Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

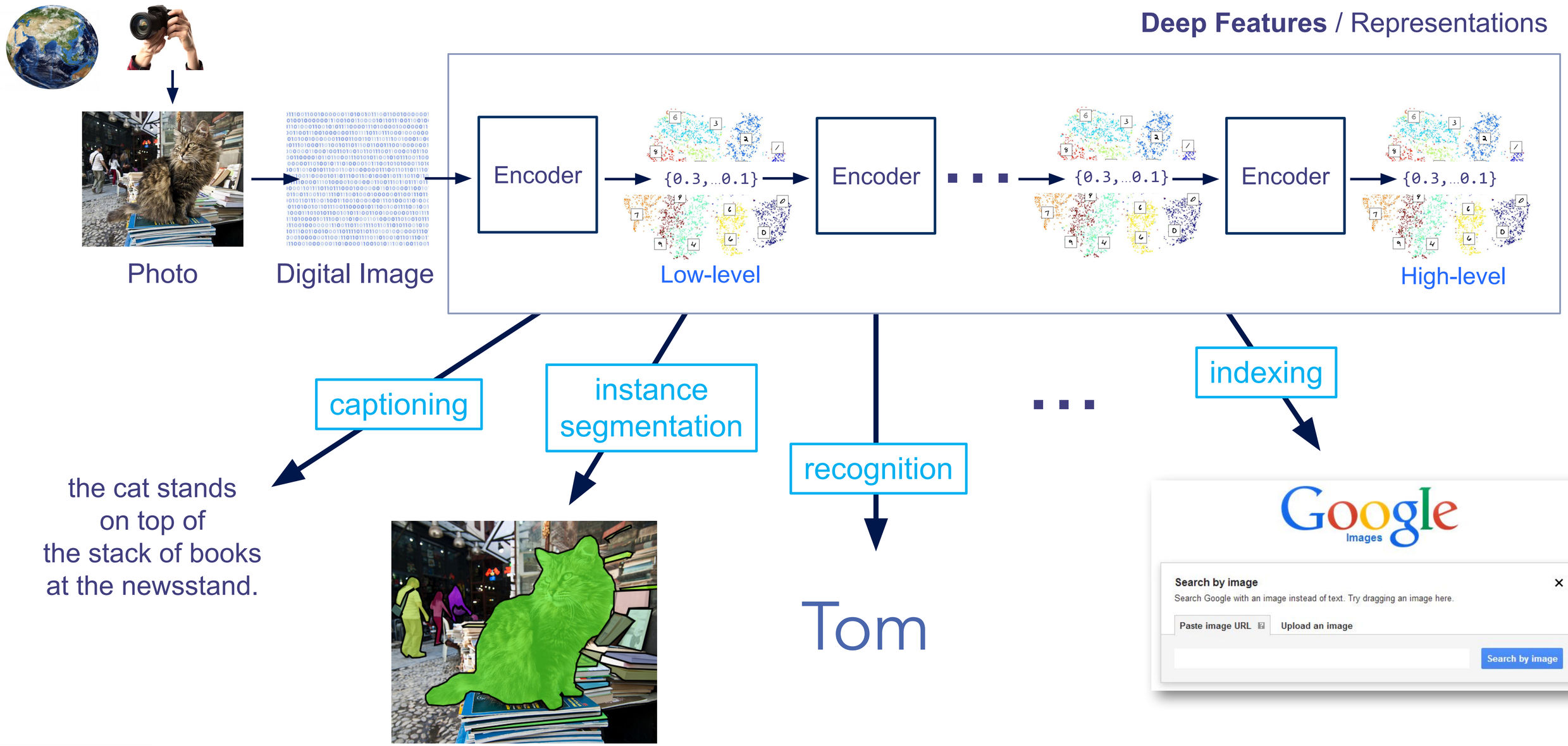
Deep Learning Tools for Image Classification and Retrieval

Fabio Carrara, Nicola Messina

Researcher @ ISTI-CNR
fabio.carrara@isti.cnr.it

PostDoc @ ISTI-CNR
nicola.messina@isti.cnr.it

The Overall Picture



Outline

- Deep Learning for Images
  - Introduction: Image Features/Representation
  - Intuition: Why we use Deep Features
  - From Light to Bits
  - Storing and Sharing Images
- Image Classification
- Image Retrieval

Introduction:
Image Features / Representations

An Image, Many Tasks!

