

Image Dataset Curation and Art Recommendation

Antonio Rueda-Toicen
AI Engineer

KI Service Zentrum, Hasso Plattner Institute
Berlin Computer Vision Group
December 2023

KI Service
Zentrum
by Hasso-Plattner-Institut

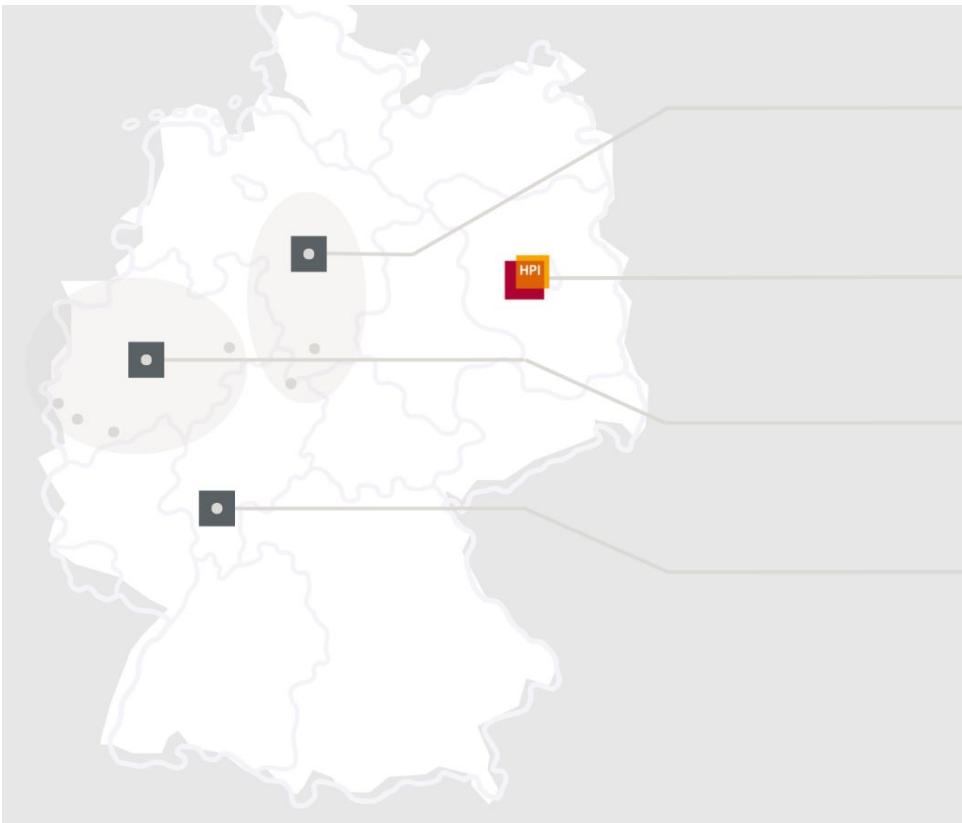


SPONSORED BY THE



Federal Ministry
of Education
and Research

<https://hpi.de/kisz/home.html>



KISSKI

Hannover | Göttingen | Kassel
Sensitive and critical infrastructures
Focus: Medicine & Energy

KISZ BB

Berlin | Brandenburg
Educational and advisory services
Use of AI in business and society

West AI

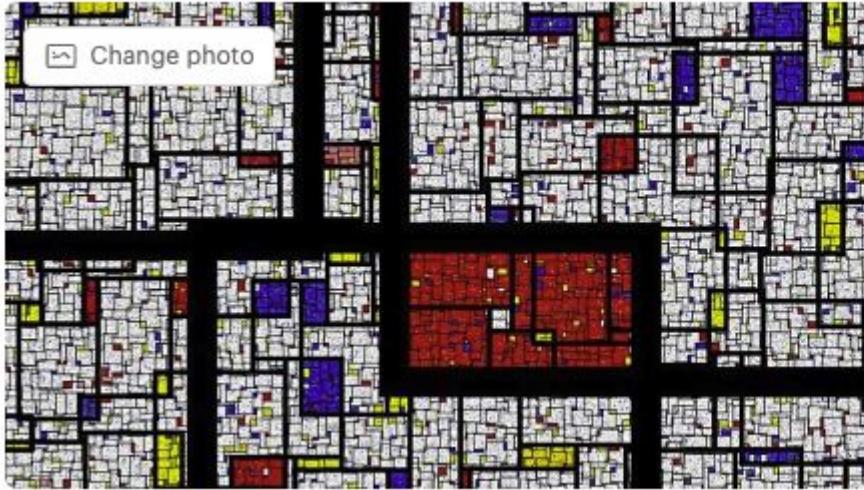
Dortmund | Bonn | Jülich | Aachen | Paderborn
Large and transferable AI models

hessian AI

Service Center

Darmstadt
Explainability, generalizability
and contextual adaptation

<https://hpi.de/kisz/home.html>



Berlin Computer Vision Group

★★★★★ (81)

