# Tutorial on Image Clustering

by Joris Guérin

# Problem definition

# Inputs: Dataset of unlabelled images


# Objective: make groups within these images

➢ Difference with Image classification?

# Motivations

# Search web image databases

**Web-scale image clustering revisited**

Yannis Avrithis[†], Yannis Kalantidis[‡], Evangelos Anagnostopoulos[†], Ioannis Z. Emiris[†]
[†]University of Athens, [‡]Yahoo! Labs

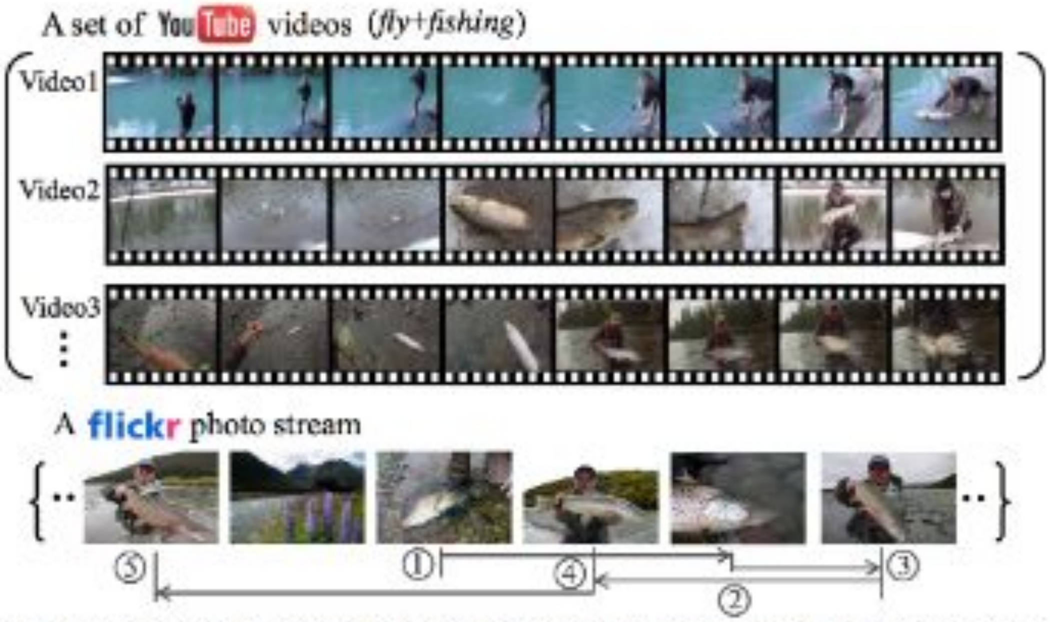
# Storyline reconstruction

**Joint Summarization of Large-scale Collections of Web Images and Videos for Storyline Reconstruction**

Gunhee Kim
Disney Research Pittsburgh
gunhee@cs.cmu.edu

Leonid Sigal
Disney Research Pittsburgh
lsigal@disneyresearch.com

Eric P. Xing
Carnegie Mellon University
epxing@cs.cmu.edu




(a) A set of videos help discover the sequential structure of a photo stream

# Concept discovery

**Shape Discovery from Unlabeled Image Collections**

Yong Jae Lee and Kristen Grauman
University of Texas at Austin

# Medical diagnoses

**Unsupervised Joint Mining of Deep Features and Image Labels for Large-scale Radiology Image Categorization and Scene Recognition**

Xiaosong Wang, Le Lu, *IEEE Senior Member*, Hoo-chang Shin, Lauren Kim,
Mohammadhadi Bagheri, Isabella Nogues, Jianhua Yao, Ronald M. Summers
Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center,
10 Center Drive, Bethesda, MD 20892

**Cluster #23**

| Word | Frequency |
| --- | --- |
| liver | 524 |
| abdomen | 337 |
| enhancement | 217 |
| mass | 198 |
| lesion | 168 |
| lobe | 161 |
| adenopathy | 119 |
| lesions | 109 |
| segment | 58 |
| bulky | 45 |

**Cluster #64**

| Word | Frequency |
| --- | --- |
| enhancement | 277 |
| cerebellar | 193 |
| lesion | 192 |
| lobe | 186 |
| flair | 173 |
| hemisphere | 155 |
| mass | 134 |
| abnormal | 119 |
| frontal | 115 |
| cerebellum | 113 |