What we want to do:

- captioning
- similarity
- segmentation
- detection
- recognition
- classification
- retrieval
- ...

From Real World to Bits

photography →

digitalization → (compression)

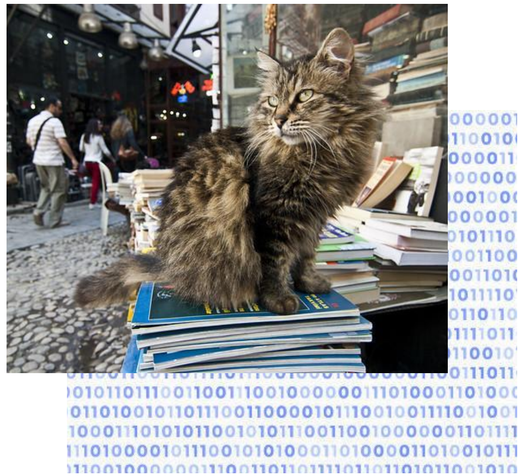
Camera Photo Bits

Analysis are performed on a digital image

A digital image (usually) is a digital photo of the real world.
(not the real world, <u>not an analog photo</u>)

The bits representing the digital images have **very low semantic**.

Representation / Features



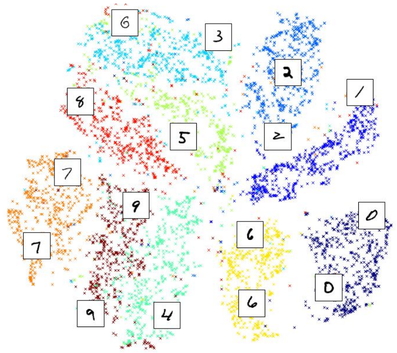
Digital Image

Encoder

feature extractor

A representation (a feature)

{0.3, 0.5 ... 0.1}



in a **latent space**

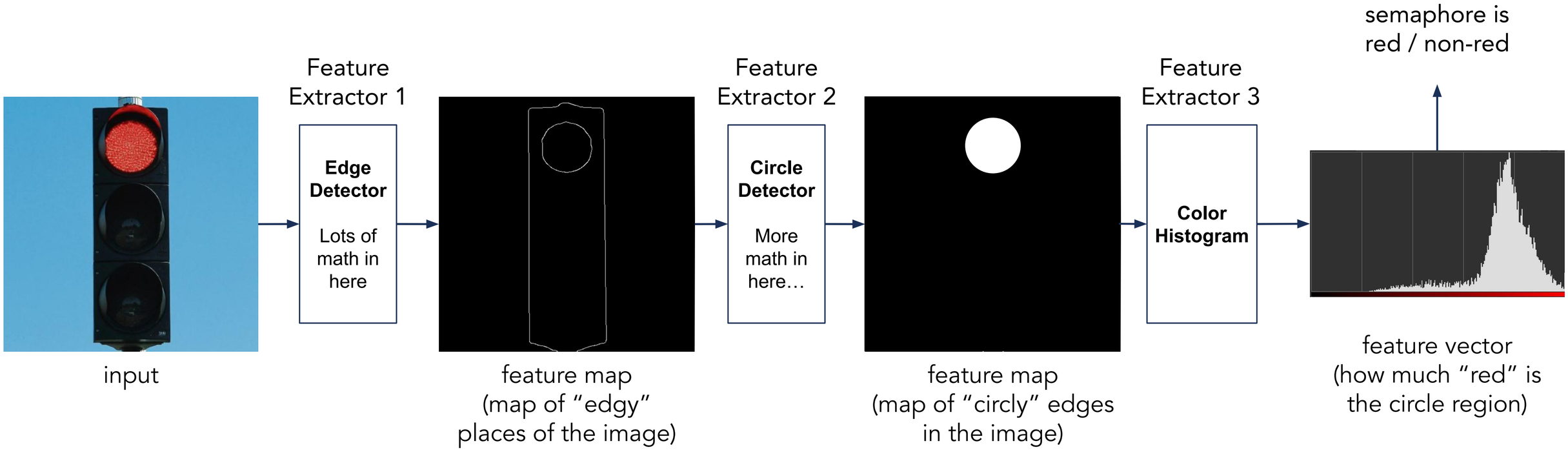
- captioning
- similarity
- segmentation
- detection
- recognition
- classification
- retrieval
- ...