Berlin, Germany

848 members · Public group

Organized by Antonio Rueda Toicen

Edit group info

Share: [n](#) [f](#) [t](#) [in](#) [e](#)

<https://www.meetup.com/berlin-computer-vision-group/>

Agenda

- Brief intro and learning objectives
- Use cases for image similarity
- Scraping images from Google Images or Bing
- Setting up Google Colab and Drive
- Understanding image embeddings
- Using pre-trained networks for image similarity
- K-nearest neighbors for dataset exploration
- Image clustering with K-medoids
- Visualizing embeddings with TensorBoard
- Transfer learning for image classification
- Review questions and discussion

What we expect you to have

- Some Python knowledge
- Curiosity :)

Learning objectives

At the end of this workshop you will be able to:

- Describe use cases for image similarity in dataset curation
- Scrape images from Google Images or Bing
- Generate embeddings for images using a pretrained neural network (Resnet)
- Compare images using cosine similarity
- Cluster images using k-medoids
- Visualize embeddings in 3D using Tensorboard
- Classify images using a pre-trained Resnet

How are we going to do this workshop

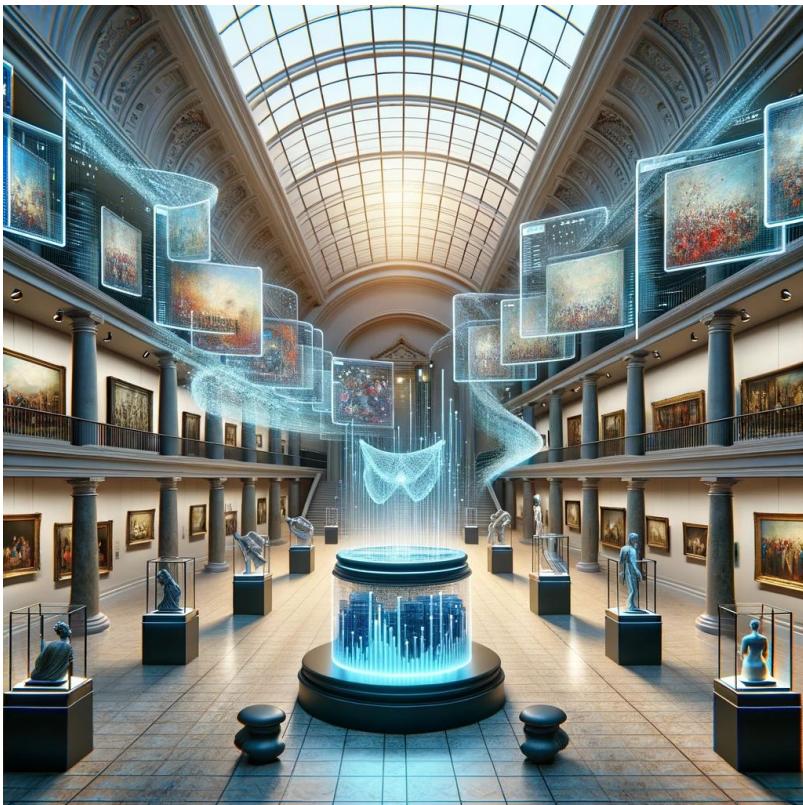
- We pick an artist
- We scrape their images from Google Images or Bing through Javascript (just copy and paste the script)
- We **type** most of the Python code from the tutorial notebooks in Google Colab

Repository: <https://github.com/andandandand/image-dataset-curation>

What you need

- A Google user account
- A Google Drive account with enough free space
- Google Chrome or Firefox

What is dataset curation?



- We want to make data **accurate** and **relevant**
- We need to clean, deduplicate, and label
- This is similar to what museum curators do

How similar are these two images?



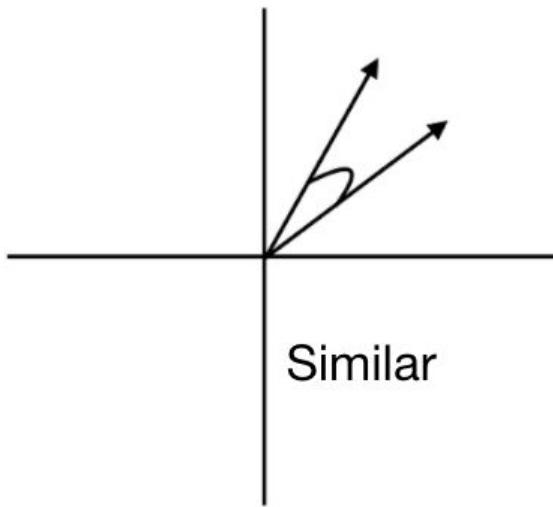
Cosine similarity = 0.65

How similar are these two images?