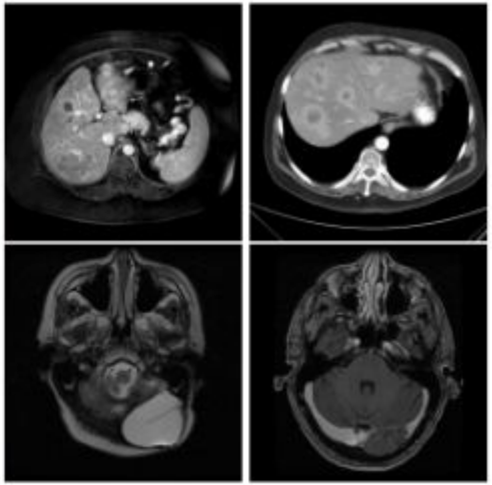
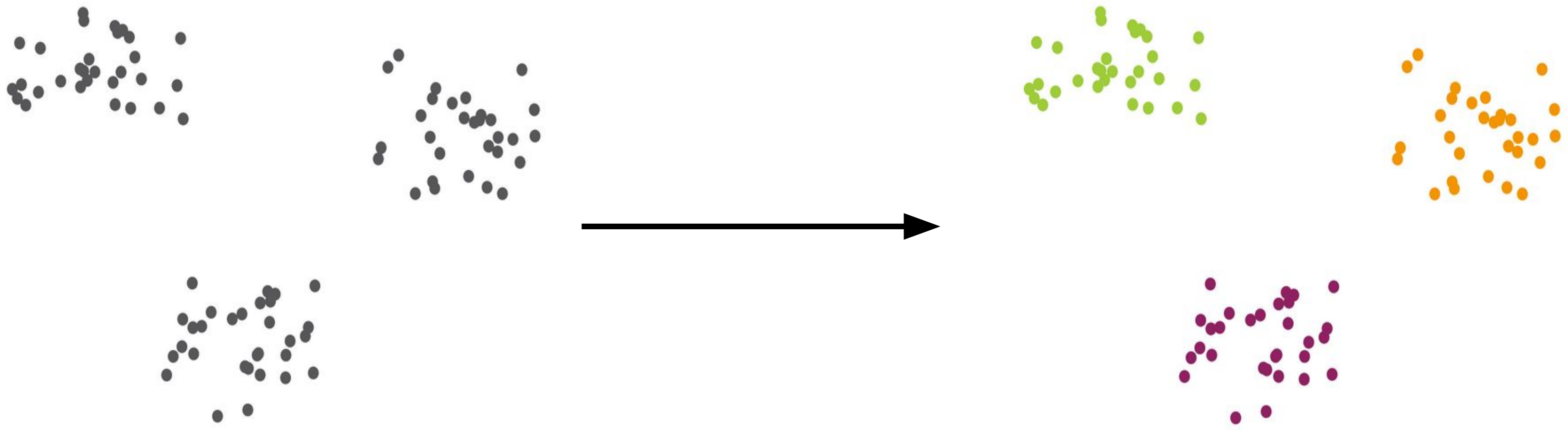


Figure 2. *Sample images of two unsupervisedly discovered image clusters with associated clinically semantic key words, containing (likely appeared) anatomies, pathologies , their attributes and imaging protocols or properties.*

# Intro to clustering

# Data clustering



# K-Means clustering

$$\underset{A,\, c}{\text{Minimize}} \quad \sum_{i=1}^{M} \sum_{k=1}^{K} a_{ik} \times d(x_i, c_k),$$

$$\text{subject to} \quad \sum_{k=1}^{K} a_{ik} = 1,\ \forall i \in \{1, ..., M\},$$

$$a_{ik} \in \{0, 1\},\ \forall i,\ \forall k.$$

With

- $c_k$ cluster centers,
- $a_{ik}$ membership binary variables,
- $d(.,.)$ distance metric used.

