Fun with content-based image retrieval (CBIR) with neural networks

stefan.otte@gmail.com

sotte.github.io


DR. SHELDON COOPER'S
FLAGS

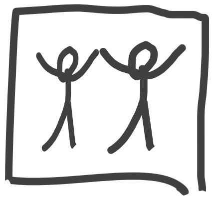
before we begin...

"*Content-based image retrieval (CBIR), [...] is the application of computer vision techniques to the image retrieval problem, that is,* ***the problem of searching for digital images in large databases***."

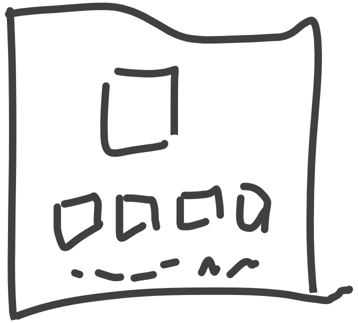
-- wikipedia.org

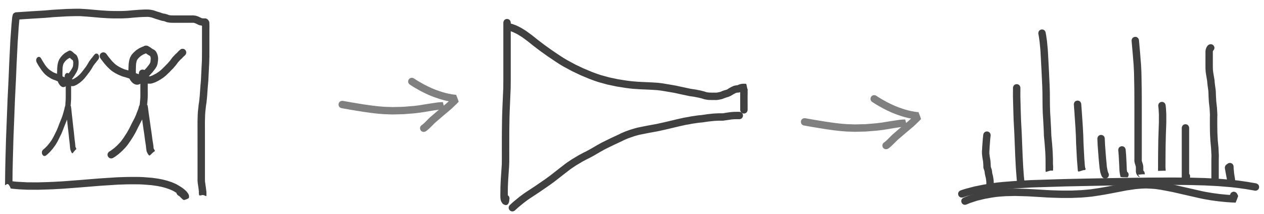
demo

Setup

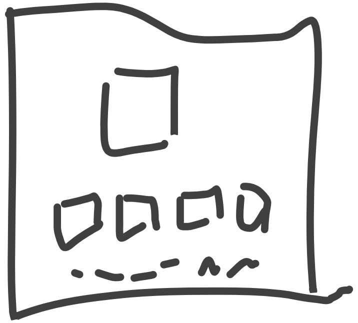


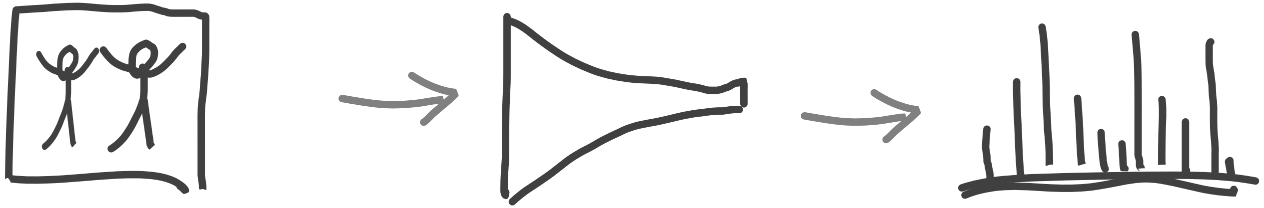
API/GUI



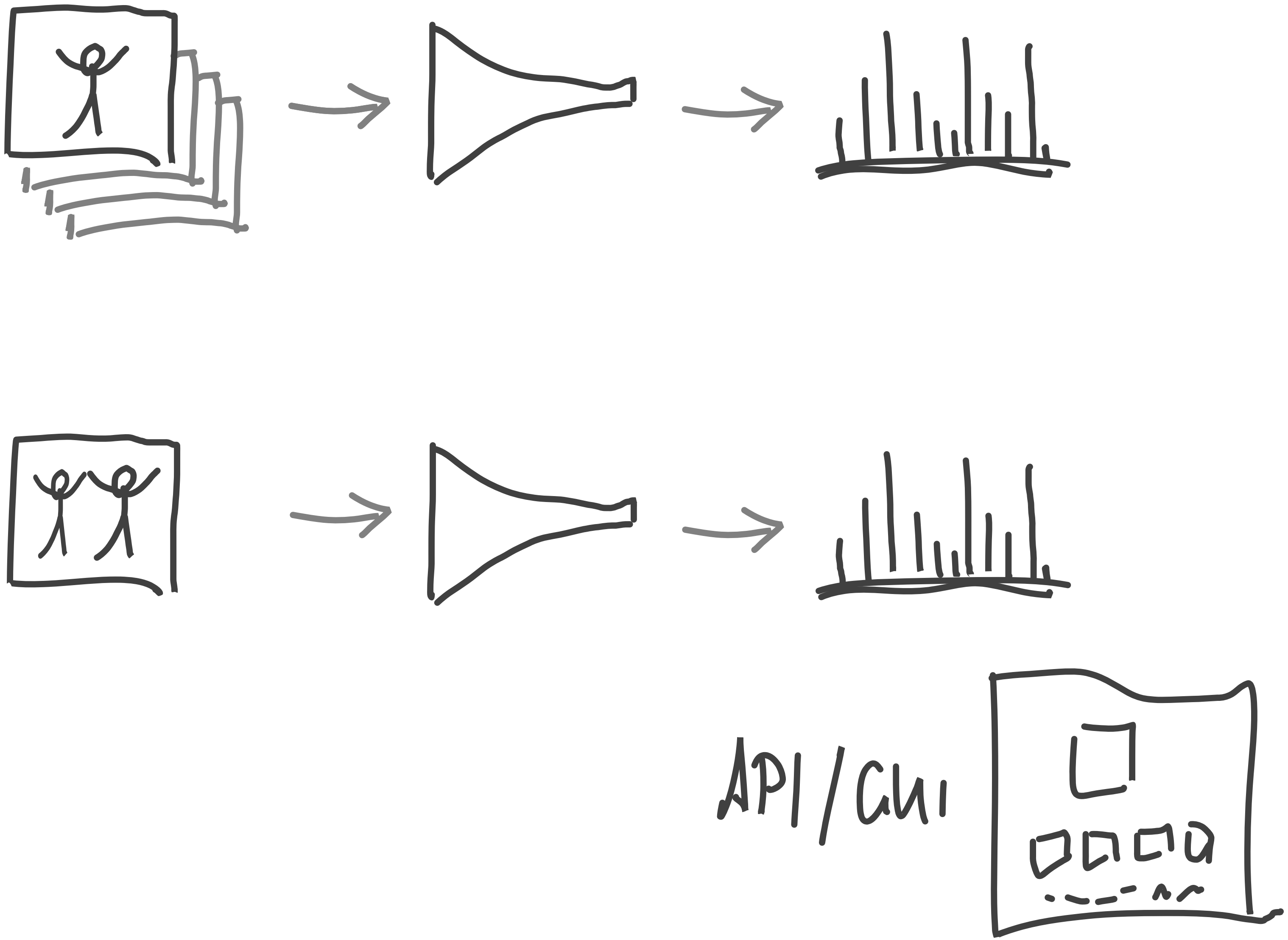


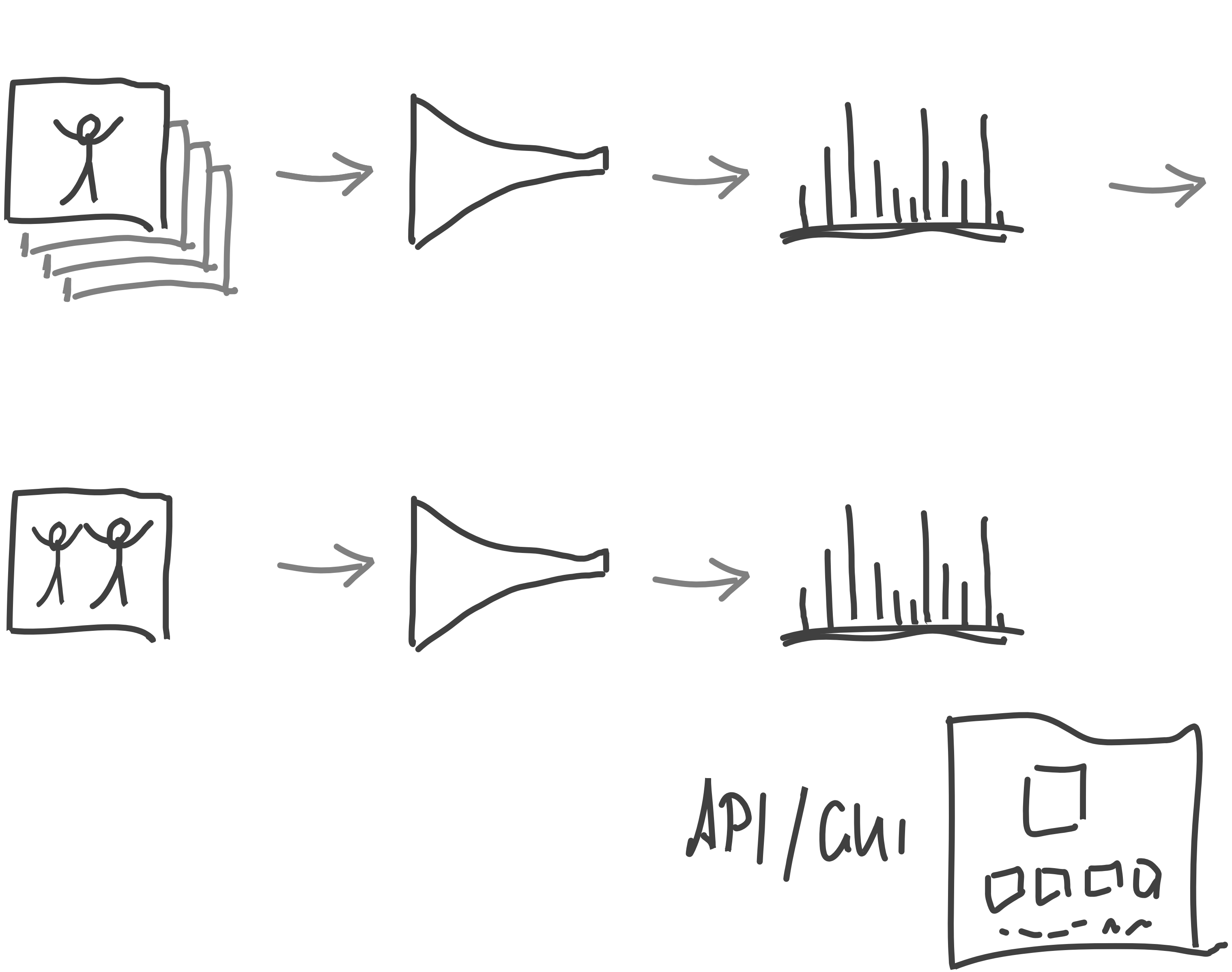
API/GUI

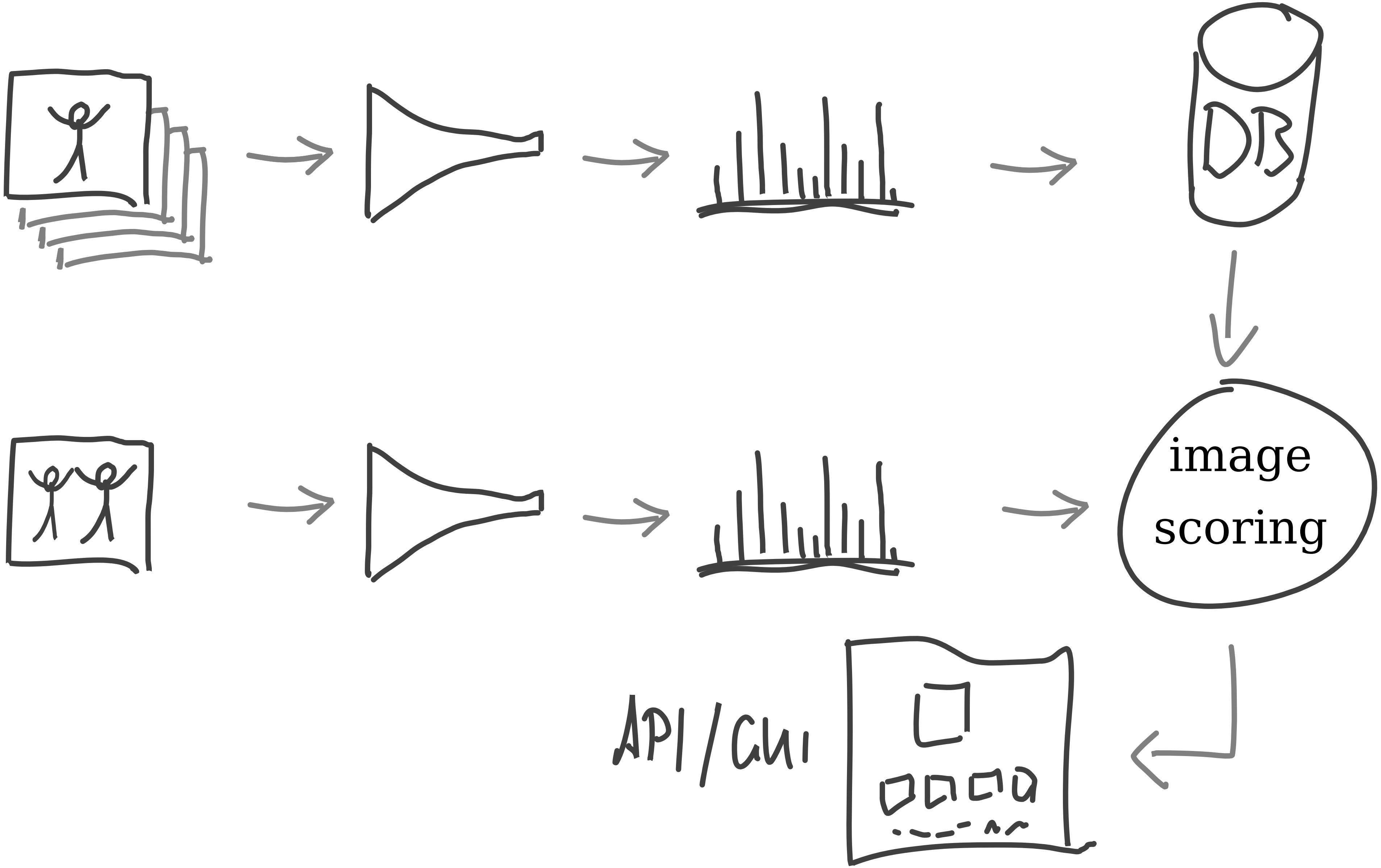


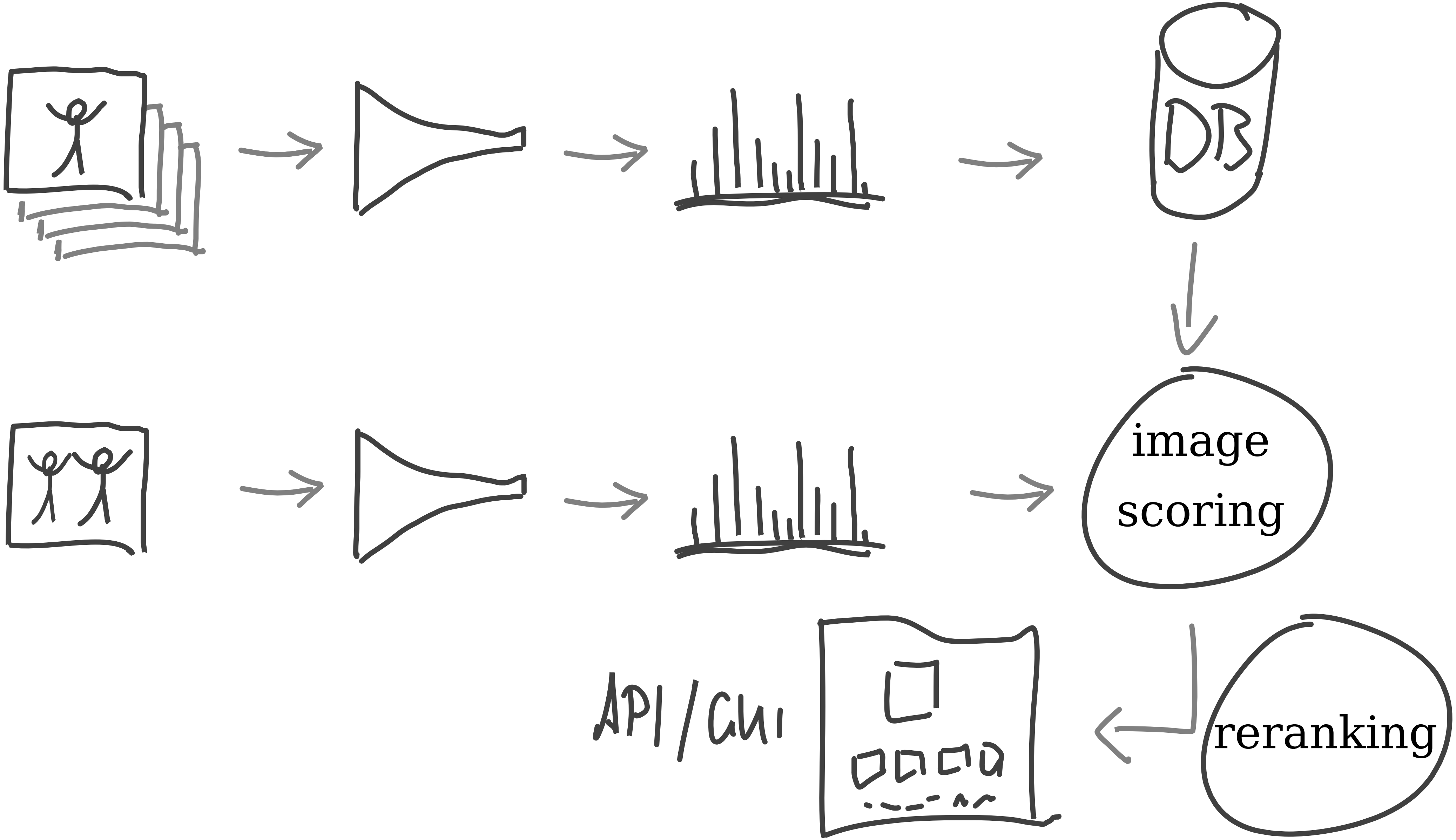


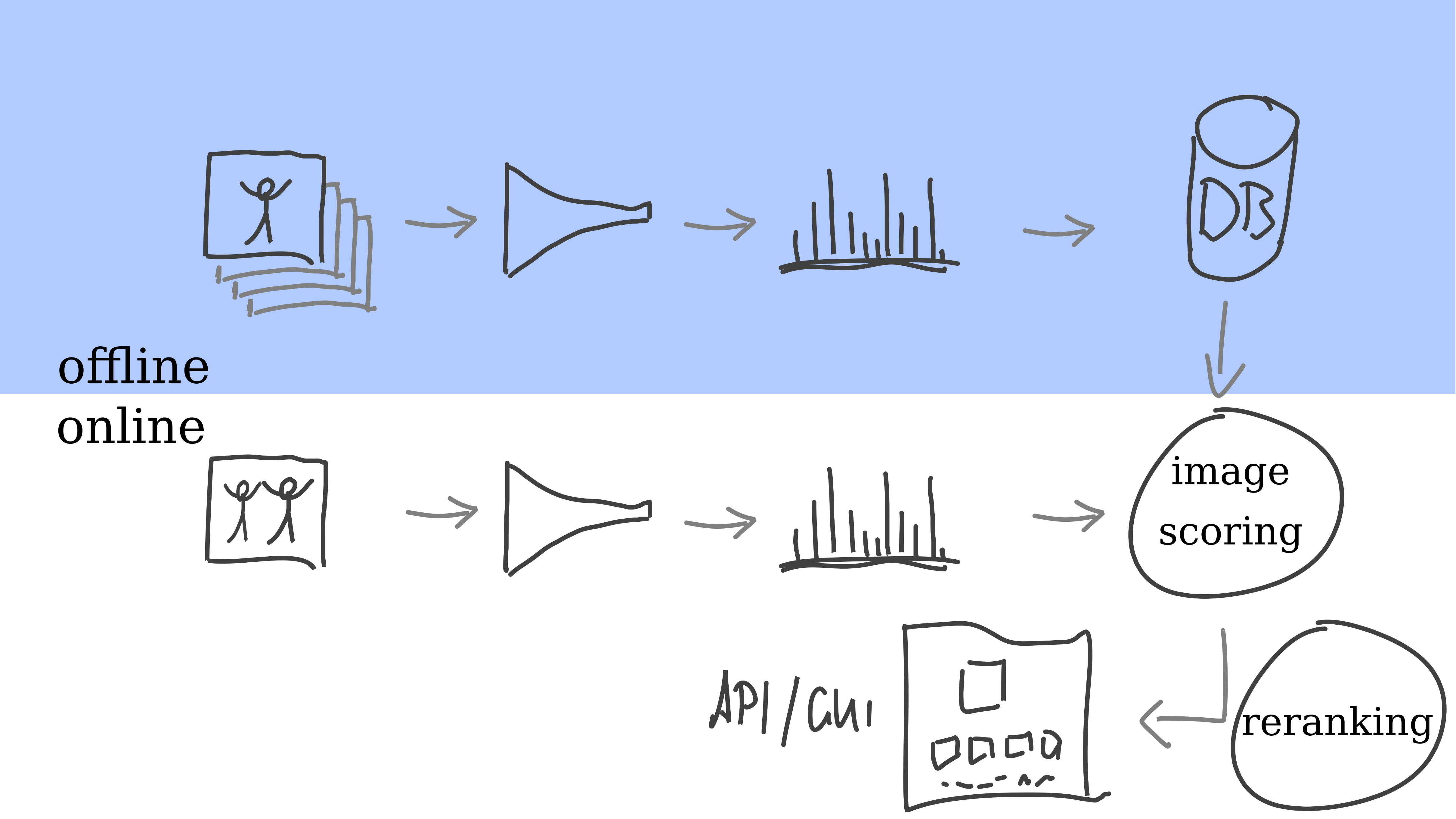

API/GUI


API/GUI


DB
API/GUI


DB
image scoring
API/GUI


DB
image scoring
