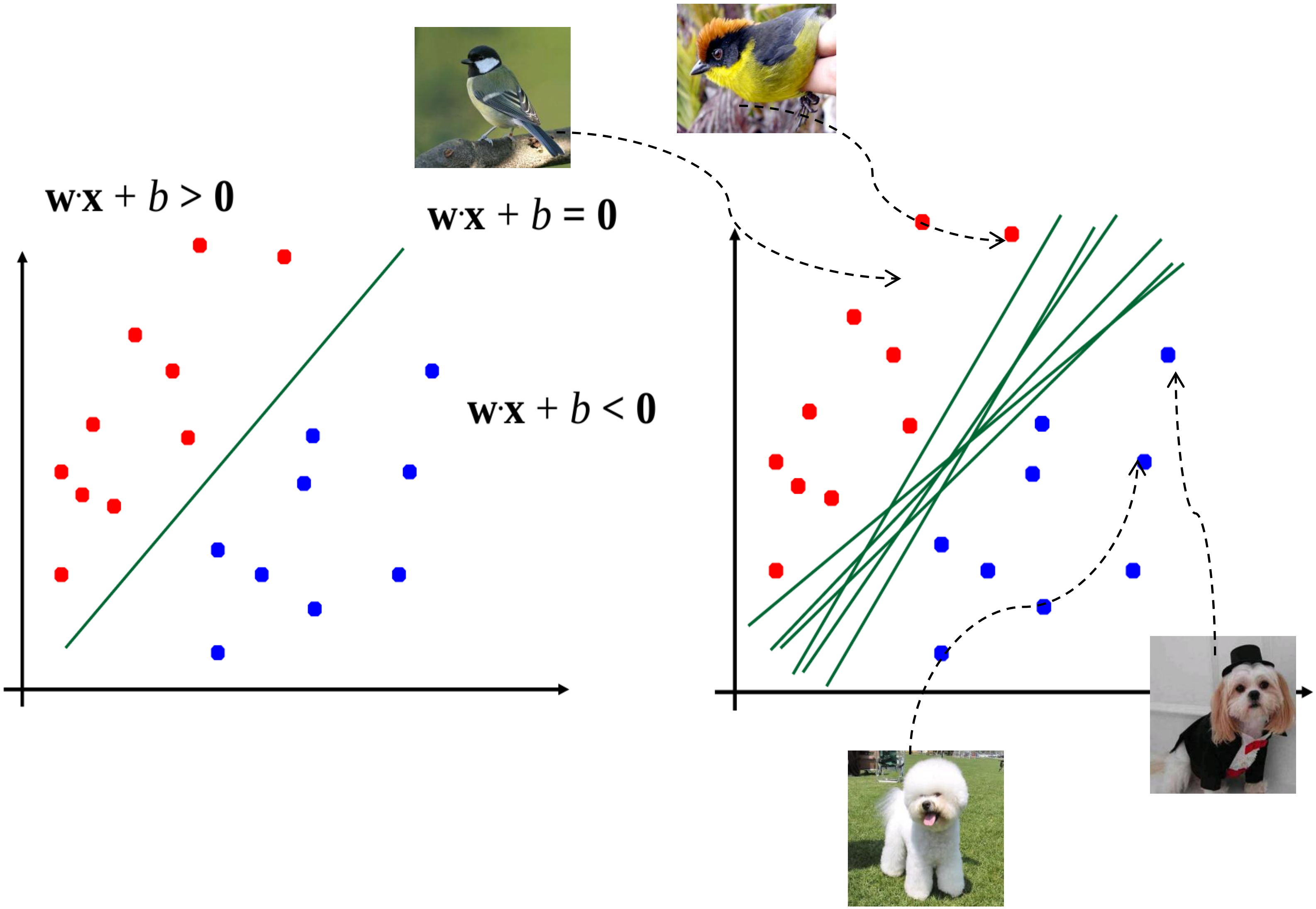
$$f(x) = \langle w, \Phi(x) \rangle + b$$