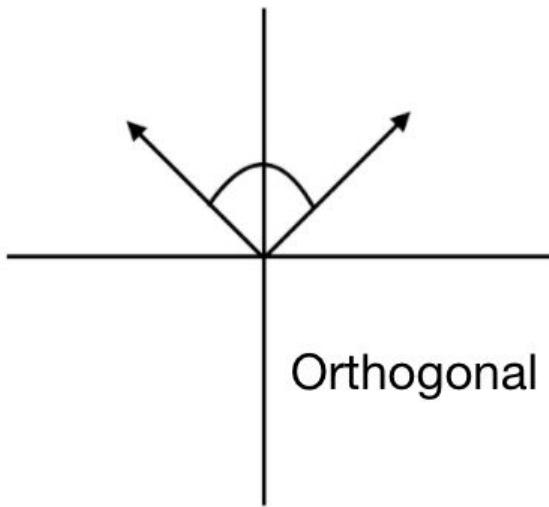


Cosine similarity = 0.85

Cosine Similarity



Similar



Orthogonal

13

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Use Case: Google's reverse image search

Google

Find image source

Search Text Translate

Related search

The Tower of Babel

Wikimedia
File:Pieter Bruegel d. Ä.
075.jpg - Wikimedia...

See exact matches >

Genially
La Tour de Babel par Lila et Suzanne

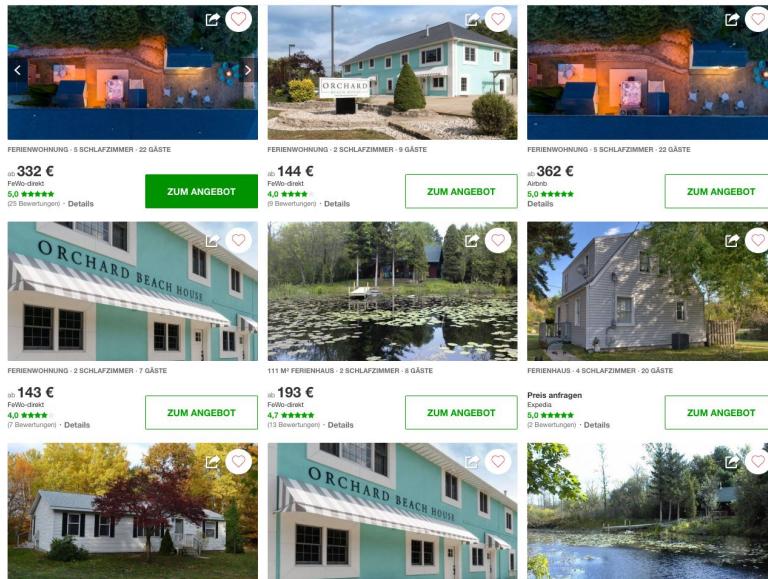
HistoryExtra
Babylon: What Happened To The...

Christian Bible Refe...
The Tower of Babel

<https://images.google.com/>

Use case: primary image matching at Hometogo

- Inventory understanding (500 million images)
- Providing the best deals to users (sample use case: strike prices)