reranking
API/GUI


DB
offline
online
image scoring
reranking
API/GUI

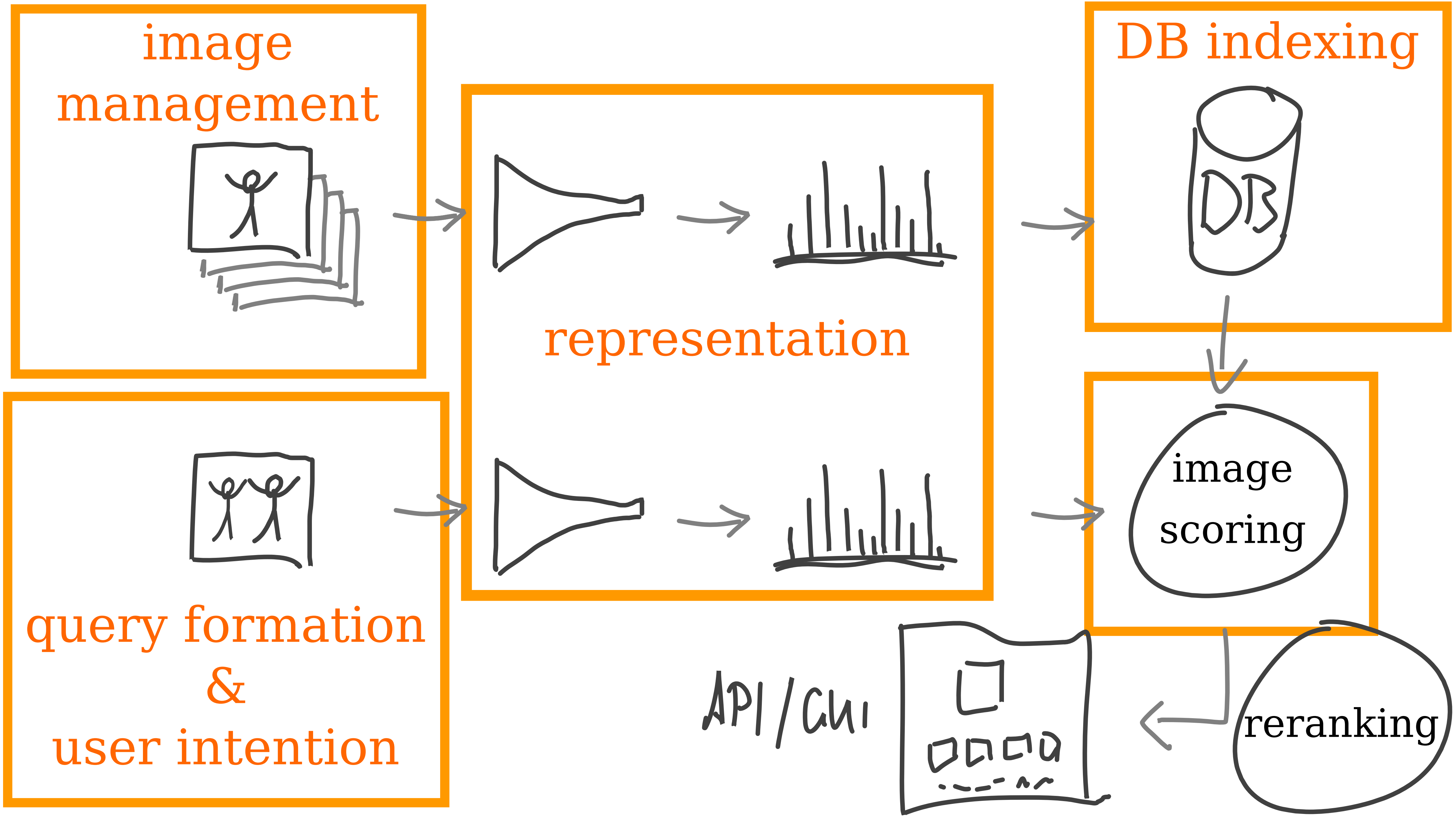


image management
representation
DB indexing
DB
query formation
&
user intention
image scoring
reranking
API/GUI

```python
# Offline
feature_extractor = resnet18(pretrained=True)
reference_features = torch.cat(
    [feature_extractor(load_image(p)) for p in image_paths]
)
```

```
# Offline
feature_extractor = resnet18(pretrained=True)
reference_features = torch.cat(
    [feature_extractor(load_image(p)) for p in image_paths]
)

# Online
query_image = load_image("sample_image.jpg")
query_feature = feature_extractor(query_image)

sim = F.cosine_similarity(query_feature, reference_features)
sorted_sim, sorted_index = torch.topk(sim, dim=top_k)
```

folding_chair 98.0% | rocking_chair 1.3% | mantis 0.2% |

Metrics

- mean average precision (mAP)
- normalized discounted cumulative gain (NDCG)