An encoder (or feature extractor)
takes an input (e.g., an image)
and produce a representation (or descriptor)
that is used (in place of the image) for the specific task.

Handcrafted Features

Task: find out automatically <u>if semaphore is red</u>

Hierarchy of features you may manually define to solve the task:



Features / Representation

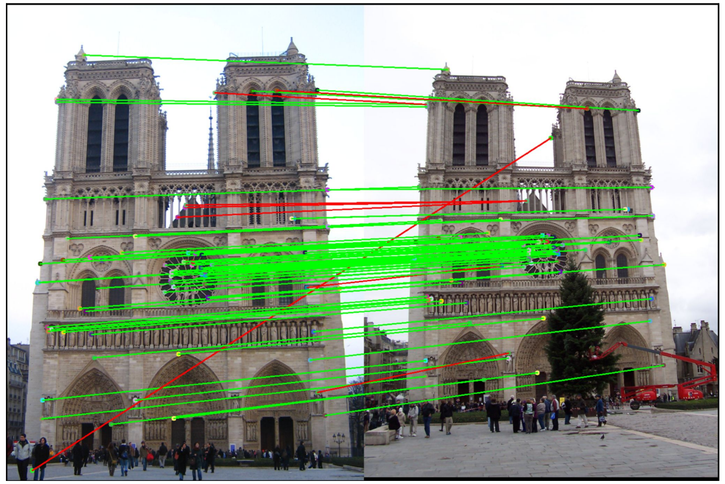
Before Deep Learning, handcrafted features/representations:

Global Features:

- color, edge, texture etc...

Local Features:

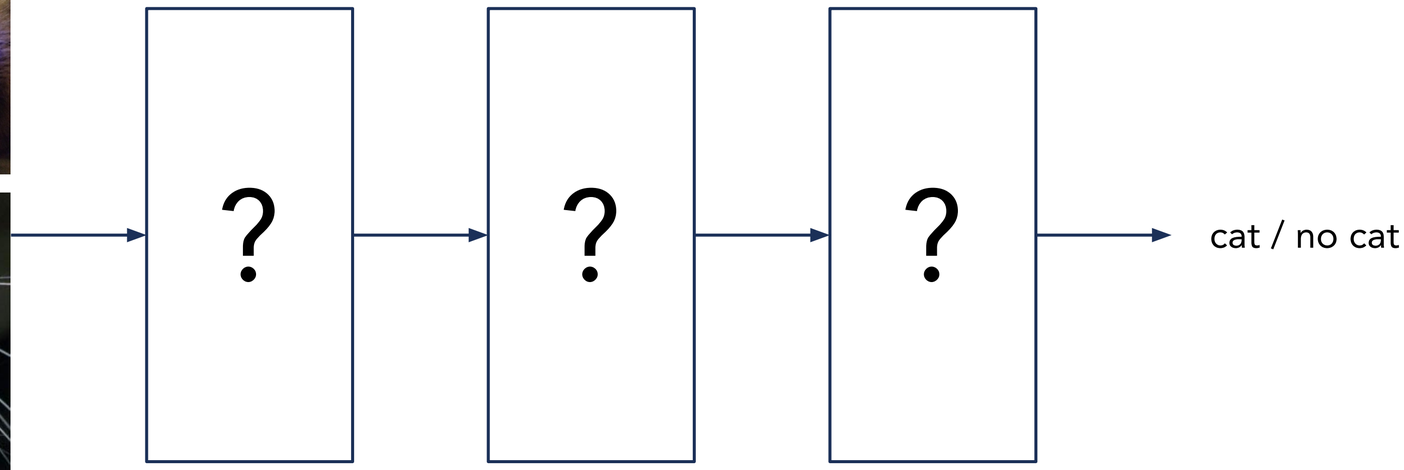
- representation of interest points/regions
- for image stitching or object recognition