3-STAR HOTEL · DOWNTOWN GRAND LAS VEGAS

\$30 ~~\$33~~
Travelocity
4.2 ★★★★★ · Details

VIEW DEAL

Use case: primary image matching at Hometogo



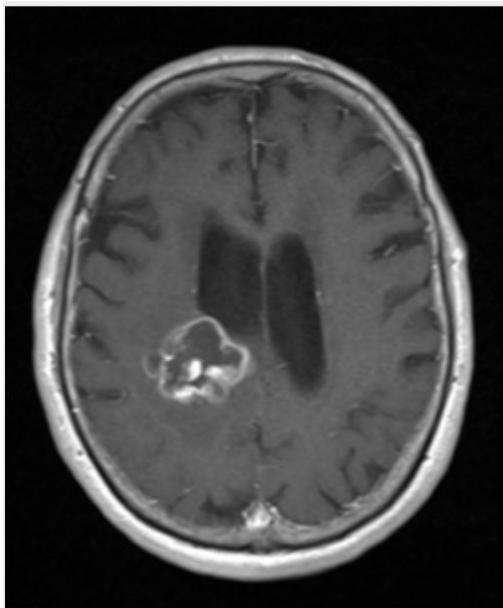
Cosine similarity = 0.65

Use case: primary image matching at Hometogo

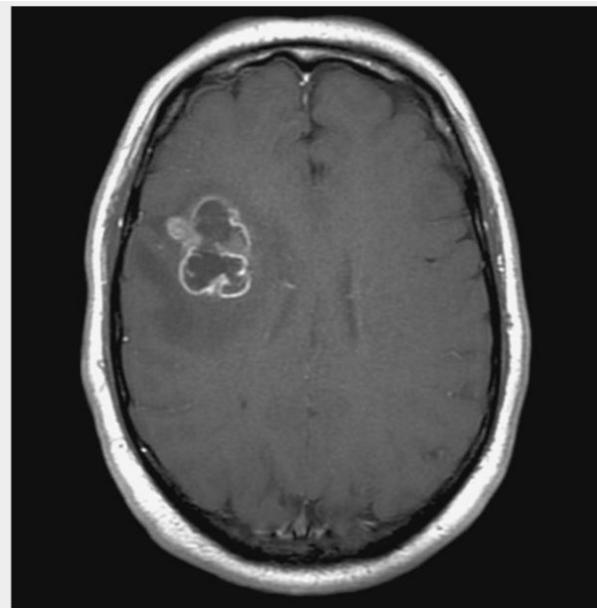


Cosine similarity = 0.99

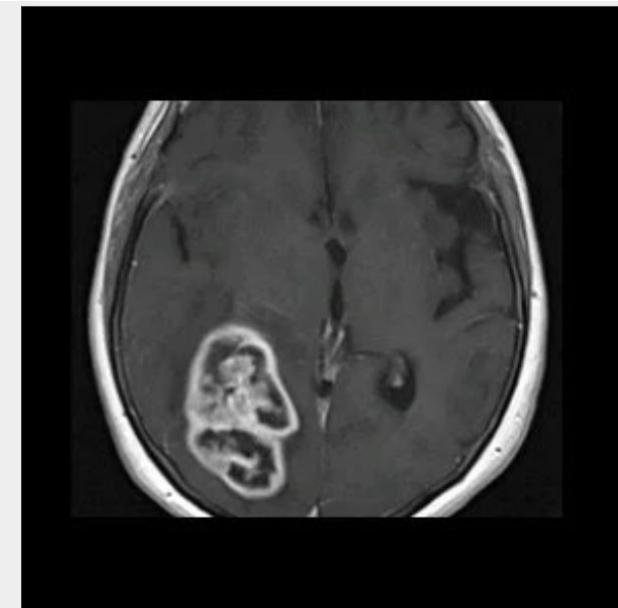
Use case: differential diagnosis



Glioma grade I



Glioma grade I



Glioma grade III

Use case: bike trip routing

Cluster 0: Safe bike lanes



Cluster 1: Unsafe bike lanes



Use case: bike trip routing



github.com/graumannm/Berlin_Bike_CV

Use case: cleaning image datasets



<https://github.com/cleanlab/cleanlab>

Art Recommendation system

This is the repository of a portfolio project at DSR. This project aims to identify similar images using pre-trained computer vision networks. For an explanation of the technology see the [technology section](#).

Contributors

- Catarina Ferreira
- Gargi Maheshwari

<https://github.com/gargimaheshwari/Wikiart-similar-art>

Motivation: enhancing Wikiart's recommendations



M.C. Escher

Maurits Cornelis Escher

Born: June 17, 1898; Leeuwarden, Netherlands [i](#)

Died: March 27, 1972; Laren, Netherlands [i](#)

Nationality: [Dutch](#)

Art Movement: [Surrealism](#), [Op Art](#)

Genre: [tessellation](#)

Field: [printmaking](#), [illustration](#), [drawing](#)

Wikipedia: en.wikipedia.org/wiki/M._C._Escher ↗

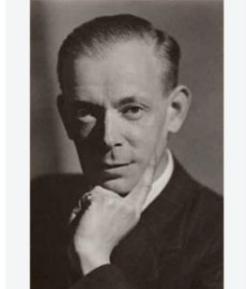
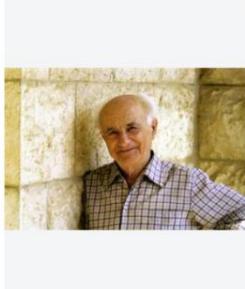
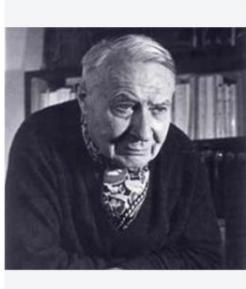