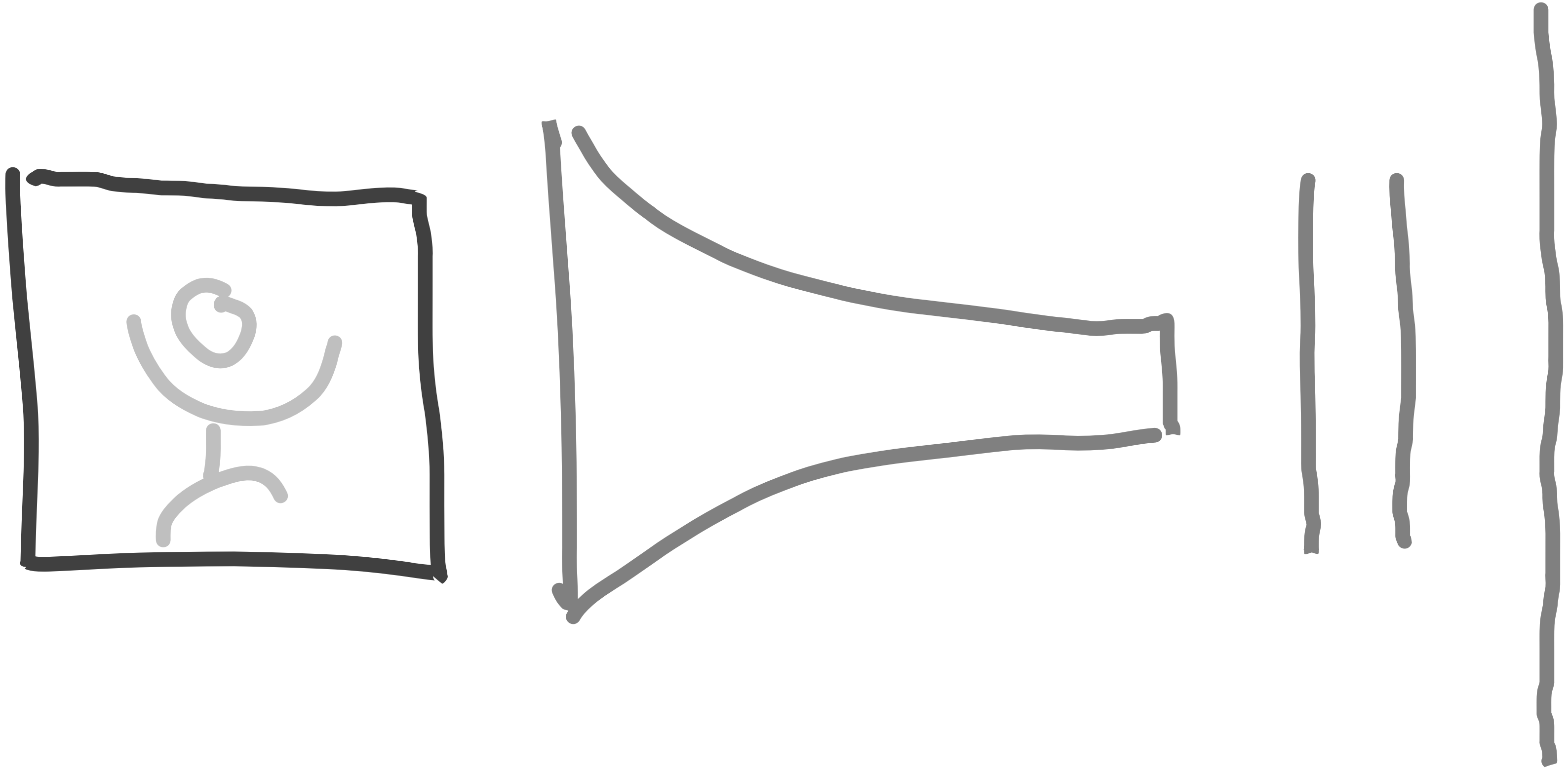
Representations

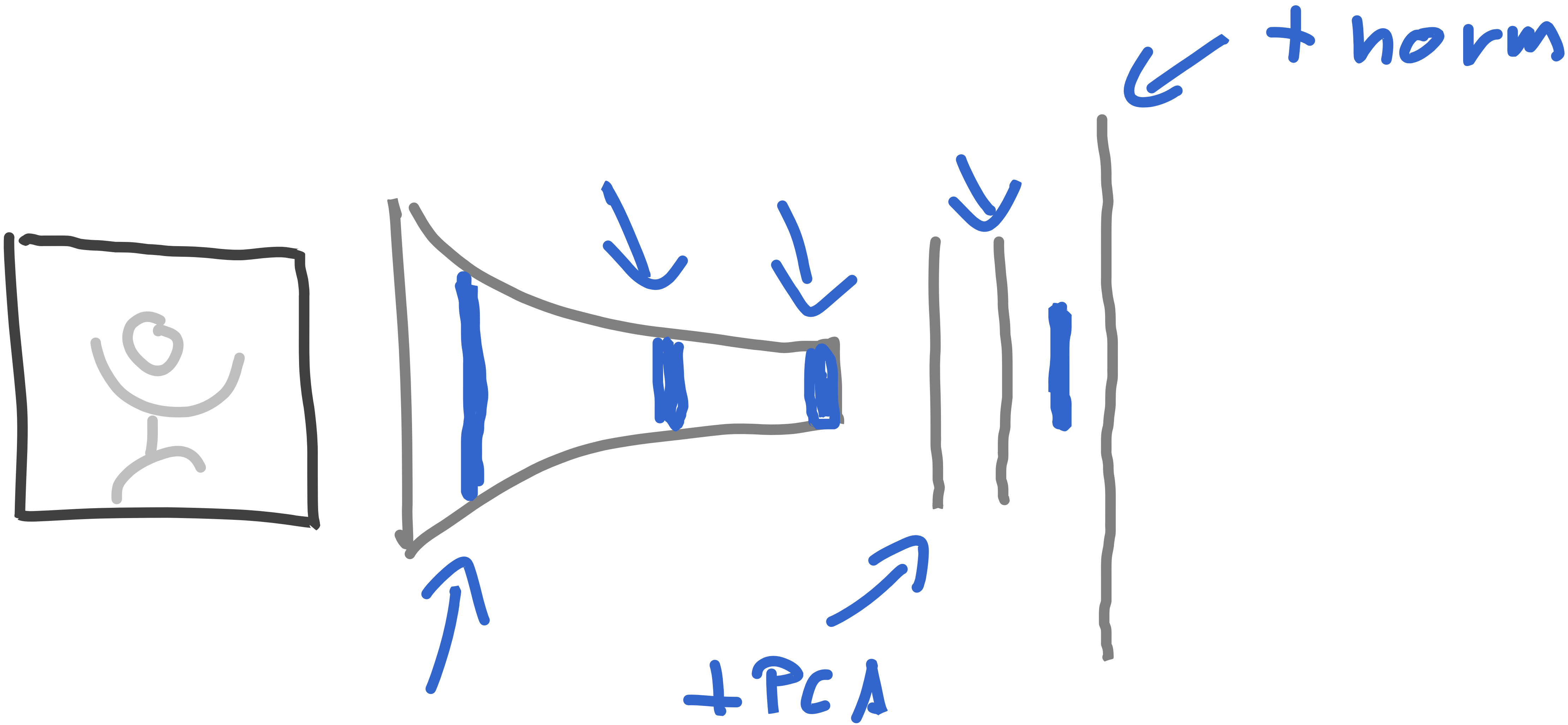
refine with
auxiliary task

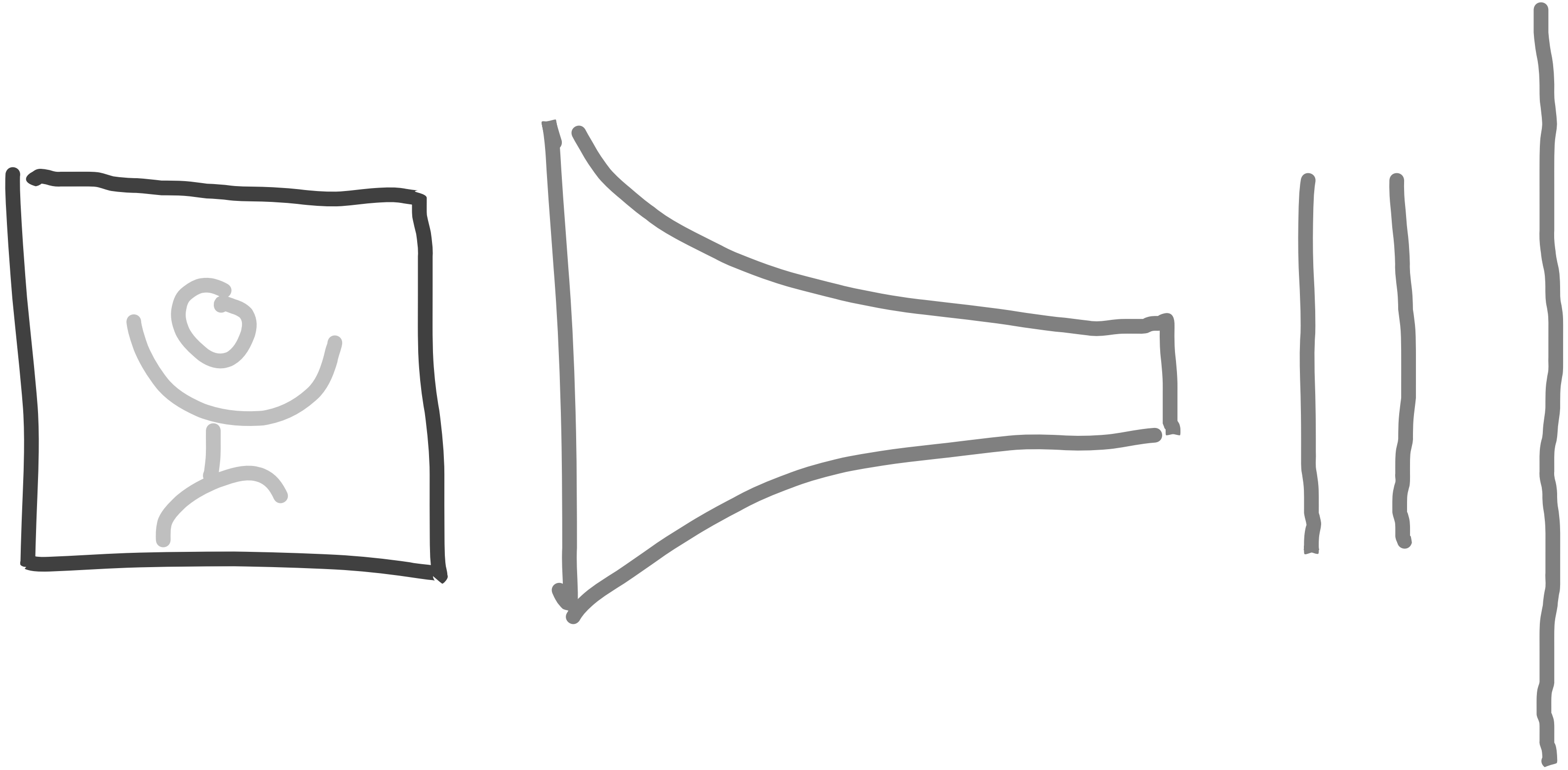
refine with
similarity learning

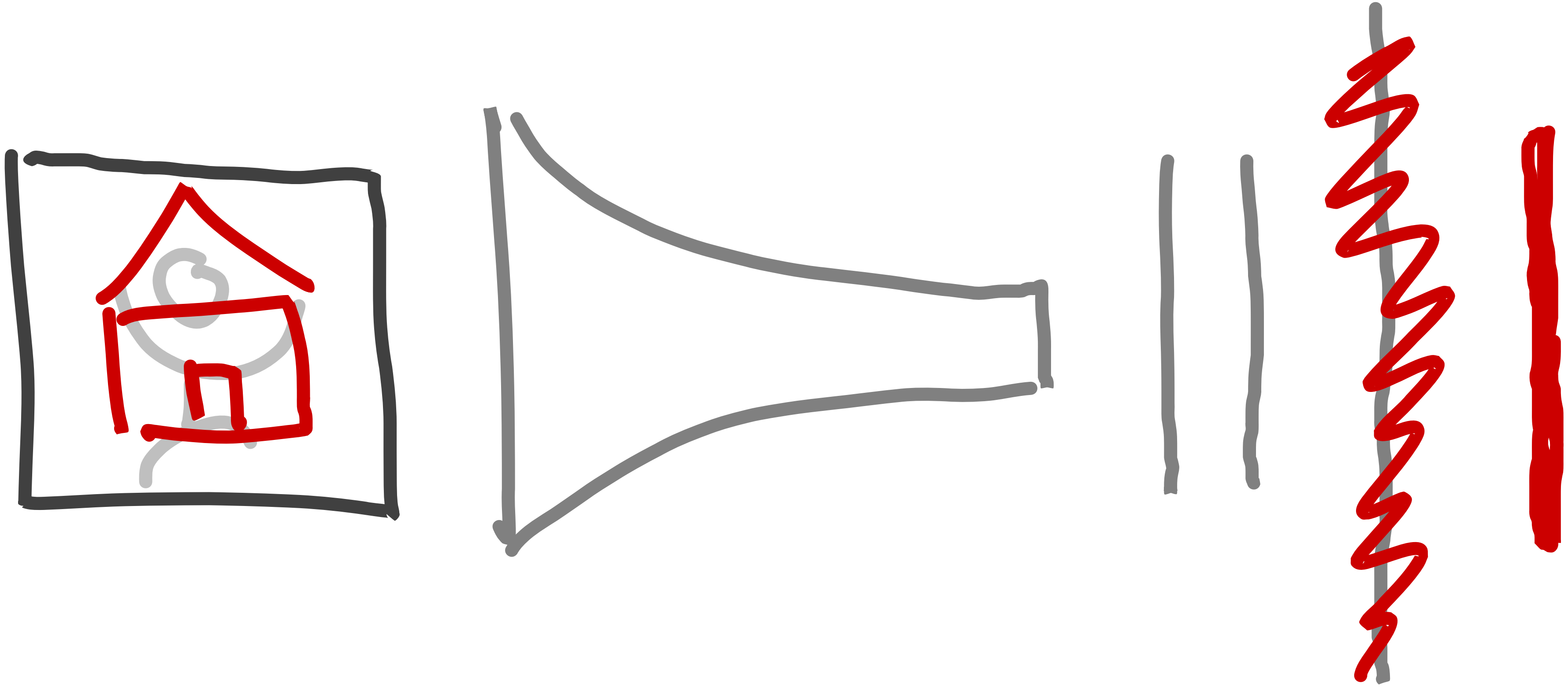
Representations

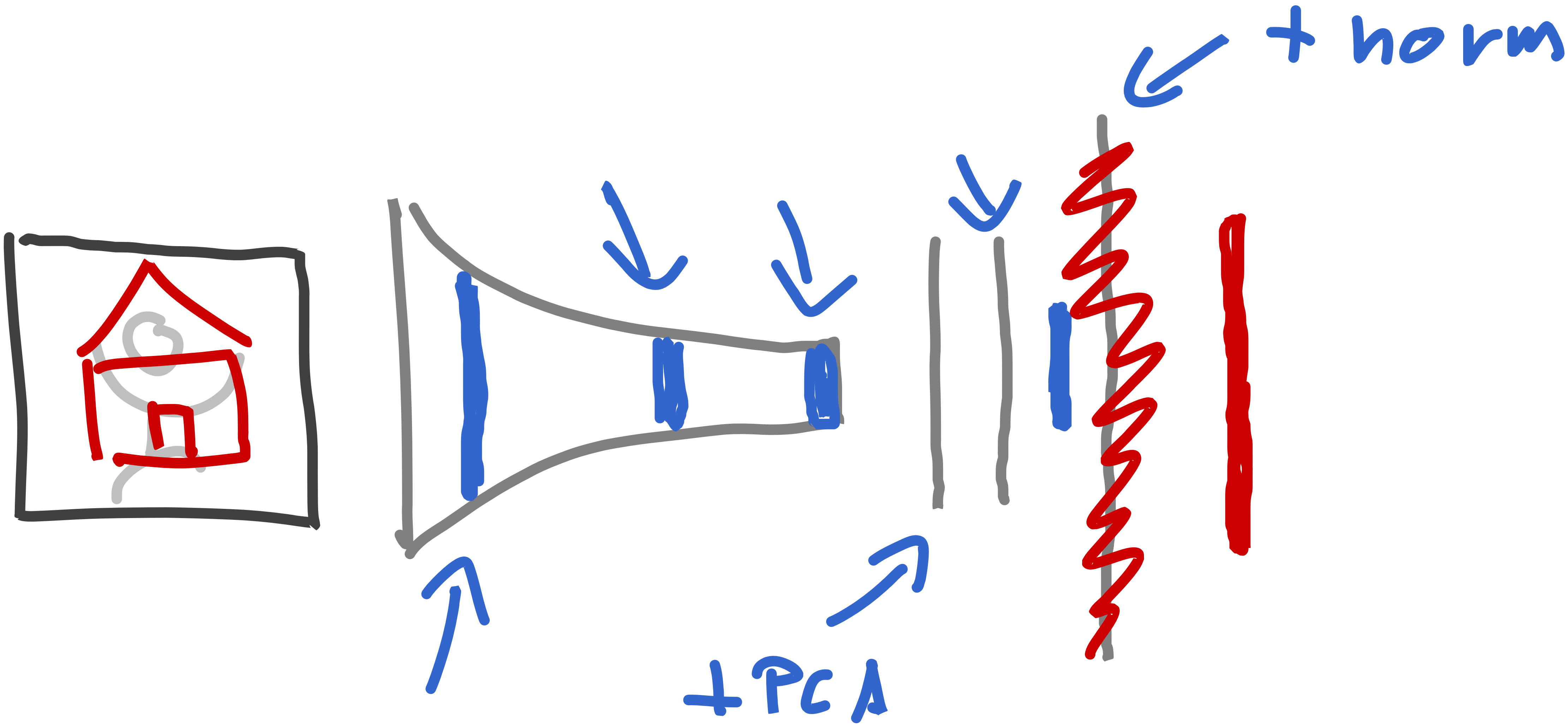
direct
representation




+ norm
+ PCA







+ norm
+ PCA

pencil_sharpener 15.5% | starfish 8.6% | triceratops 7.