→ Alternating optimization : EM algorithm

# External metrics

Purity

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

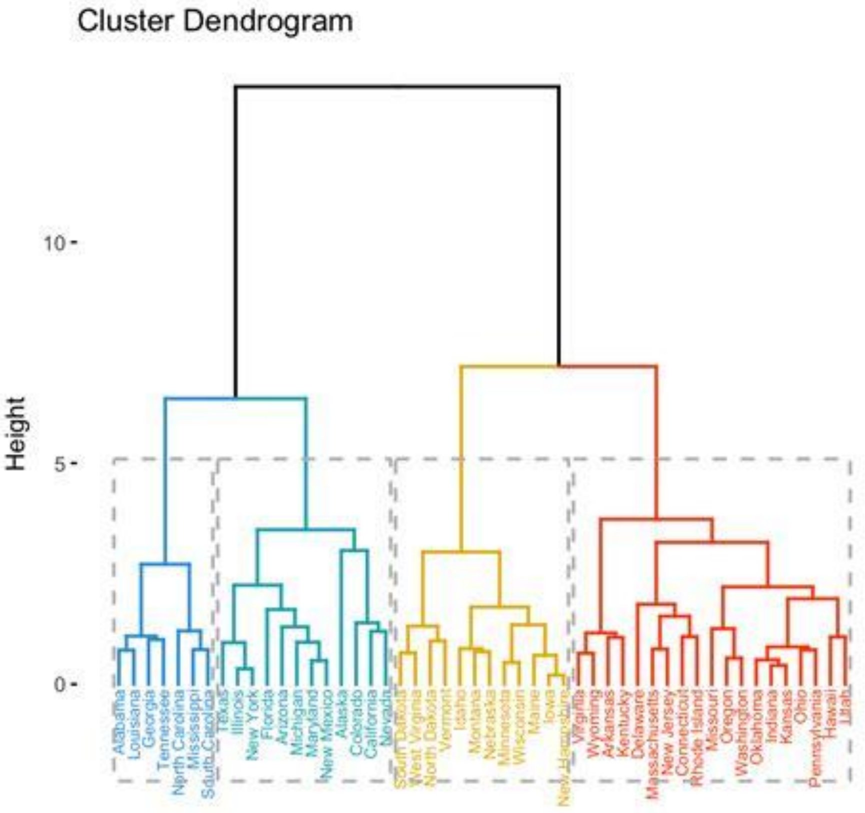
Normalized Mutual Information

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2}$$

Fowlkes-Mallows index

$$FM = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$$

# Agglomerative clustering

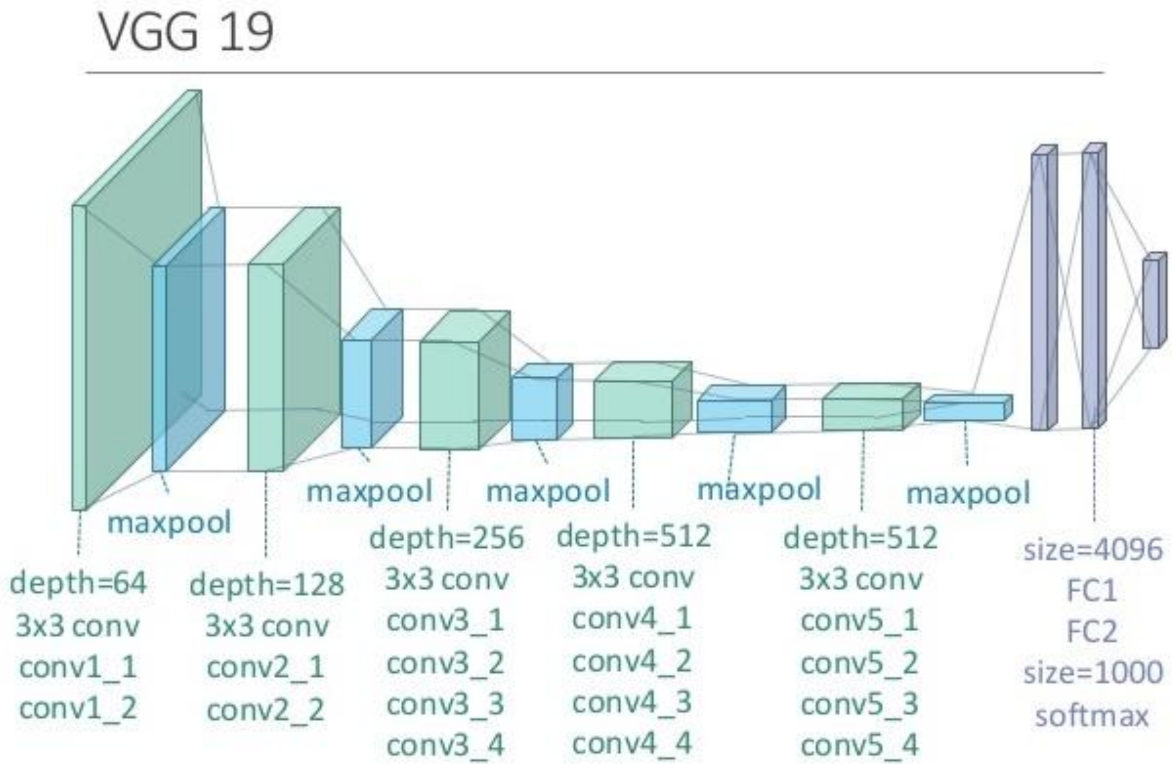


# Clustering on raw images

# Clustering on standard CV features

# Clustering on deep CNN features

# CNN feature extractor



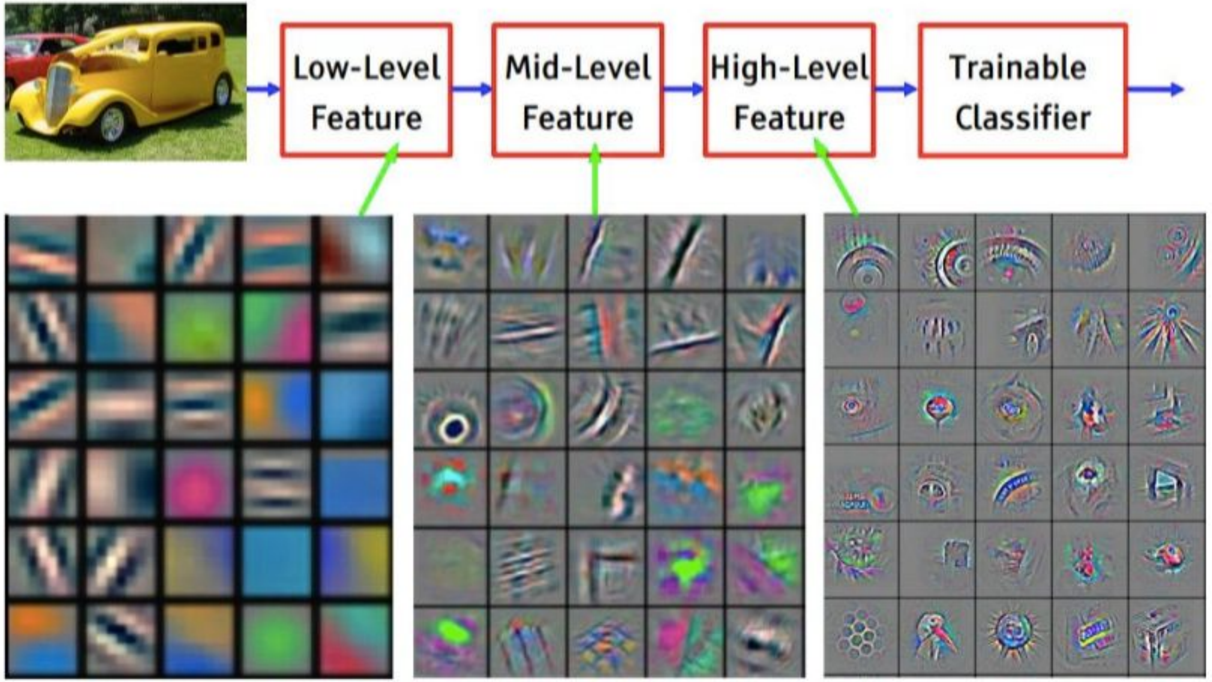

- Complex model
- Solve many computer vision tasks recently
- Trained in a supervised way

# CNN feature extractor - pre-training

**ImageNet**

- Several million images
- 1000 categories

# CNN feature extractor - image processing




Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

**ImageNet**

- Several million images
- 1000 categories

# For next time (21/03/2019)

- Choose an image dataset.

- Repeat the experiments on you dataset with various feature extractors
    - 10 available in Keras: https://github.com/keras-team/keras-applications/tree/master/keras_applications

- You can try different architectures, different layers, different clustering algorithms ...
    - https://scikit-learn.org/stable/modules/clustering.html

- Record and send all your results and send them to me 24 hours before the class.

# For next time (21/03/2019)

- Code available at :
  https://github.com/jorisguerin/tutorial_IC
- Need GPU?
  https://colab.research.google.com/
- Send your results to:
  jorisguerin.research@gmail.com