ORDER OIL PAINTING
REPRODUCTION

<https://www.wikiart.org/en/m-c-escher>

Motivation: improving Wikiart's art recommendation

RELATED ARTISTS i



**Carlos Orozco
Romero**

1896 - 1984

Andre Masson

1896 - 1987

Mordecai Ardon

1896 - 1992

Jock Macdonald

1897 - 1960

Victor Servranckx

1897 - 1965

Paul Delvaux

1897 - 1994

Lorser Feitelson

1898 - 1978

<https://www.wikiart.org/en/m-c-escher>

Motivation: improving Wikiart's art recommendation

ARTISTS BY ART MOVEMENT

Art historians employ a number of ways to group world arts into systems of classification. They subdivide the continuous flow of artworks through time and space into groupings. These groupings are defined by the perception that the artworks within them share a single quality or a set of qualities that are significant. Significant qualities reflect a specific approach of an artist; they can include the formal, stylistic, iconographic, thematic, or other aspects of art. The definition of a grouping reflects judgments about the nature of meaningful connections between artworks, and between art and its larger context. Western arts are usually structured by art movements, using mostly cultural and aesthetic criteria, while Eastern arts are subdivided into periods according to political-dynastic markers.

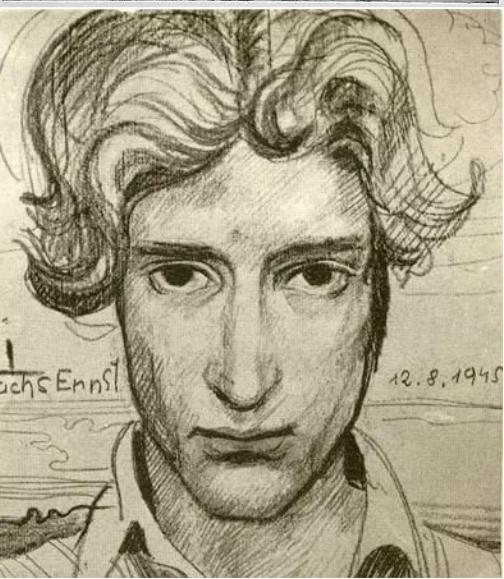
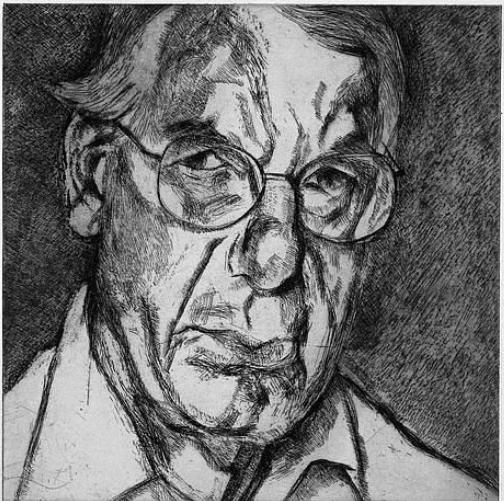
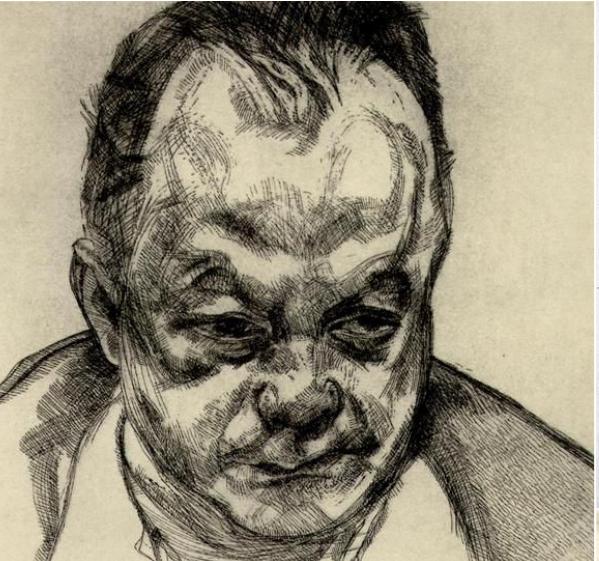
by time