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b$$

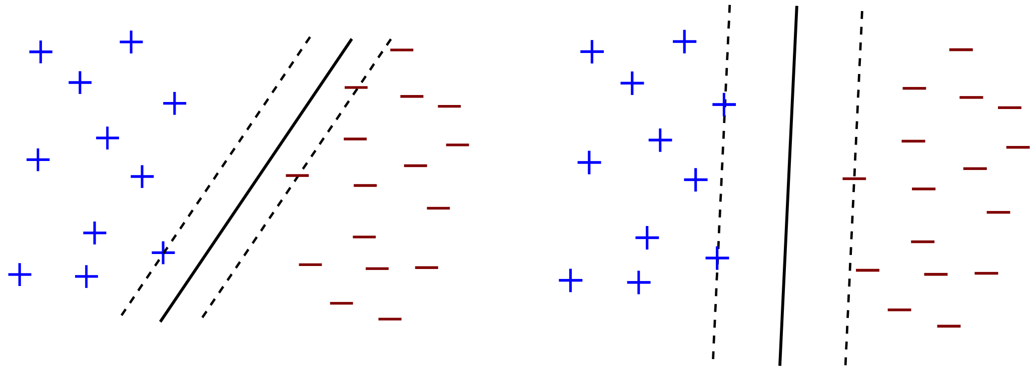
Which hyperplane w? and b?


$\mathbf{w}\cdot\mathbf{x} + b > 0$
$\mathbf{w}\cdot\mathbf{x} + b = 0$
$\mathbf{w}\cdot\mathbf{x} + b < 0$

Which hyperplane w? and b?


$\mathbf{w} \cdot \mathbf{x} + b > 0$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b < 0$

SVM



SVM optimization: maximizing the margin between + and -

Def.: Margin = distance between the hyperplanes $f(x) = 1$and $f(x) = -1$ (dashed lines in Figure).

Intuitively, a classifier with a larger margin is more robust to fluctuations

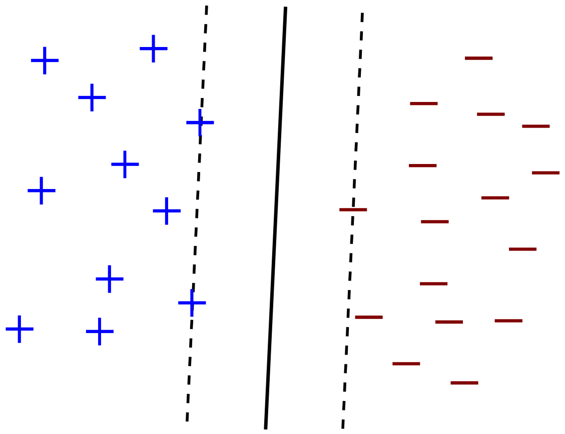
Hard Margin

Final expression for the Hard Margin SVM optimization:

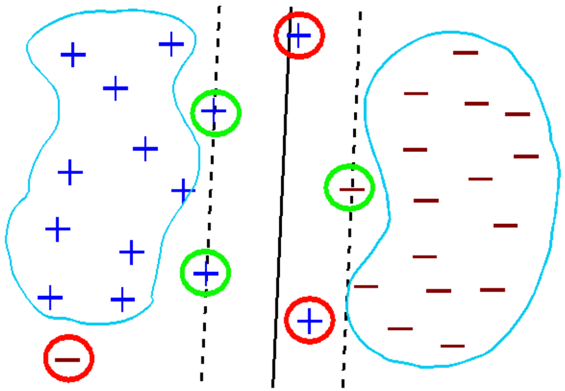
$$\min_{w,b}\ P(w,b) = \frac{1}{2}\|w\|^2 \quad \text{with} \quad \forall\, i \quad y_i\, f(x_i) \geq 1$$

SVM

- Hard Margin: OK if data are linearly separated



- Otherwise: noisy data (in red) disrupt the optim.
- Solution: Soft SVM



SVM: Soft Margin

Introducing the slack variables ξ_i, one usually gets rid of the inconvenient max of the loss and rewrite the problem as

$$\min_{w,b} \; P(w,b) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{with} \quad \left\{ \begin{array}{ll} \forall\, i & y_i\, f(x_i) \geq 1 - \xi_i \\ \forall\, i & \xi_i \geq 0 \end{array} \right.$$