Similarity Learning

contrastive loss

triplet loss

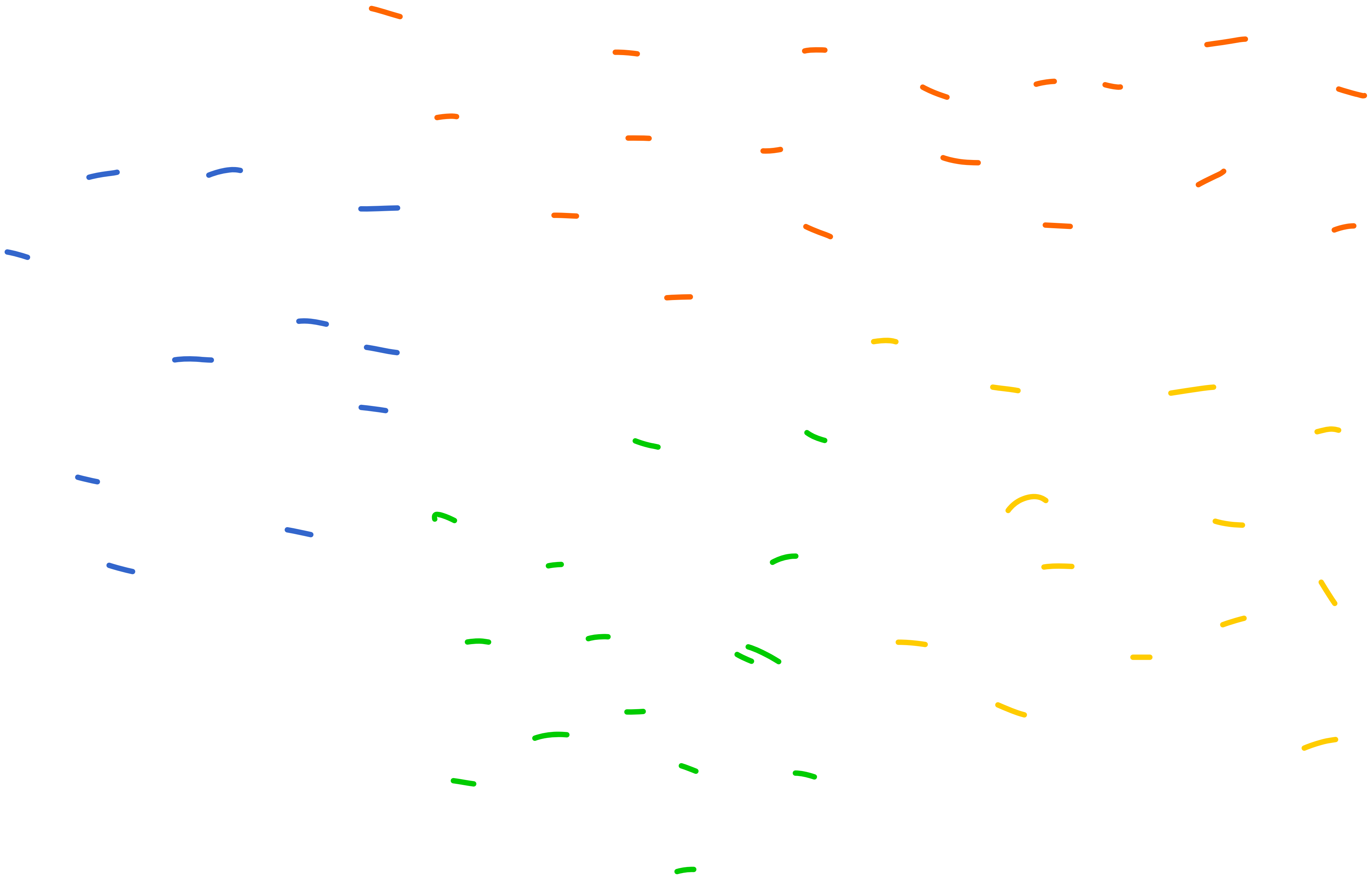
max margin

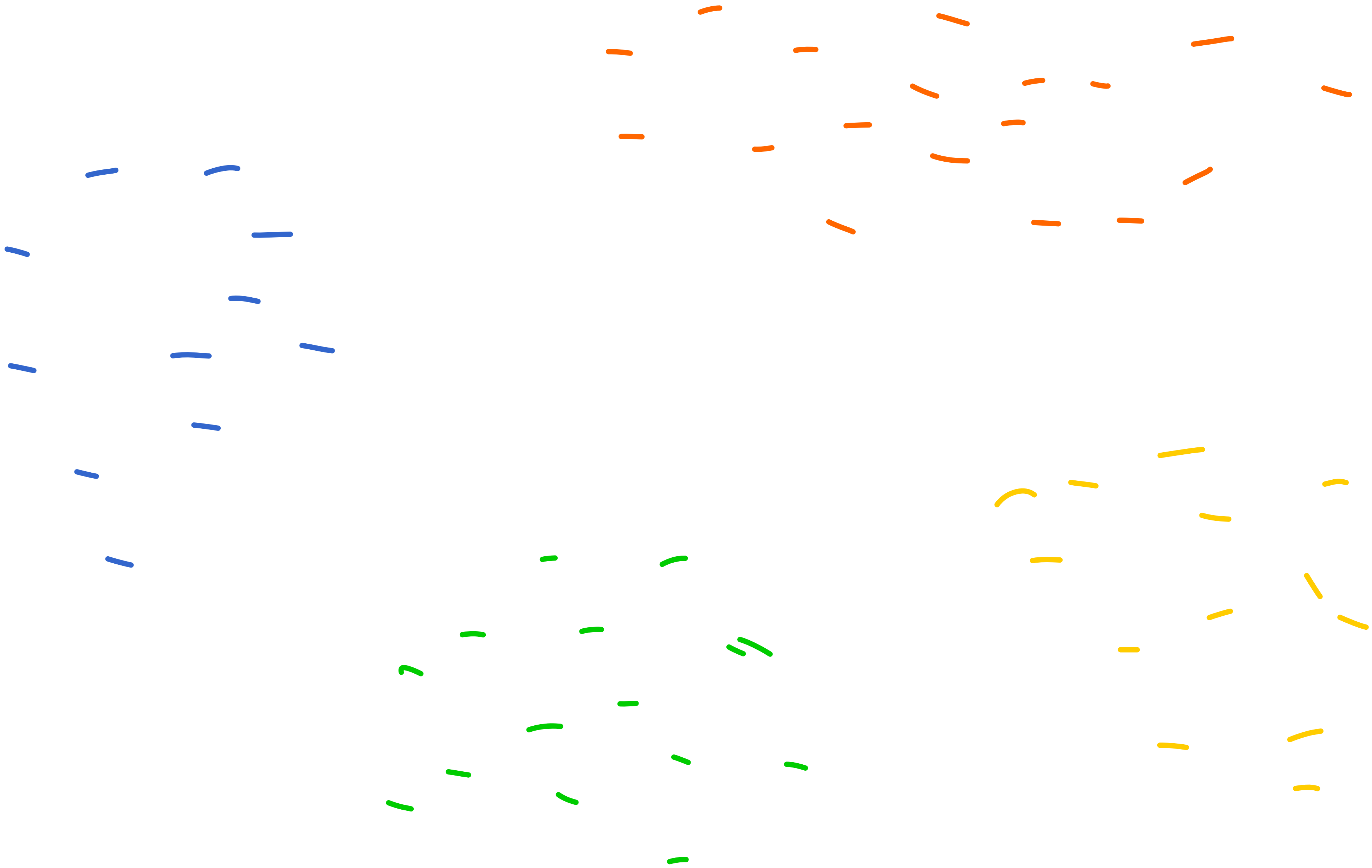
Similarity Learning

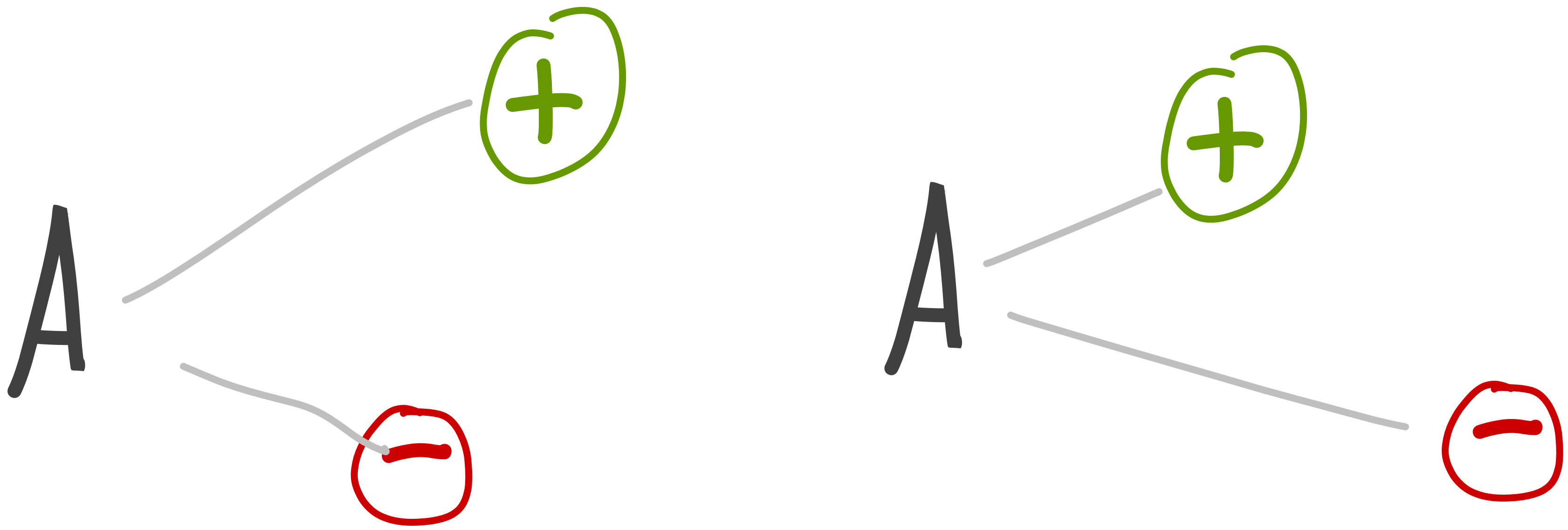
magnet loss

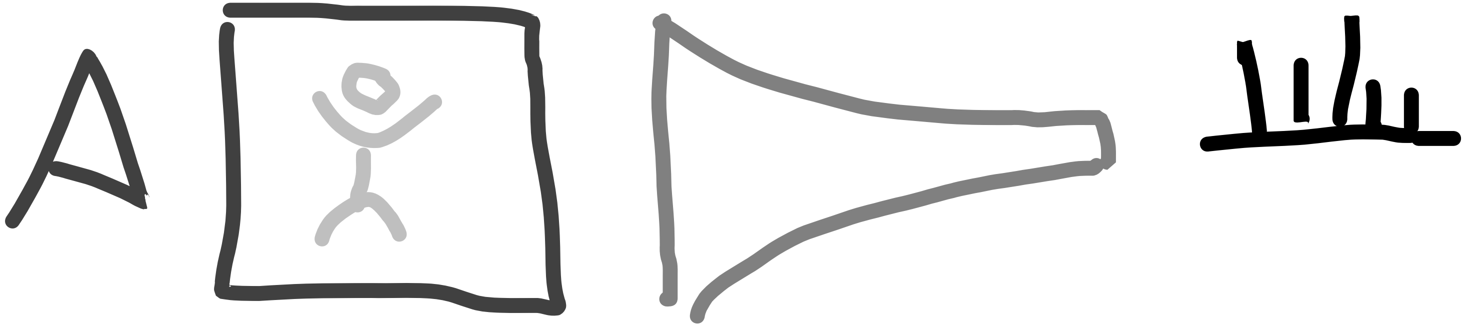
siamese networks