by name

by count

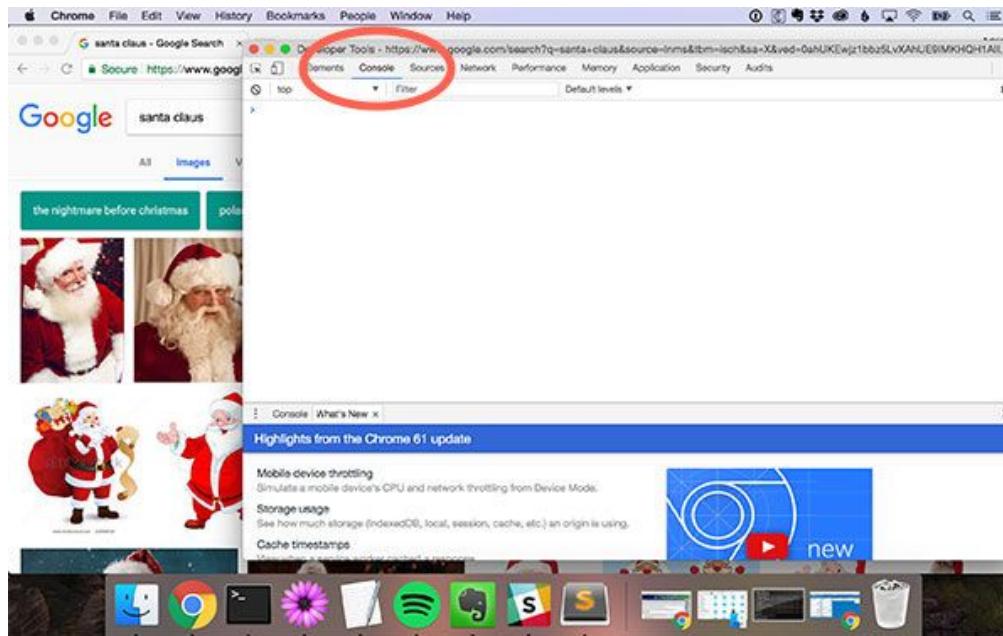






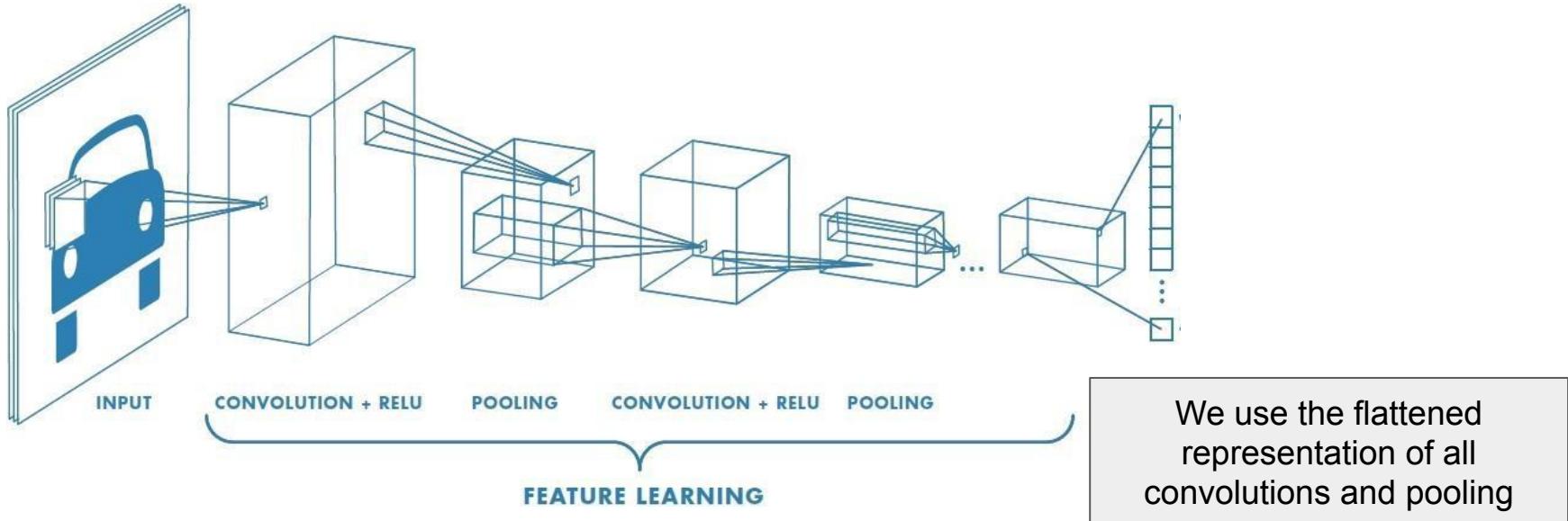


Scraping a dataset from Google Images



[PylImageSearch guide](#)

Image Embeddings from Convolutional Networks



The ImageNet dataset



14,197,122 images, 21841 synsets indexed

[Home](#) [Download](#) [Challenges](#) [About](#)

Not logged in. [Login](#) | [Signup](#)

ImageNet is an image database organized according to the **WordNet** hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. The project has been **instrumental** in advancing computer vision and deep learning research. The data is available for free to researchers for non-commercial use.