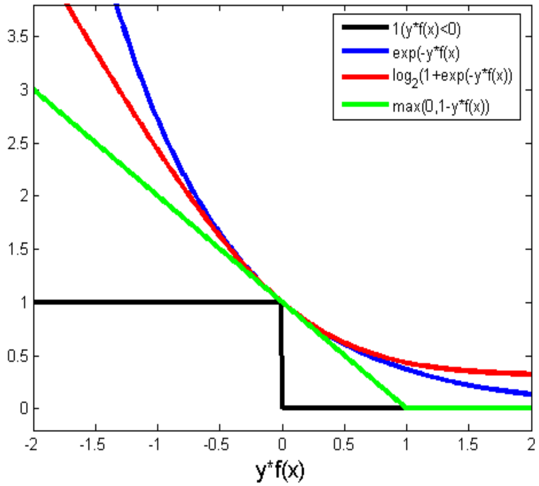
For very large values of the hyper-parameter C, **Hard Margin** case:

– Minimization of ||w|| (ie margin maximization) under the constraint that all training examples are correctly classified with a loss equal to zero.

Smaller values of C relax this constraint: **Soft Margin** case

– SVMs that produces markedly better results on noisy problems.

SVM learning scheme



Equivalently, minimizing the following objective function in feature space with the hinge loss function:

$$\ell(y_i\, f(x_i)) = \max\left(0, 1 - y_i\, f(x_i)\right)$$

$$\min_{w,b}\ P(w,b) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\ell(y_i\, f(x_i))$$

Regularization

Margin Maximization

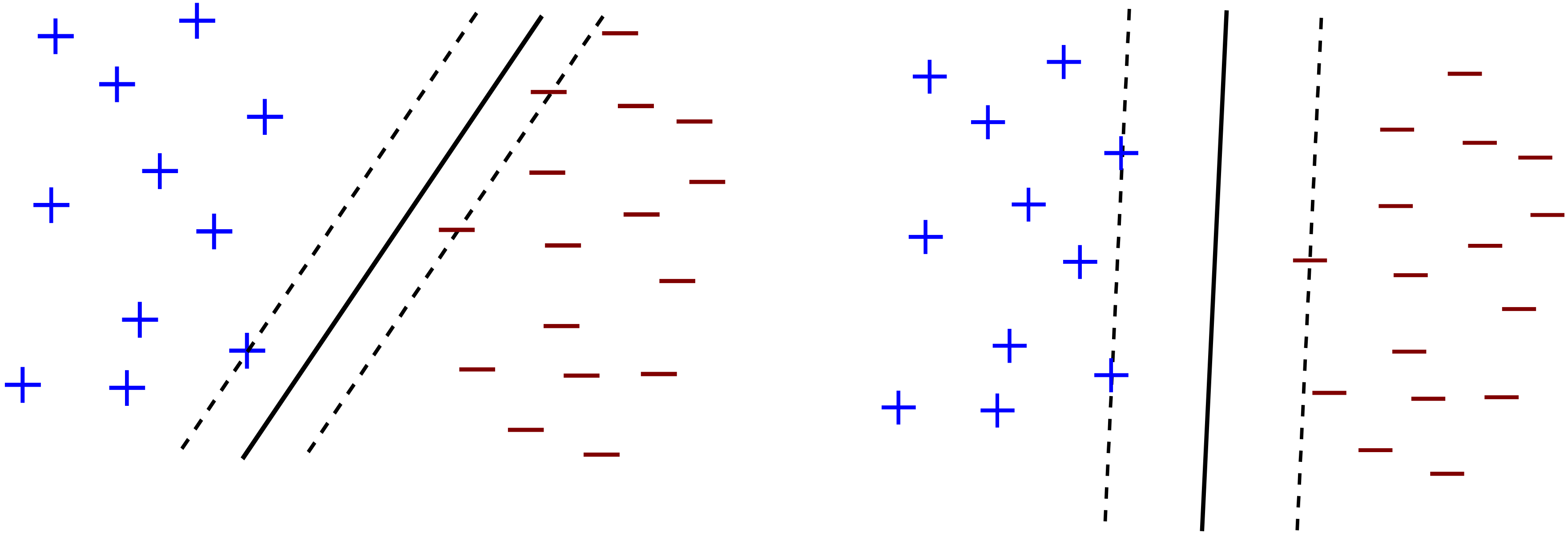
Data fitting

Constraint satisfaction

Solving equation: SVM

Support Vector Machines (SVM) defined by three incremental steps:

1. [Vapnik63]: linear classifier / separates the training examples with the **widest margin** => Optimal Hyperplane