Handcrafted Features

Task: find out automatically if image contains a cat
Hierarchy of features: ???

Define good and robust features for this task manually is way too hard.



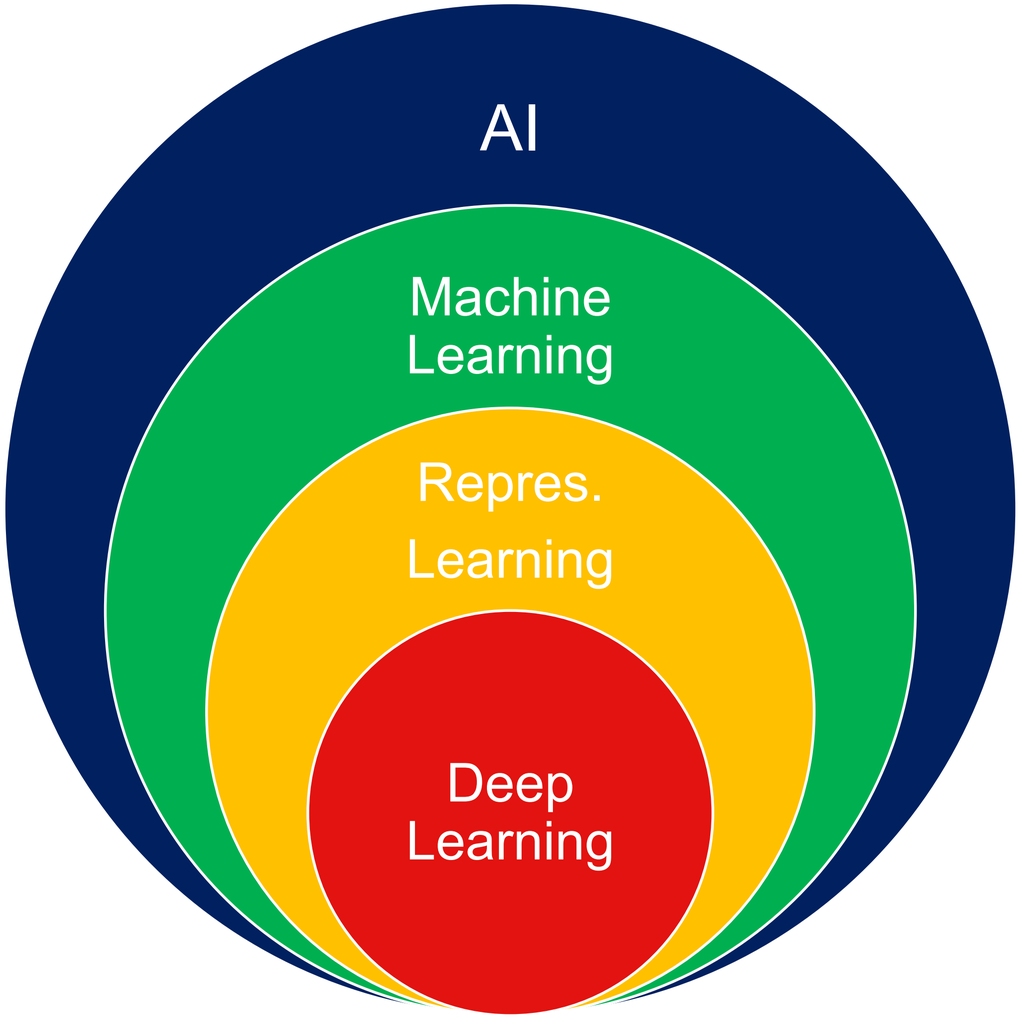
You LEARNT to recognize a cat by examples and experience. You are an expert cat-recognizer. What do you think are the elements that let you recognize a cat?