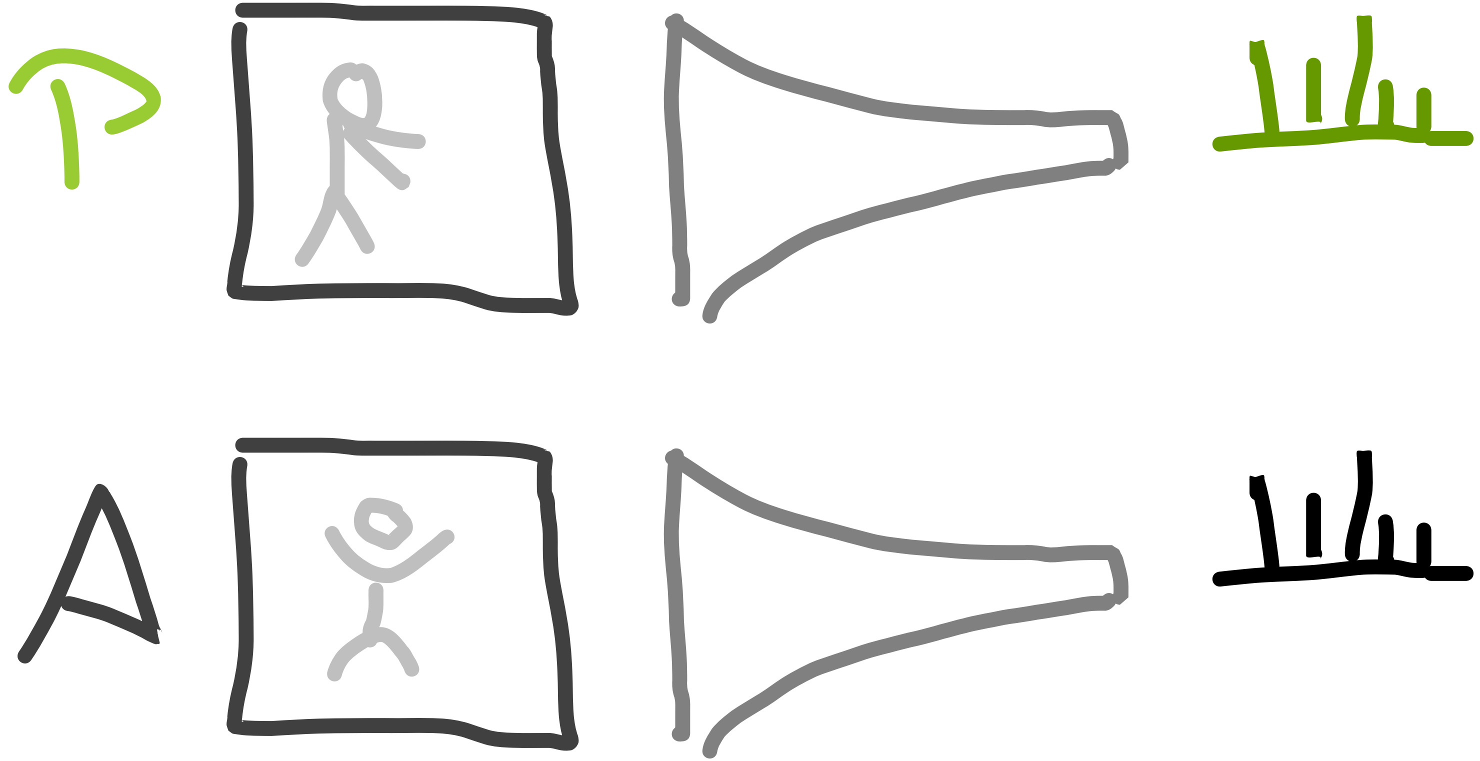
sampling

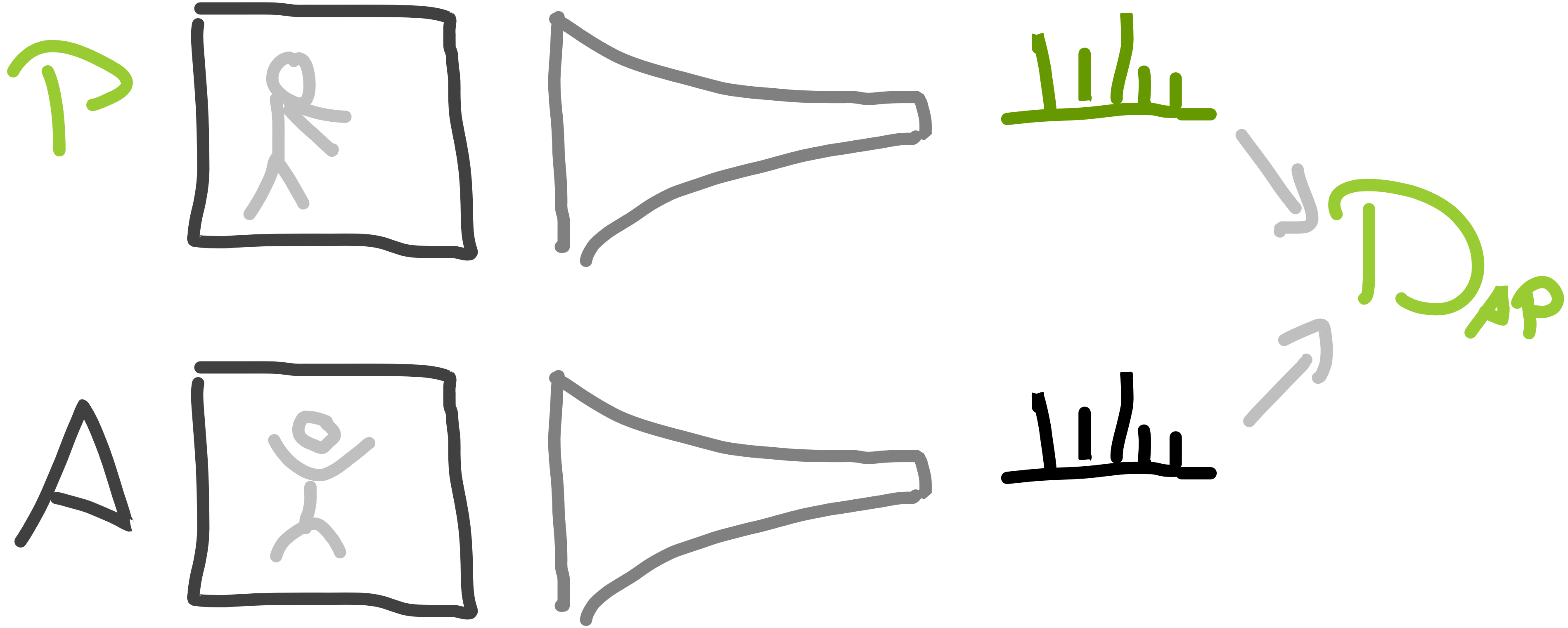




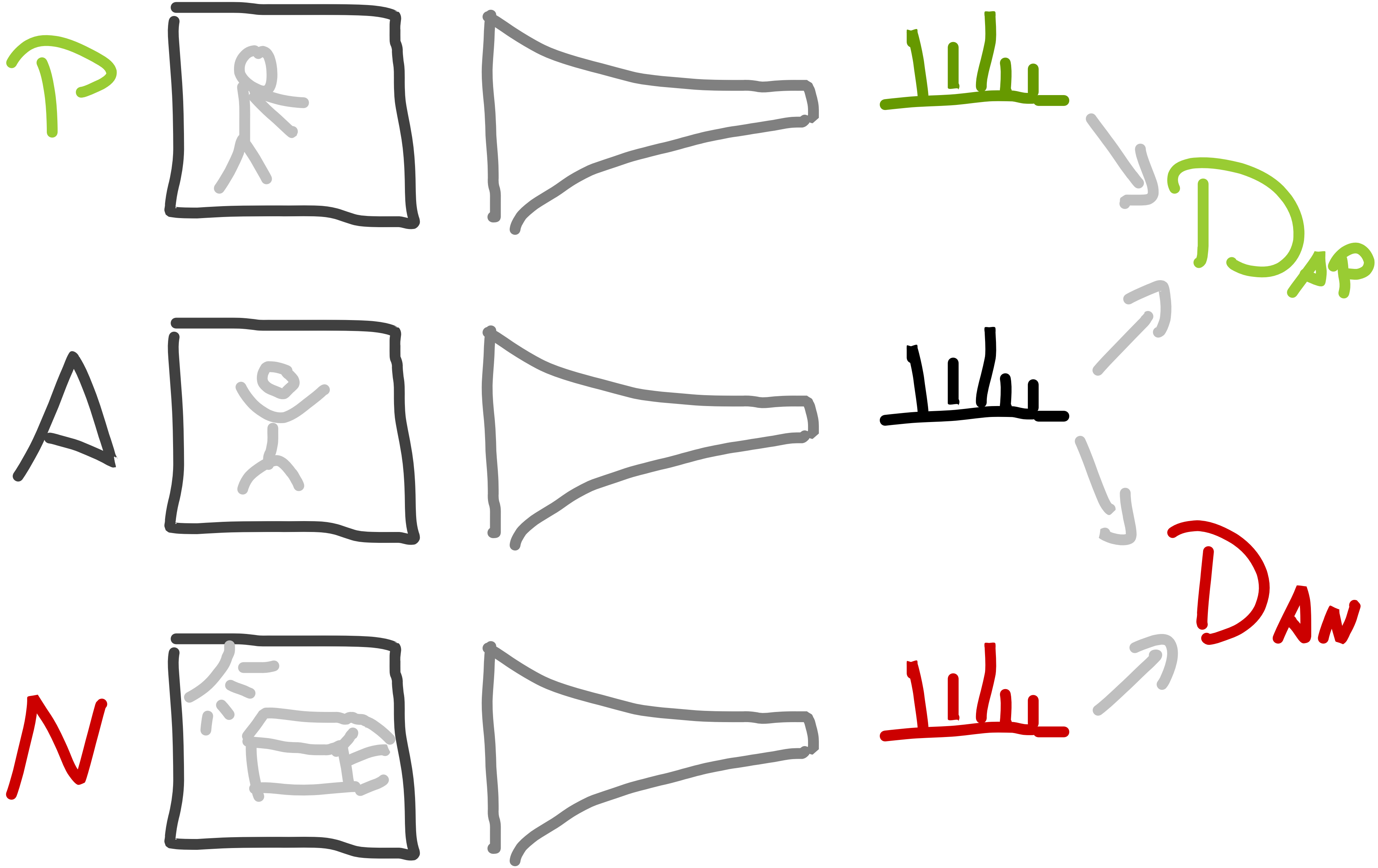

A
A

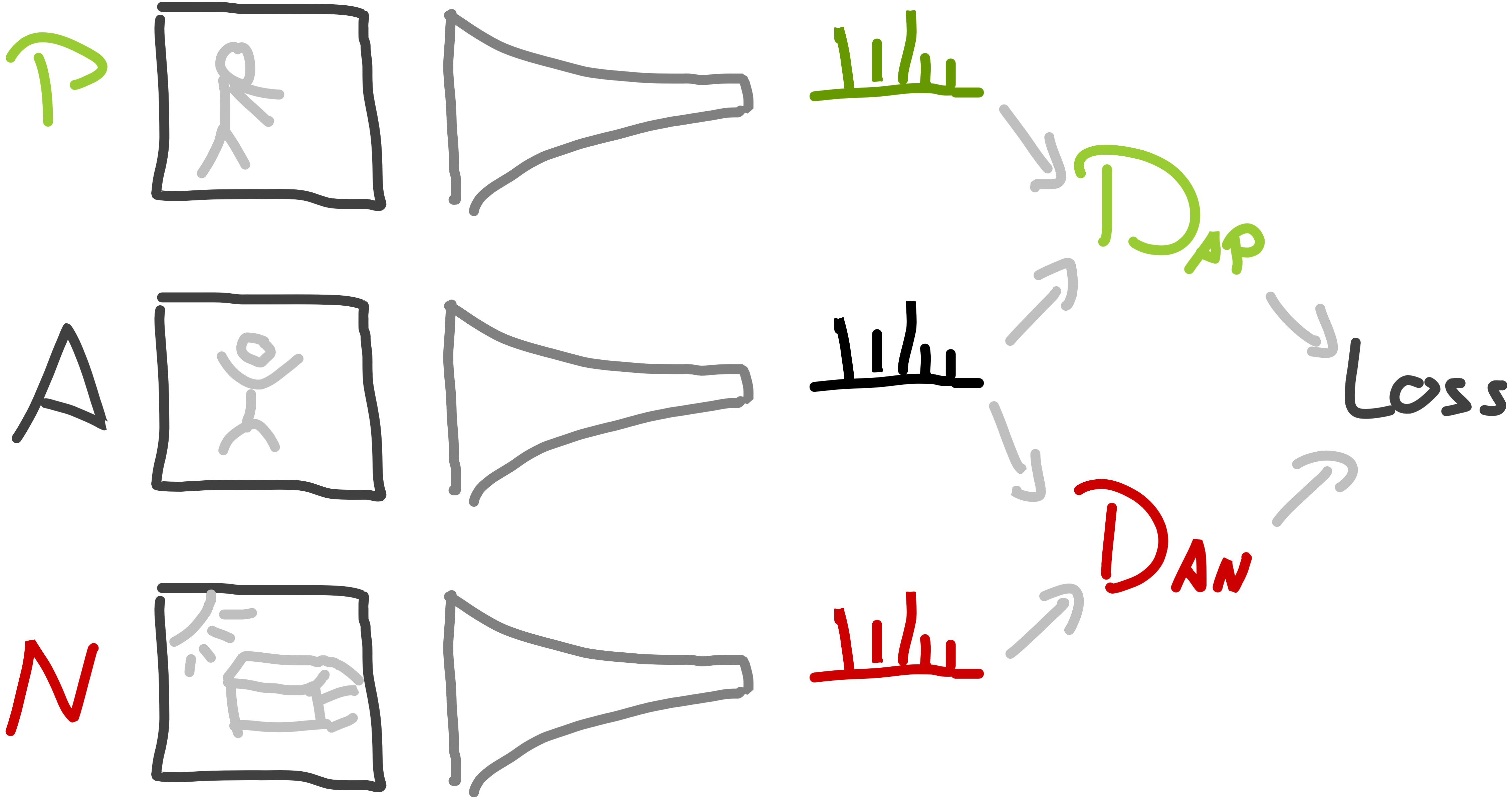

A

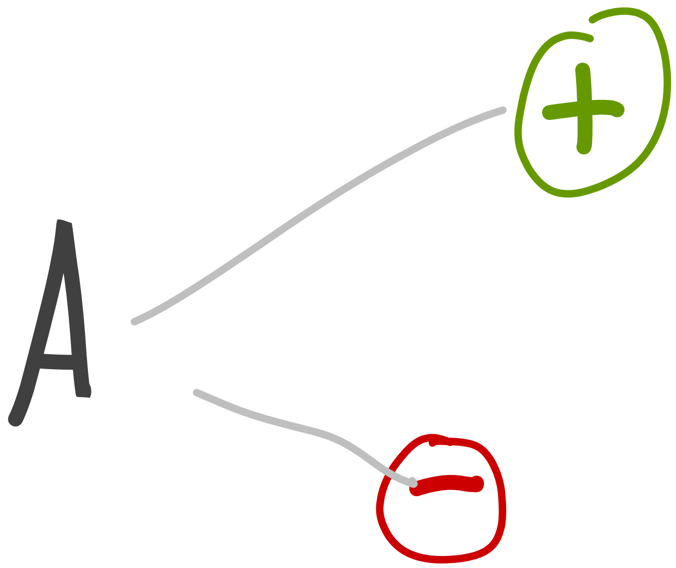
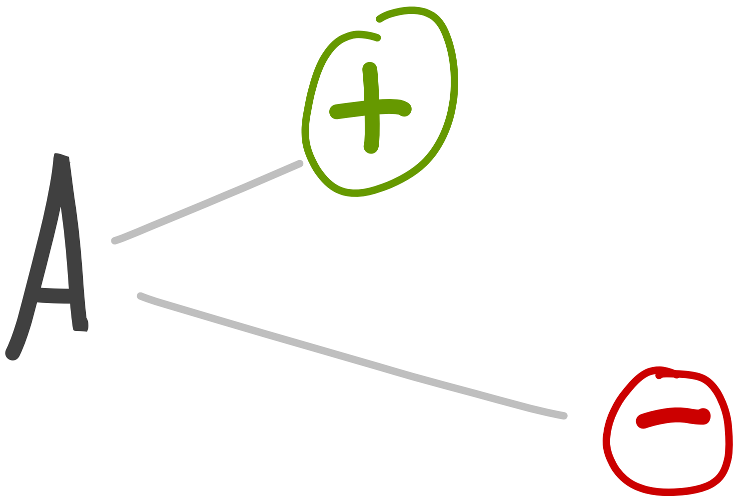