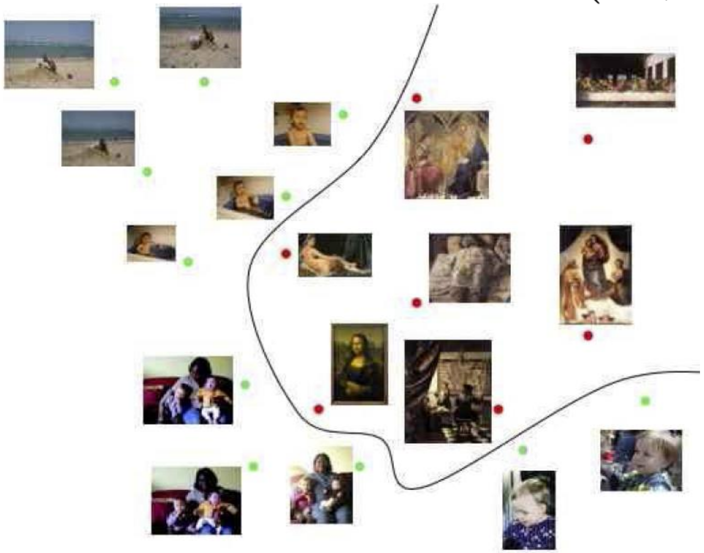


Solving equation: SVM

Support Vector Machines (SVM) defined by three incremental steps:

1. [Vapnik63]: linear classifier / separates the training examples with the widest margin =>Optimal Hyperplane
2. **[Guyon93] Optimal Hyperplane built in the feature space induced by a kernel function**

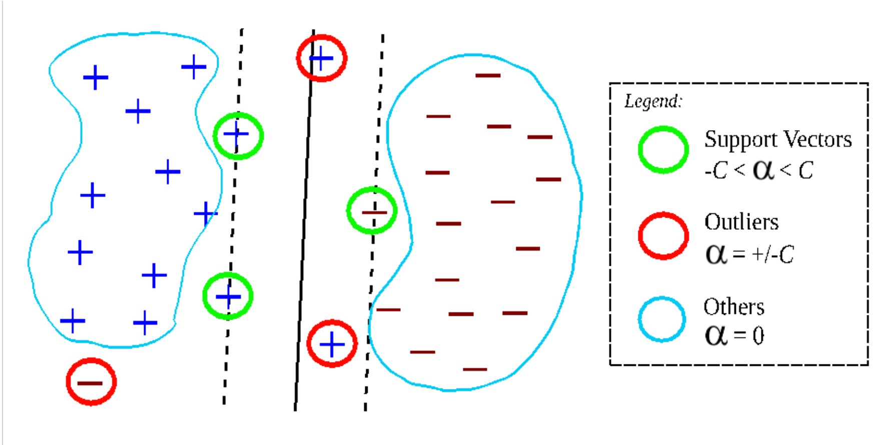
$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$



Solving equation: SVM

Support Vector Machines (SVM) defined by three incremental steps:

1. [Vapnik63]: linear classifier / separates the training examples with the widest margin =>Optimal Hyperplane
2. [Guyon93] Optimal Hyperplane built in the feature space induced by a kernel function
3. **[Cortes95] soft version: noisy problems addressed by allowing some examples to violate the margin constraint**



Classification pipeline

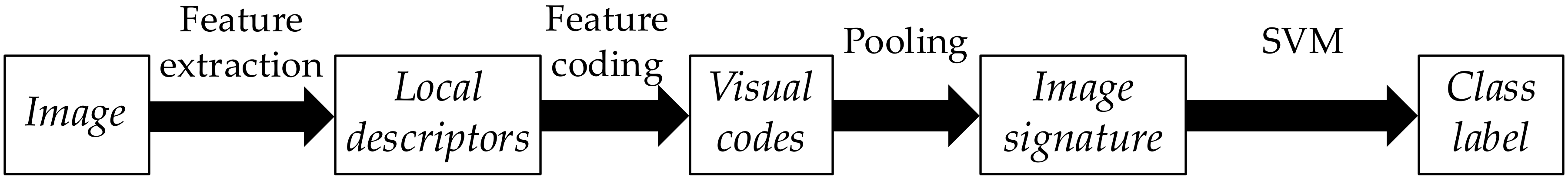


Image
Feature extraction
Local descriptors
Feature coding
Visual codes
Pooling
Image signature
SVM
Class label