<https://www.image-net.org/>

Image neighborhoods

1.0



0.73



0.73



0.72



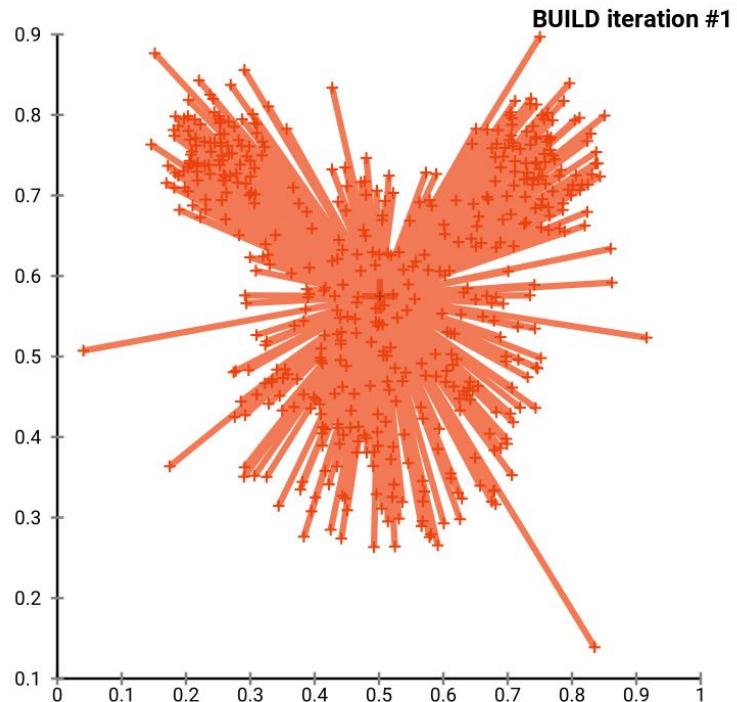
0.7



0.7



K-medoids clustering



Visualizing embeddings in TensorBoard

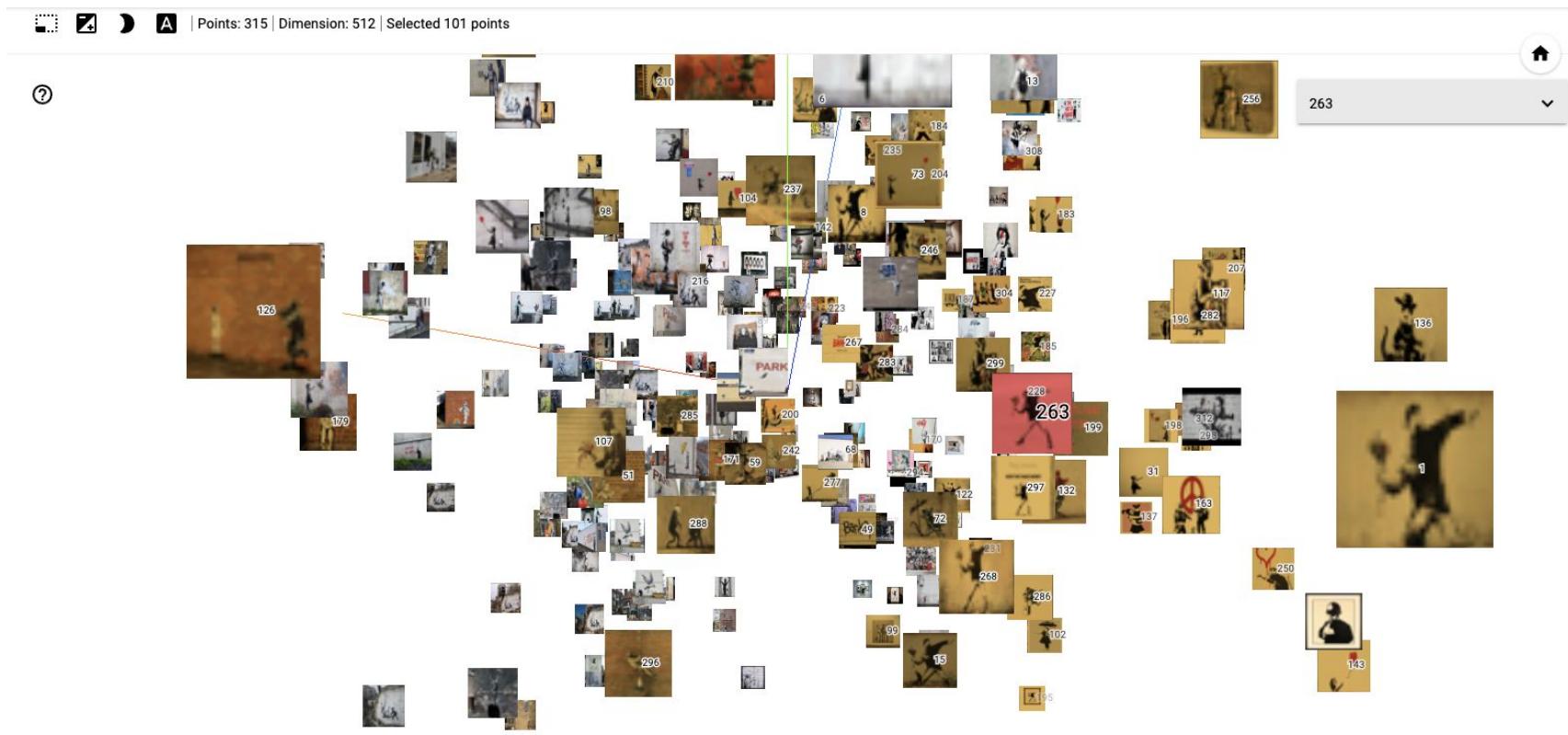
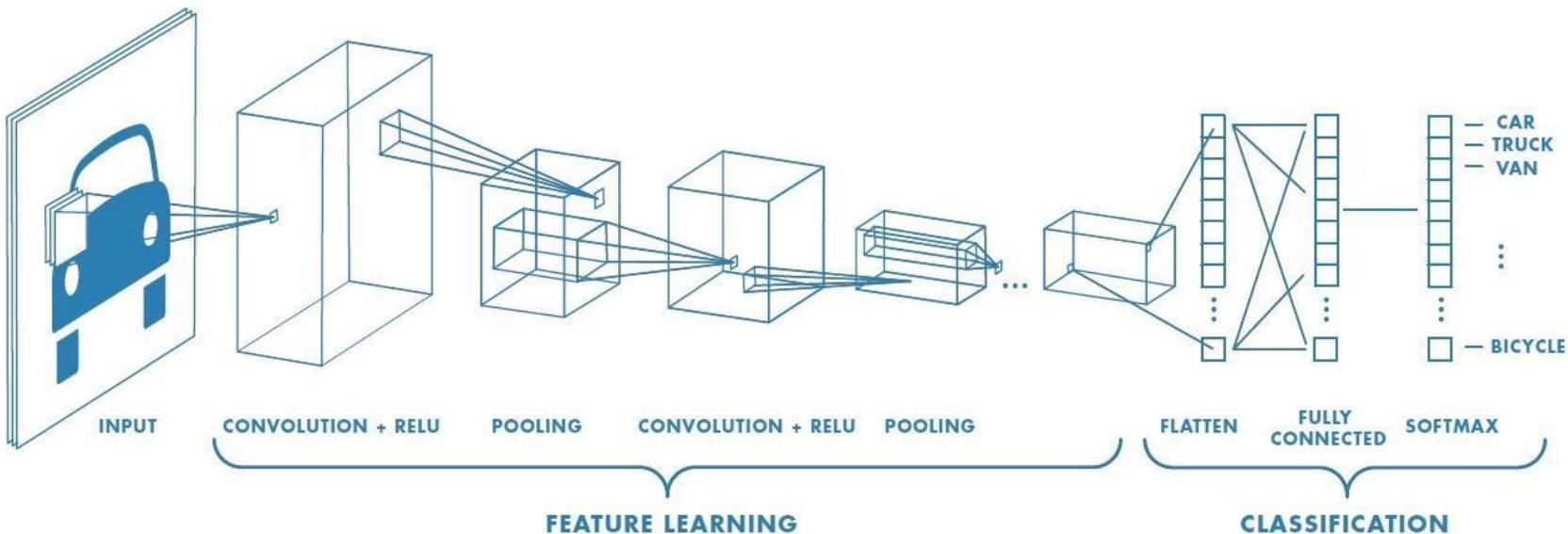


Image classification through transfer learning



Review questions

- How can we access an image array from its file's url?
- What is an embedding?
- What is maximum possible value of cosine similarity between two vectors?
- What's the cosine similarity between two orthogonal vectors?
- How does the k-medoids algorithm work?
- What is transfer learning?
- What is Imagenet?

Join us for the next workshops!

- [AI Service Center - Berlin Brandenburg](#)
- [Berlin Computer Vision Group](#)

Next topics:

- Fine tuning models for image classification
- Image analysis with CLIP
- Meta's Segment Anything
- Segmentation and object detection with Detectron
- Model deployment with FastAPI and Docker