P
A



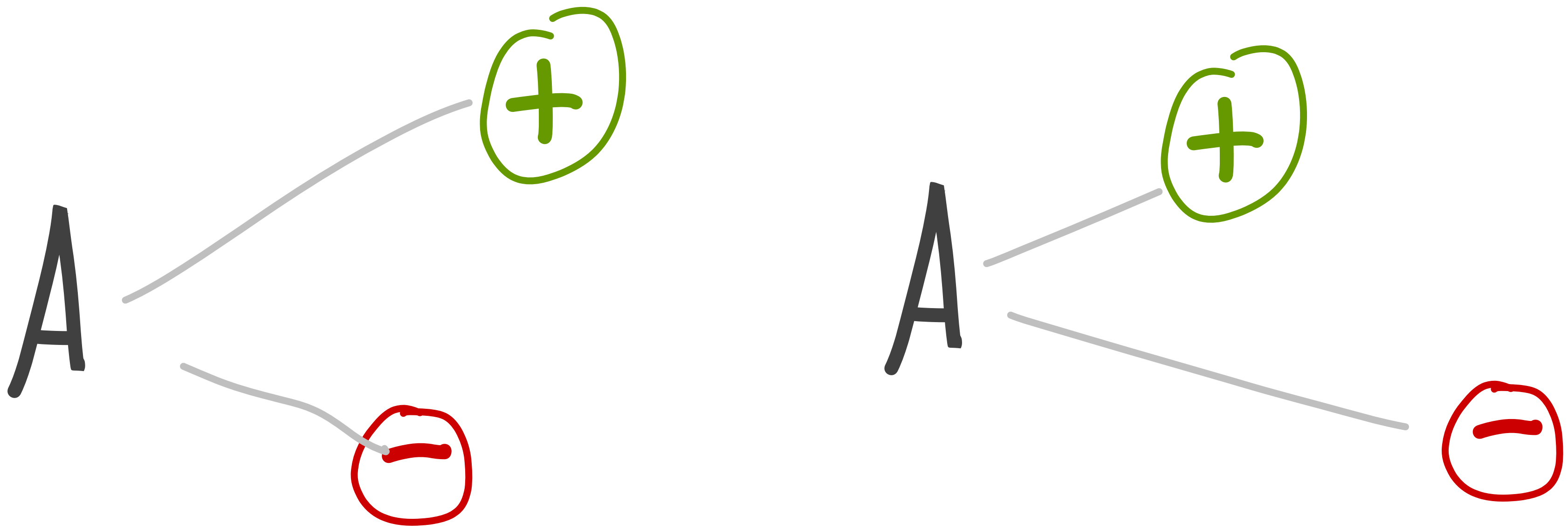
P
A
D_AP


P
A
N
D_{AP}
D_{AN}

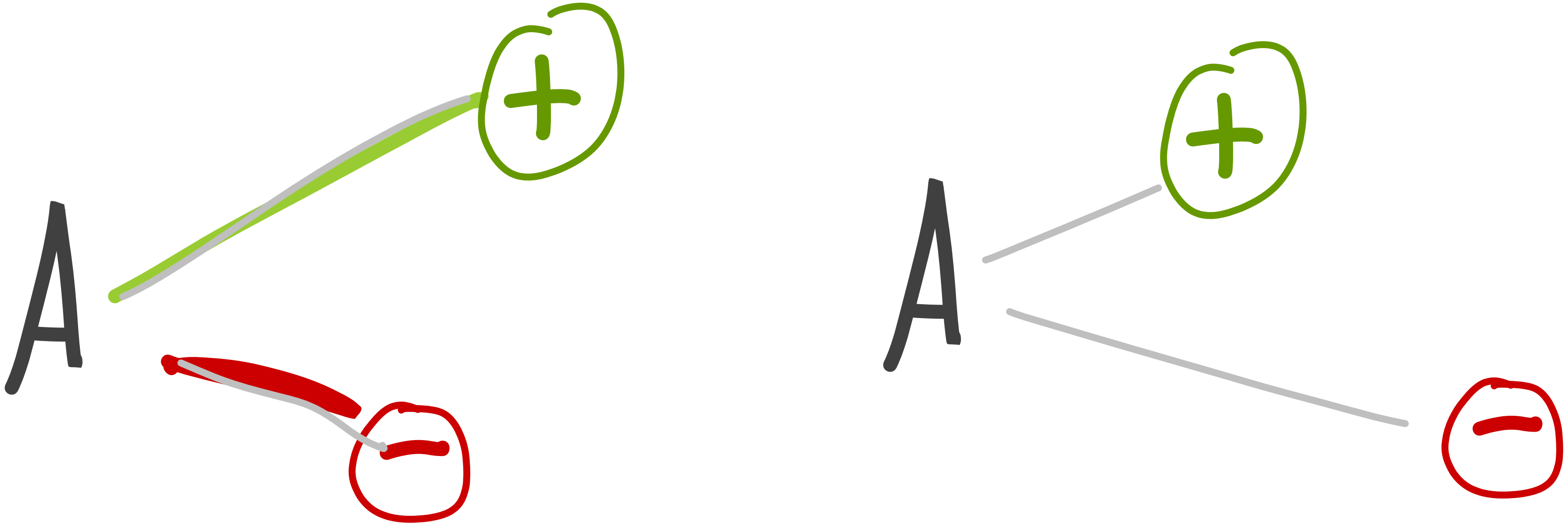

P
A
N
D
AP
D
AN
Loss


A
+
−

A
+
−

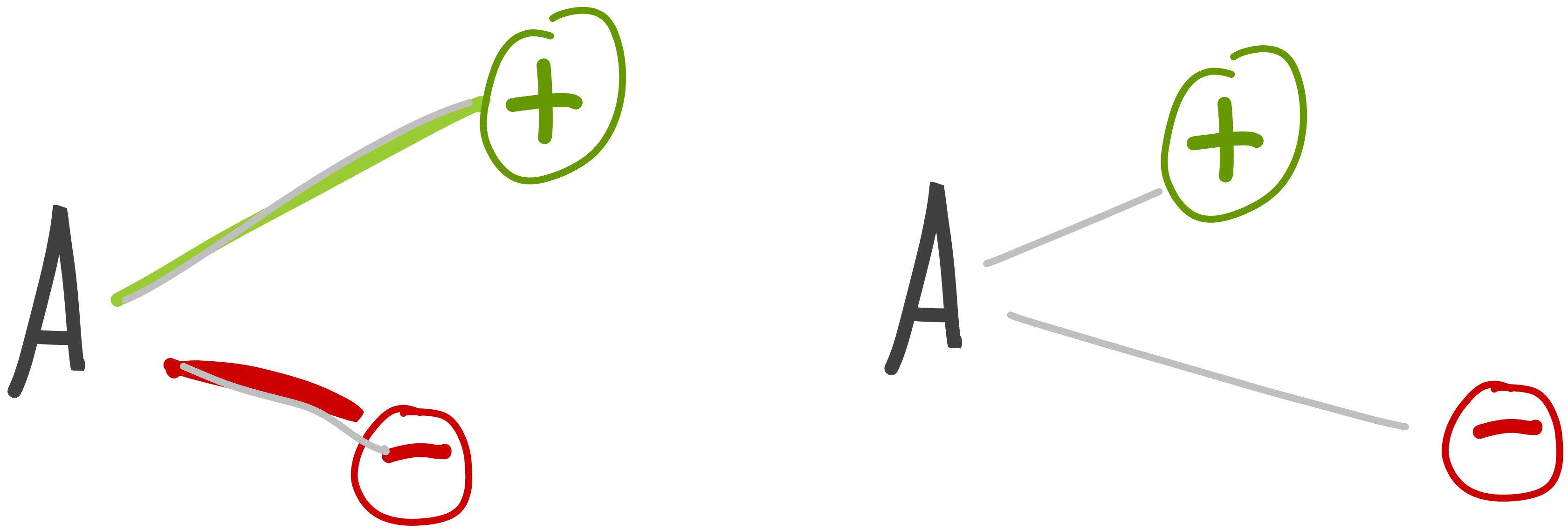
$$D_{AP} \leq D_{AN}$$



$$D_{AP} \le D_{AN}$$

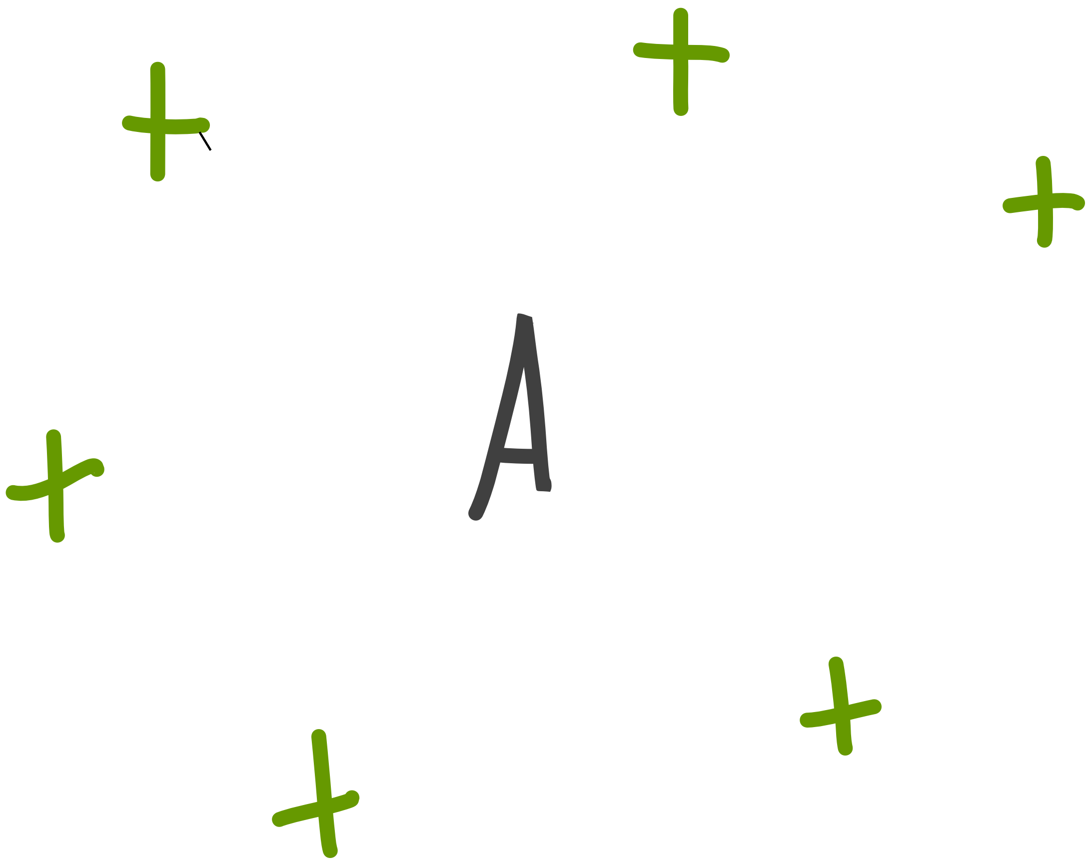


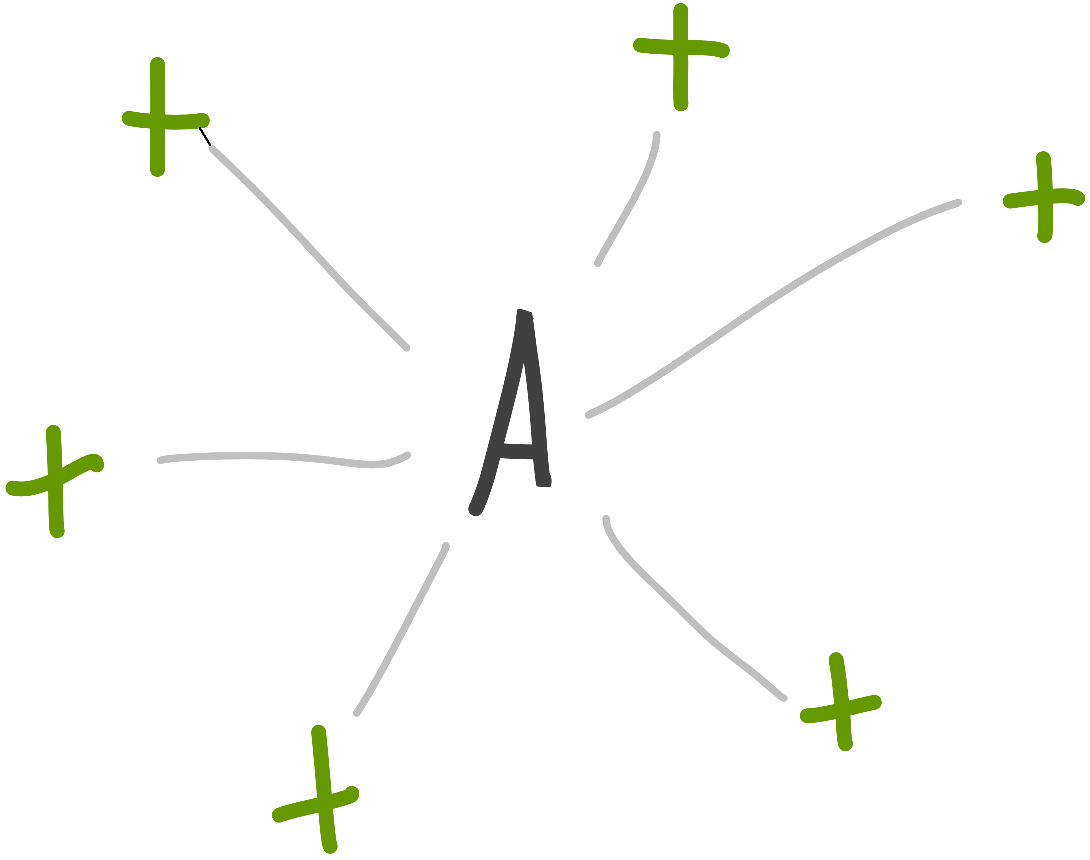
$$D_{AP} \leq D_{AN}$$

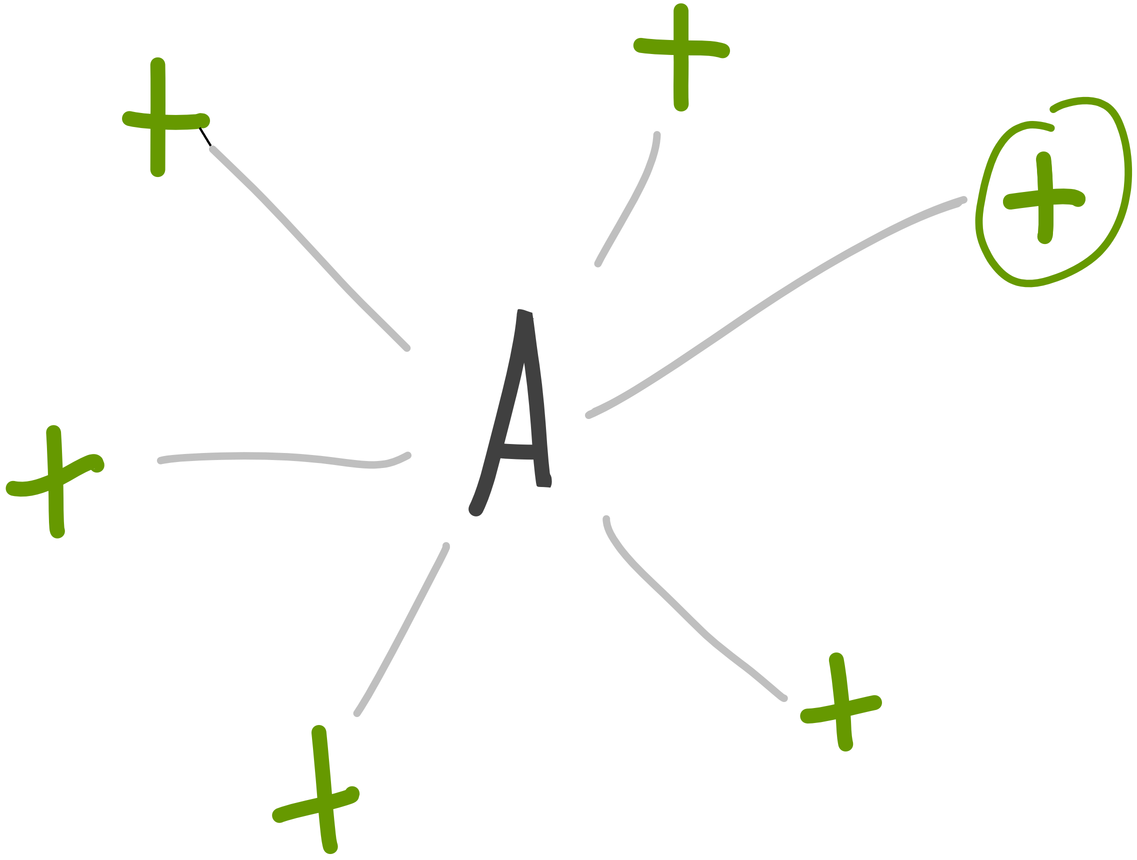


$$L = \max\{0, D_{AP} - D_{AN} + \alpha\}$$

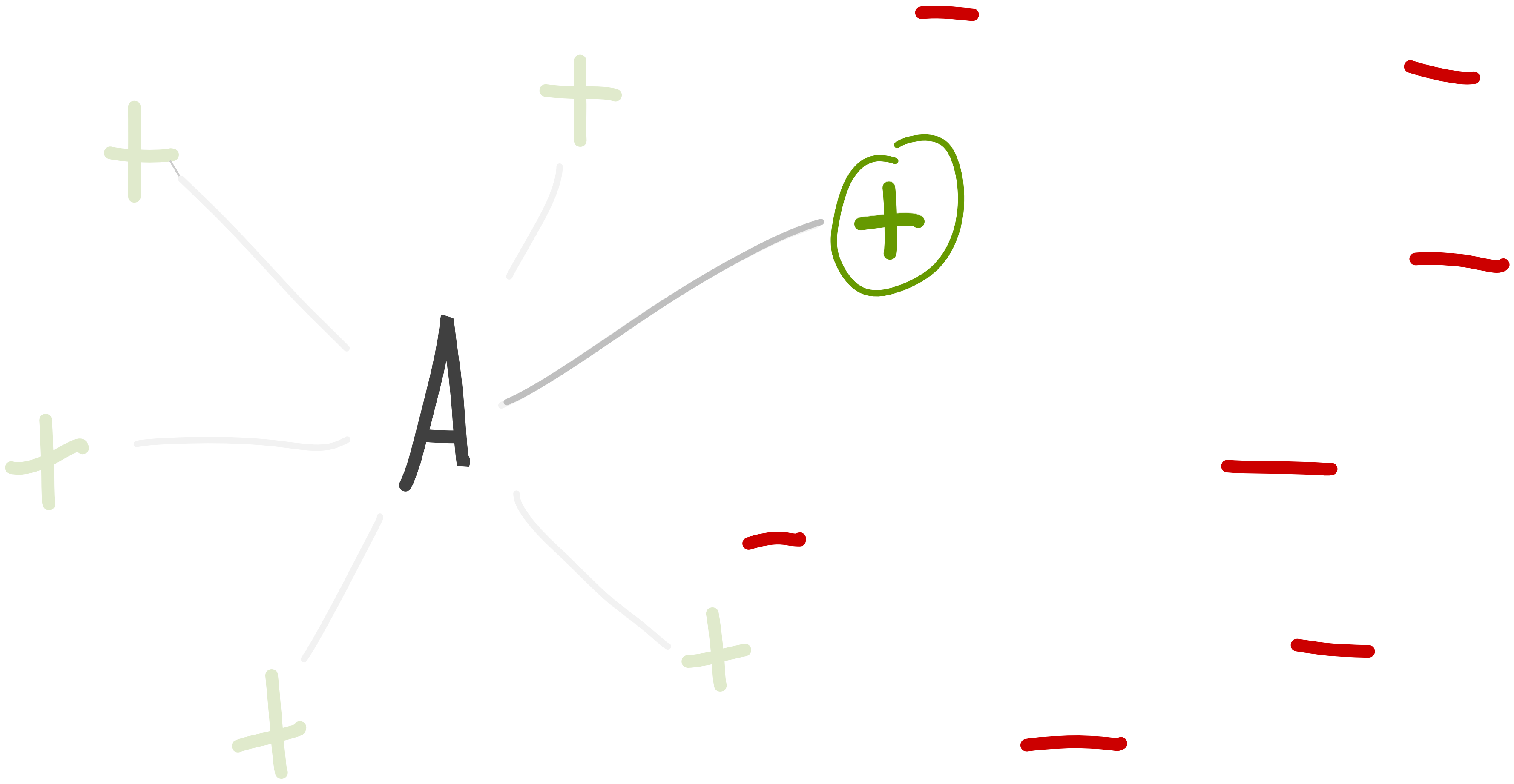
A

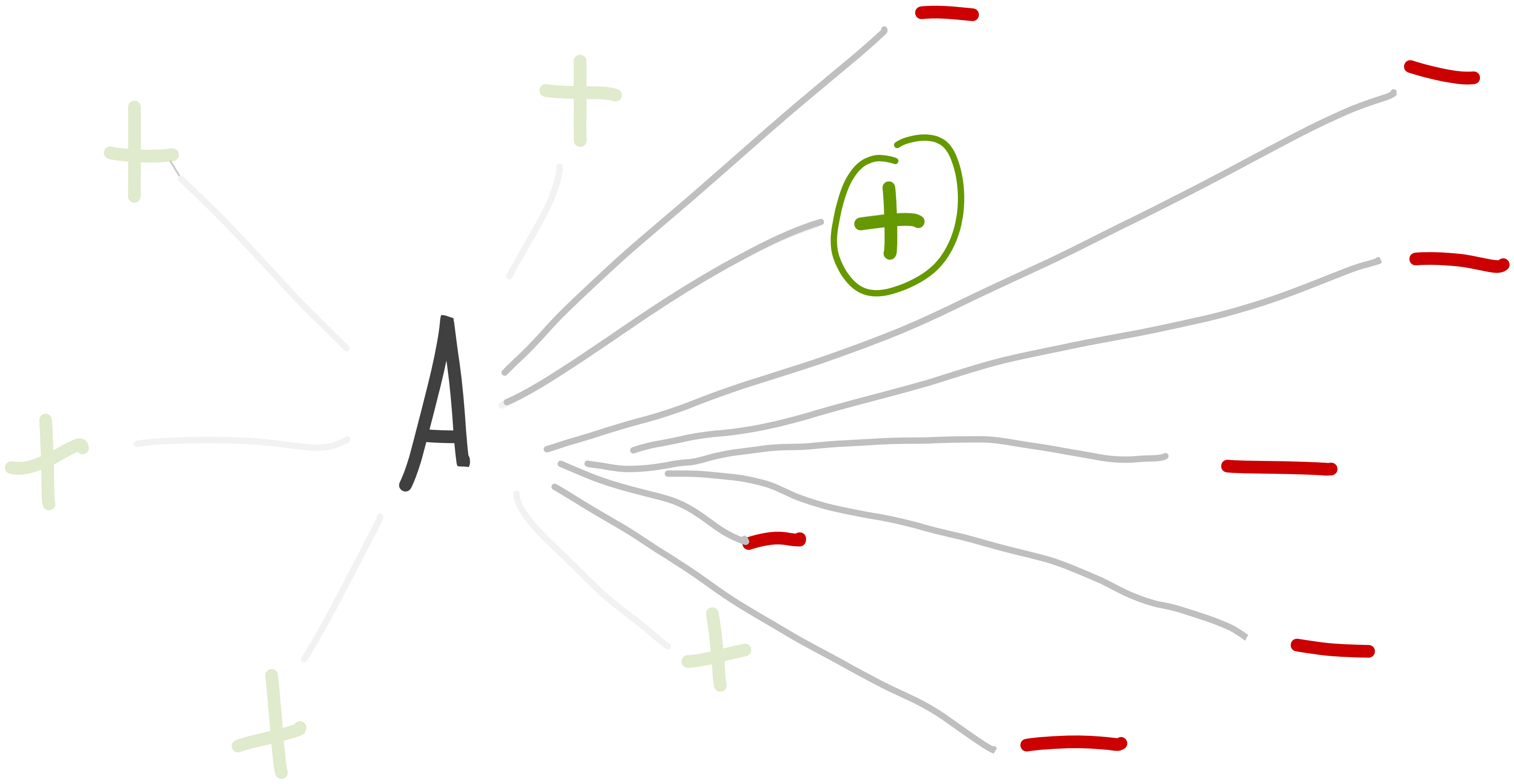

A

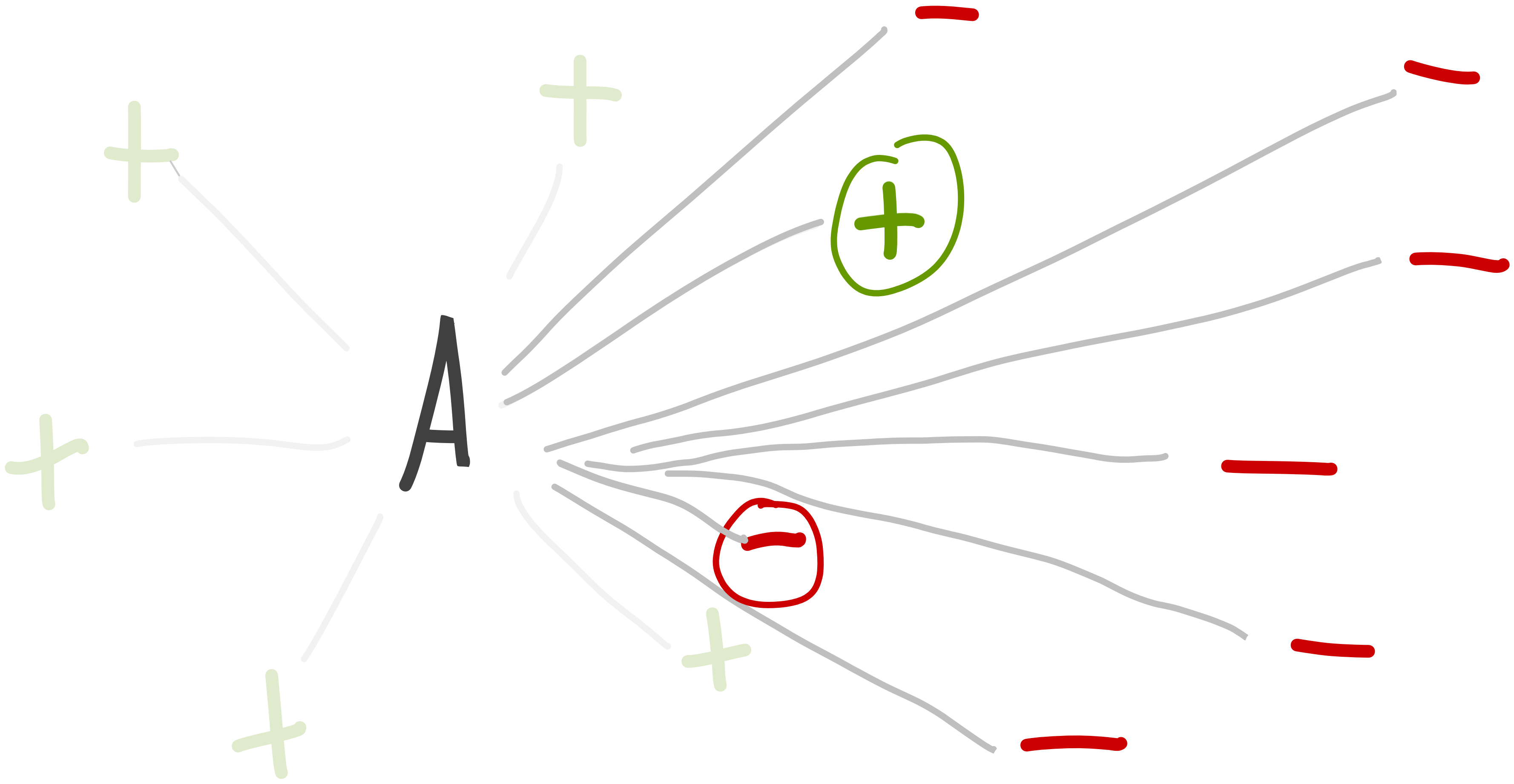

A

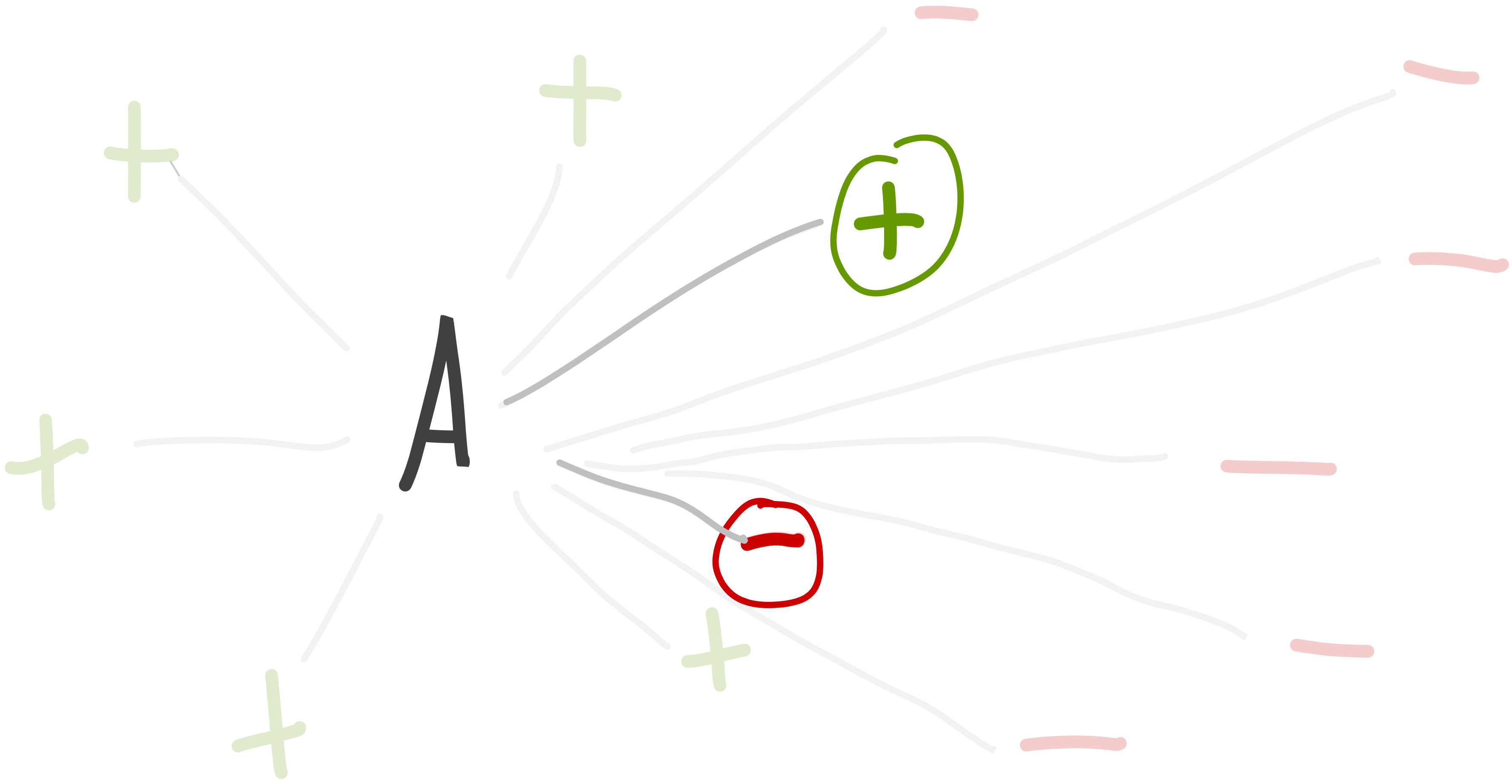

A

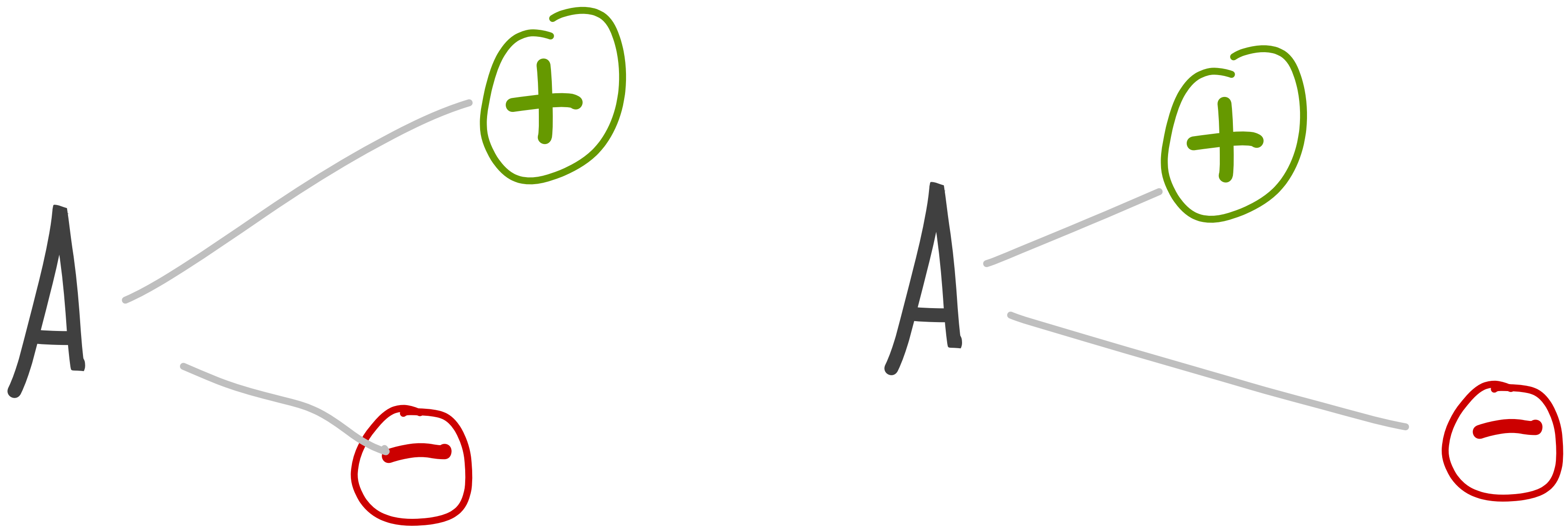

A


A


A


A


A



$$L = \max\{0, D_{AP} - D_{AN} + \alpha\}$$

FaceNet: A Unified Embedding for Face Recognition and Clustering

Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

Abstract

Despite significant recent advances in the field of face recognition [10, 14, 15, 17], implementing face verification and recognition efficiently at scale presents serious challenges to current approaches. In this paper we present a system, called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition, verification and clustering can be easily implemented

Published as a conference paper at ICLR 2016

METRIC LEARNING WITH ADAPTIVE DENSITY DISCRIMINATION

Oren Rippel
MIT, Facebook AI Research
rippel@math.mit.edu

Manohar Paluri
Facebook AI Research
mano@fb.com

Piotr Dollar
Facebook AI Research
pdollar@fb.com

Lubomir Bourdev
UC Berkeley
lubomir.bourdev@gmail.com

ABSTRACT

Distance metric learning (DML) approaches learn a transformation to a representation space where distance is in correspondence with a predefined notion of similarity. While such models offer a number of compelling benefits, it has been difficult for these to compete with modern classification algorithms in performance and even in feature extraction.

In this work, we propose a novel approach explicitly designed to address a num-

Sampling Matters in Deep Embedding Learning

Chao-Yuan Wu*
UT Austin
cywu@cs.utexas.edu

R. Manmatha
A9/Amazon
manmatha@a9.com

Alexander J. Smola
Amazon
smola@amazon.com

Philipp Krähenbühl
UT Austin
philkr@cs.utexas.edu

Abstract

Deep embeddings answer one simple question: How similar are two images? Learning these embeddings is the bedrock of verification, zero-shot learning, and visual search. The most prominent approaches optimize a deep convolutional network with a suitable loss function, such as contrastive loss or triplet loss. While a rich line of work focuses solely on the loss functions, we show in this paper that selecting training examples plays an equally important role. We propose distance weighted sampling, which selects more infor-

among the best-performing losses on standard embedding tasks [22, 25, 45]. Unlike pairwise losses, the triplet loss does not just change the loss function in isolation, it changes the way positive and negative example are selected. This provides us with two knobs to turn: the loss and the sampling strategy. See Figure 1 for an illustration.

In this paper, we show that sample selection in embedding learning plays an equal or more important role than the loss. For example, different sampling strategies lead to

Significance of Softmax-based Features in Comparison to Distance Metric Learning-based Features

Shota Horiguchi, Daiki Ikami, Kiyoharu Aizawa

Abstract—End-to-end distance metric learning (DML) has been applied to obtain features useful in many CV tasks. However, these DML studies have not provided equitable comparisons between features extracted from DML-based networks and softmax-based networks. In this paper, we present objective comparisons between these two approaches under the same network architecture.

Index Terms—deep learning, distance metric learning, classification, retrieval

1 Introduction

Recent developments in deep convolutional neural networks have made it possible to classify many classes of images with high accuracy. It has also been shown that such classification networks work well as feature extractors. Features extracted from classification networks show excellent performance in image classification [1], detection, and retrieval [2] [3], even when they have been trained to classify 1000 classes of the ImageNet dataset [4]. It has also been shown that fine-tuning

technically not novel, but they must be used for fair comparison between the image representations.

- We demonstrate that deep features extracted from softmax-based classification networks show competitive, or better results on clustering and retrieval tasks comparing to those from state-of-the-art DML-based networks [9], [10], [11] in the Caltech UCSD Birds 200-2011 dataset and the Stanford Cars 196 dataset.
- We show how the clustering and retrieval performances of softmax-based features and DML features change according to the size of the dataset. DML features show competitive or better performance in the stanford Online Product dataset which consists of very small number of samples per class.

In order to align the condition of the network architecture, we restrict the network architecture to GoogLeNet [14] which has been used in state-of-the-art of DML studies [9], [10], [11].

2 Background

2.1 Previous Work

2.1.1 *Softmax-Based Classification and Repurposing of the Classifier as a Feature Extractor*

Convolutional neural networks have demonstrated great potential for highly accurate image recognition [15] [16] [14] [17]. It has been shown that features extracted from classification networks can be repurposed as a good feature representation

Speed-up Search

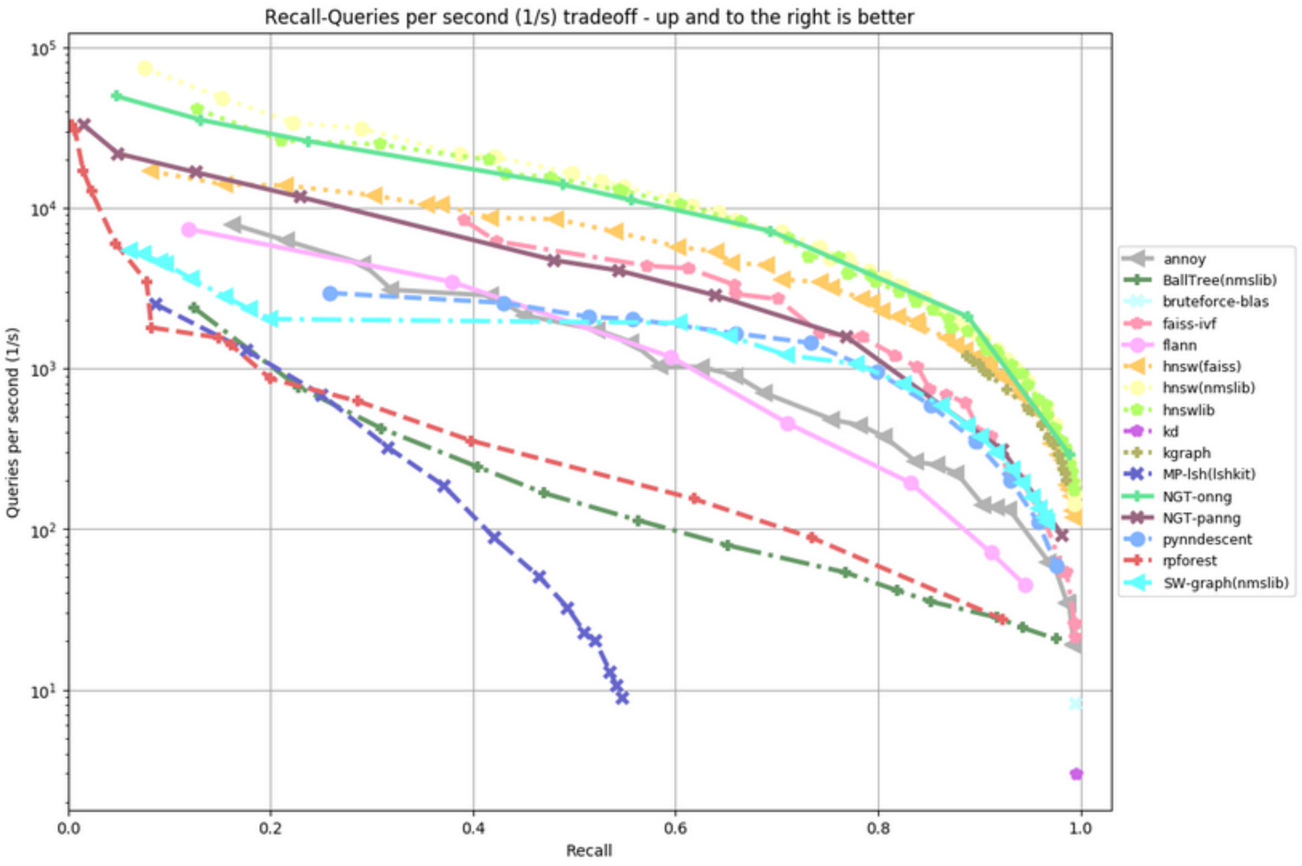
```
sim = F.cosine_similarity(
    query_feature, reference_features,
)

sorted_sim, sorted_index = torch.topk(
    sim, k=top_k,
)
```

Benchmarks for Single Queries

Results by Dataset

Distance: Angular

glove-100-angular (k = 10)



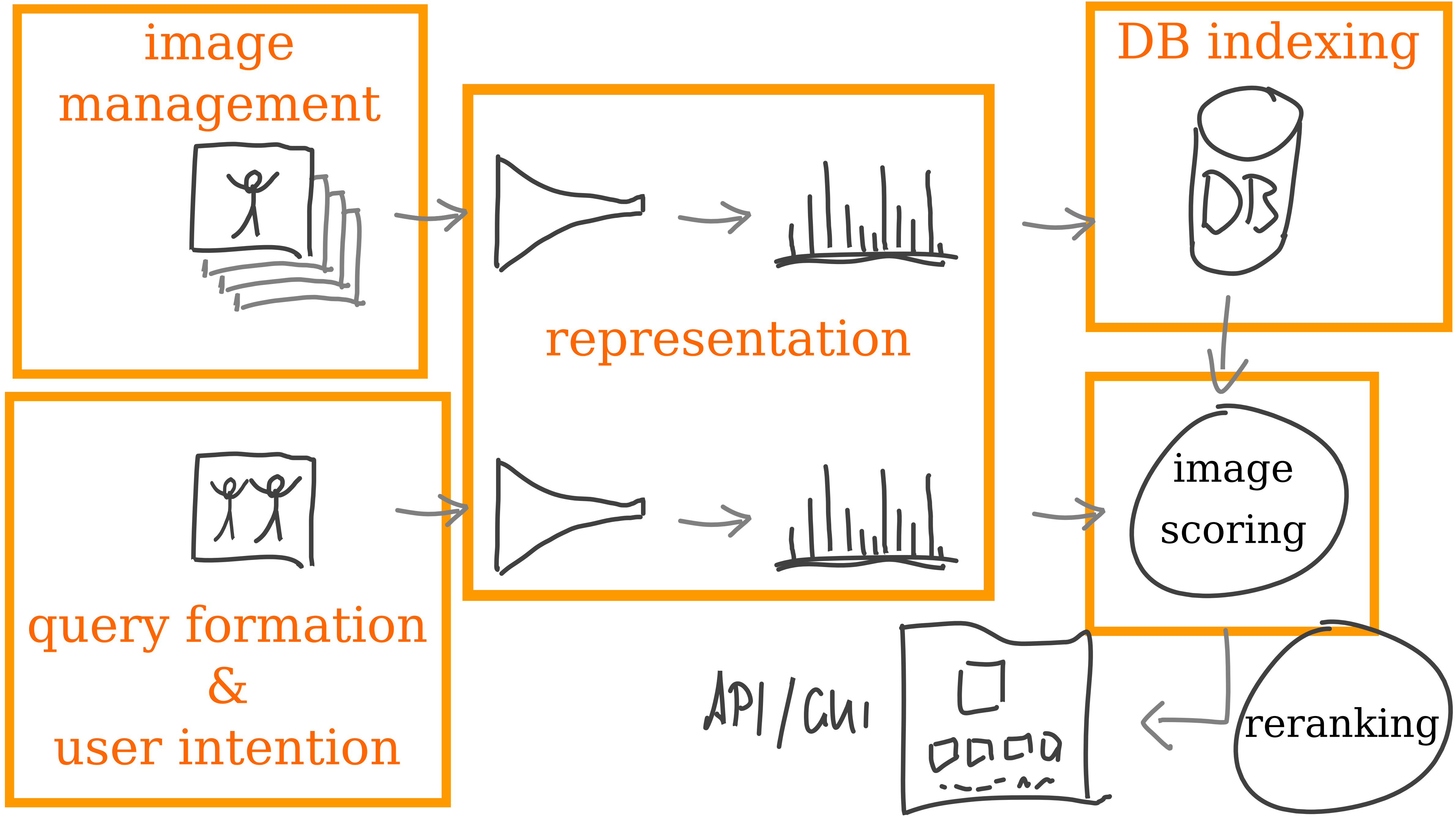
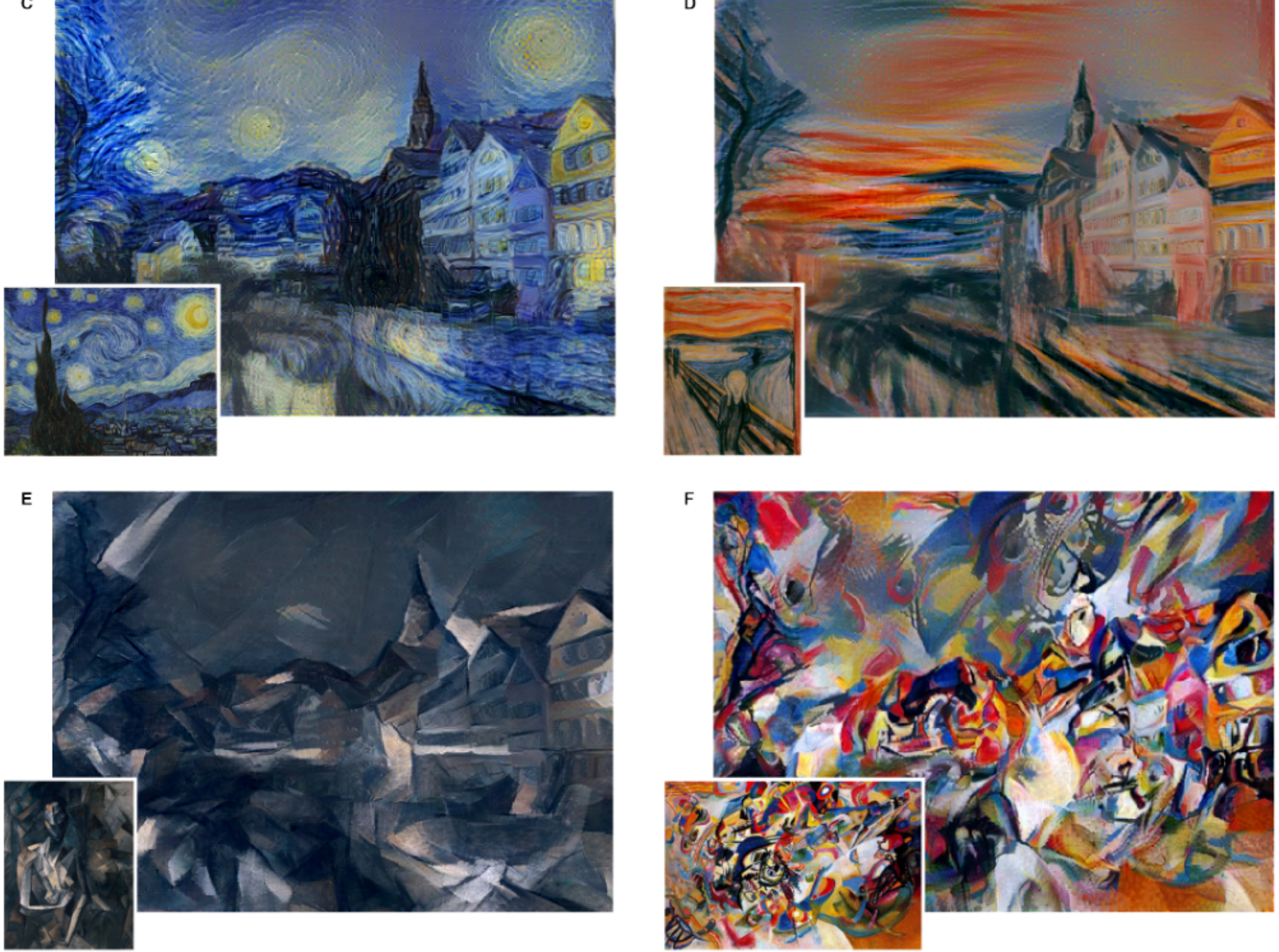


image
management
representation
DB indexing
DB
query formation
&
user intention
image
scoring
API/GUI
reranking

fun


C
D
E
F

Zero-Shot Learning by Convex Combination of Semantic Embeddings

Mohammad Norouzi*, **Tomas Mikolov**, **Samy Bengio**, **Yoram Singer**,
Jonathon Shlens, **Andrea Frome**, **Greg S. Corrado**, **Jeffrey Dean**

norouzi@cs.toronto.edu, {tmikolov, bengio, singer}@google.com
{shlens, afrome, gcorrado, jeff}@google.com

*University of Toronto
ON, Canada

Google, Inc.
Mountain View, CA, USA

Abstract

Several recent publications have proposed methods for mapping images into continuous semantic embedding spaces. In some cases the embedding space is trained

WE CREATE VISIBILLITY

Q & A

Fine-tuning CNN Image Retrieval with No Human Annotation

Filip Radenović Giorgos Tolias Ondřej Chum

Abstract—Image descriptors based on activations of Convolutional Neural Networks (CNNs) have become dominant in image retrieval due to their discriminative power, compactness of representation, and search efficiency. Training of CNNs, either from scratch or fine-tuning, requires a large amount of annotated data, where a high quality of annotation is often crucial. In this work, we propose to fine-tune CNNs for image retrieval on a large collection of unordered images in a fully automated manner. Reconstructed 3D models obtained by the state-of-the-art retrieval and structure-from-motion methods guide the selection of the training data. We show that both hard-positive and hard-negative examples, selected by exploiting the geometry and the camera positions available from the 3D models, enhance the performance of particular-object retrieval. CNN descriptor whitening discriminatively learned from the same training data outperforms commonly used PCA whitening. We propose a novel trainable Generalized-Mean (GeM) pooling layer that generalizes max and average pooling and show that it boosts retrieval performance. Applying the proposed method to the VGG network achieves state-of-the-art performance on the standard benchmarks: Oxford Buildings, Paris, and Holidays datasets.

1 Introduction

In instance image retrieval an image of a particular object, depicted in a query, is sought in a large, unordered collection of images. Convolutional neural networks (CNNs)

the similarity measure to be used in the final task by selecting *matching* and *non-matching* pairs to perform the training.

In contrast to previous methods of training-data acqui-

Published in MM '14: Proceedings of the ACM International Conference on Multimedia: November 3-7, 2014, Orlando. pp. 157-166.
http://doi.org/10.1145/2647868.2654948

Deep Learning for Content-Based Image Retrieval: A Comprehensive Study

Ji Wan[1,2,5], Dayong Wang[3], Steven C.H. Hoi[2], Pengcheng Wu[3],
Jianke Zhu[4], Yongdong Zhang[1], Jintao Li[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, China
[2]School of Information Systems, Singapore Management University, Singapore
[3]School of Computer Engineering, Nanyang Technological University, Singapore
[4]College of Computer Science, Zhejiang University, Hangzhou, China
[5]University of Chinese Academy of Sciences, Beijing, China
chhoi@smu.edu.sg; {dywang,wupe0003}@ntu.edu.sg; {wanji,zhyd,jtli}@ict.ac.cn; jkzhu@zju.edu.cn

ABSTRACT

Learning effective feature representations and similarity measures are crucial to the retrieval performance of a content-based image retrieval (CBIR) system. Despite extensive research efforts for decades, it remains one of the most challenging open problems that considerably hinders the successes of real-world CBIR systems. The key challenge has been attributed to the well-known "semantic gap" issue that exists between low-level image pixels captured by machines and high-level semantic concepts perceived by

1. INTRODUCTION

The retrieval performance of a content-based image retrieval system crucially depends on the feature representation and similarity measurement, which have been extensively studied by multimedia researchers for decades. Although a variety of techniques have been proposed, it remains one of the most challenging problems in current content-based image retrieval (CBIR) research, which is mainly due to the well-known "semantic gap" issue that exists between low-level image pixels captured by machines and high-level

Recent Advance in Content-based Image Retrieval: A Literature Survey

Wengang Zhou, Houqiang Li, and Qi Tian *Fellow, IEEE*

Abstract—The explosive increase and ubiquitous accessibility of visual data on the Web have led to the prosperity of research activity in image search or retrieval. With the ignorance of visual content as a ranking clue, methods with text search techniques for visual retrieval may suffer inconsistency between the text words and visual content. Content-based image retrieval (CBIR), which makes use of the representation of visual content to identify relevant images, has attracted sustained attention in recent two decades. Such a problem is challenging due to the intention gap and the semantic gap problems. Numerous techniques have been developed for content-based image retrieval in the last decade. The purpose of this paper is to categorize and evaluate those algorithms proposed during the period of 2003 to 2016. We conclude with several promising directions for future research.

Index Terms—content-based image retrieval, visual representation, indexing, similarity measurement, spatial context, search re-ranking.

1 Introduction

With the universal popularity of digital devices embedded

From the early 1990s to the early 2000s, there have been extensive study on content-based image search. The progress in those years has been comprehensively discussed