Deep Learning (from Nature)

nature International weekly journal of science

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

Yann LeCun, Yoshua Bengio & Geoffrey Hinton



Representation learning methods that

"allow a machine to be fed with raw data and to automatically discover the representations needed for detection or classification."

Deep-learning are representation learning methods

- with multiple levels of representation, obtained by
- composing simple but non-linear modules that each
- transform the representation at one level into a representation at a higher, slightly more abstract level.

Deep Learning & Artificial Neural Networks

Deep Learning Models are often Artificial Neural Networks:

- Loosely-inspired by biological learning in mammal brains
- An artificial neuron can learn to recognize a pattern
- Several neurons are organized in layers
- Each layer can be thought as a learnable feature extractor
- Several layers, one feeding on the output of the previous one, form an artificial neural network
- Given inputs and a desired outputs, in the training phase neurons adapt to align the network output to the desired one

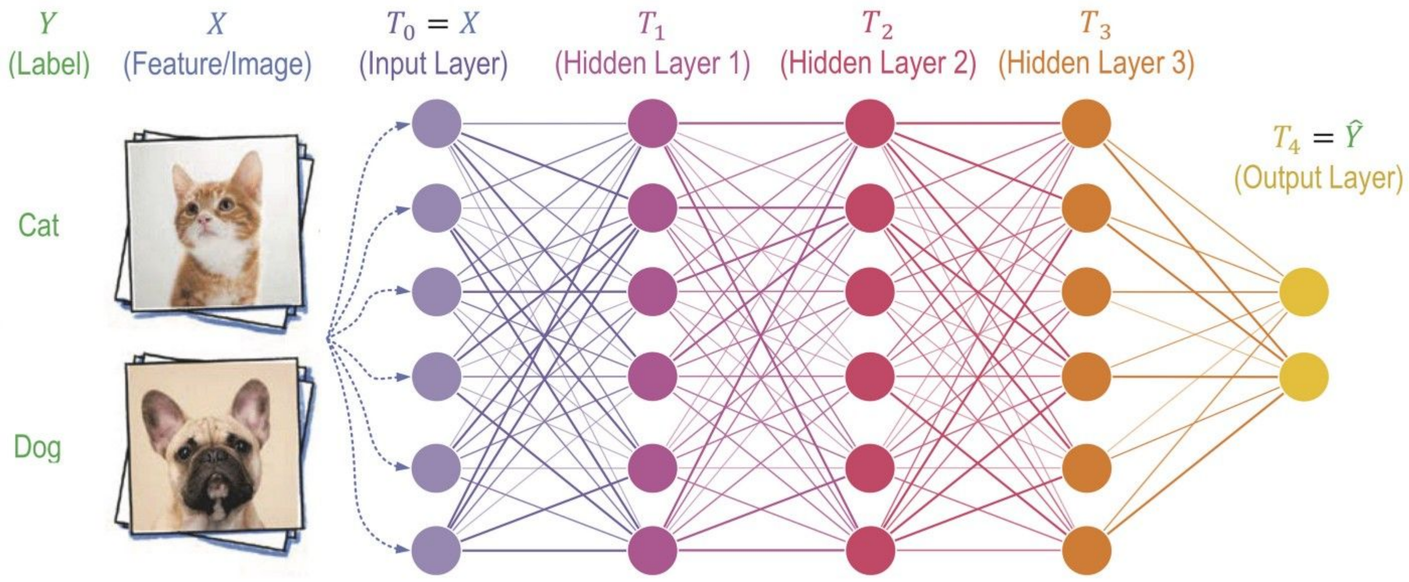
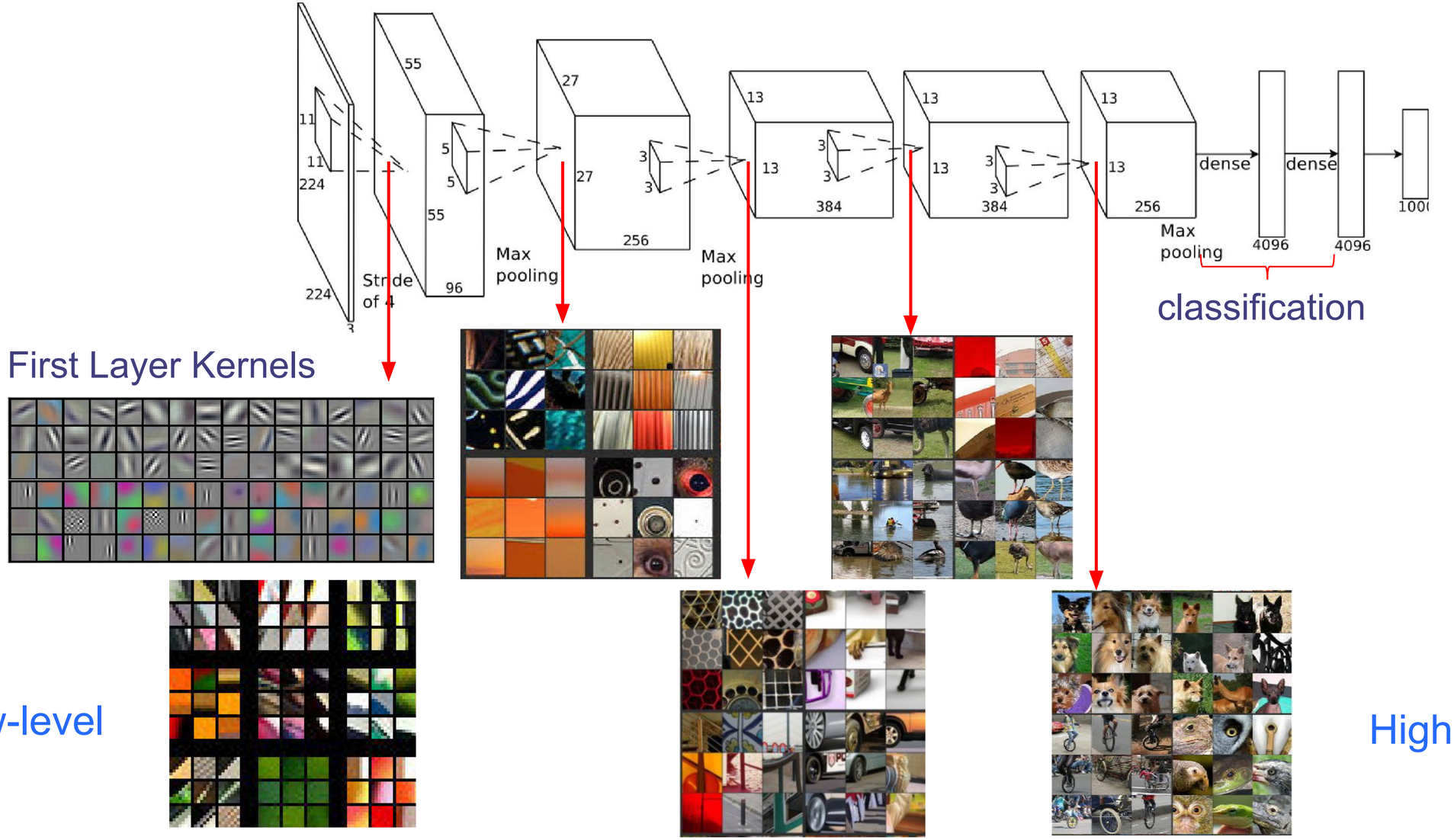


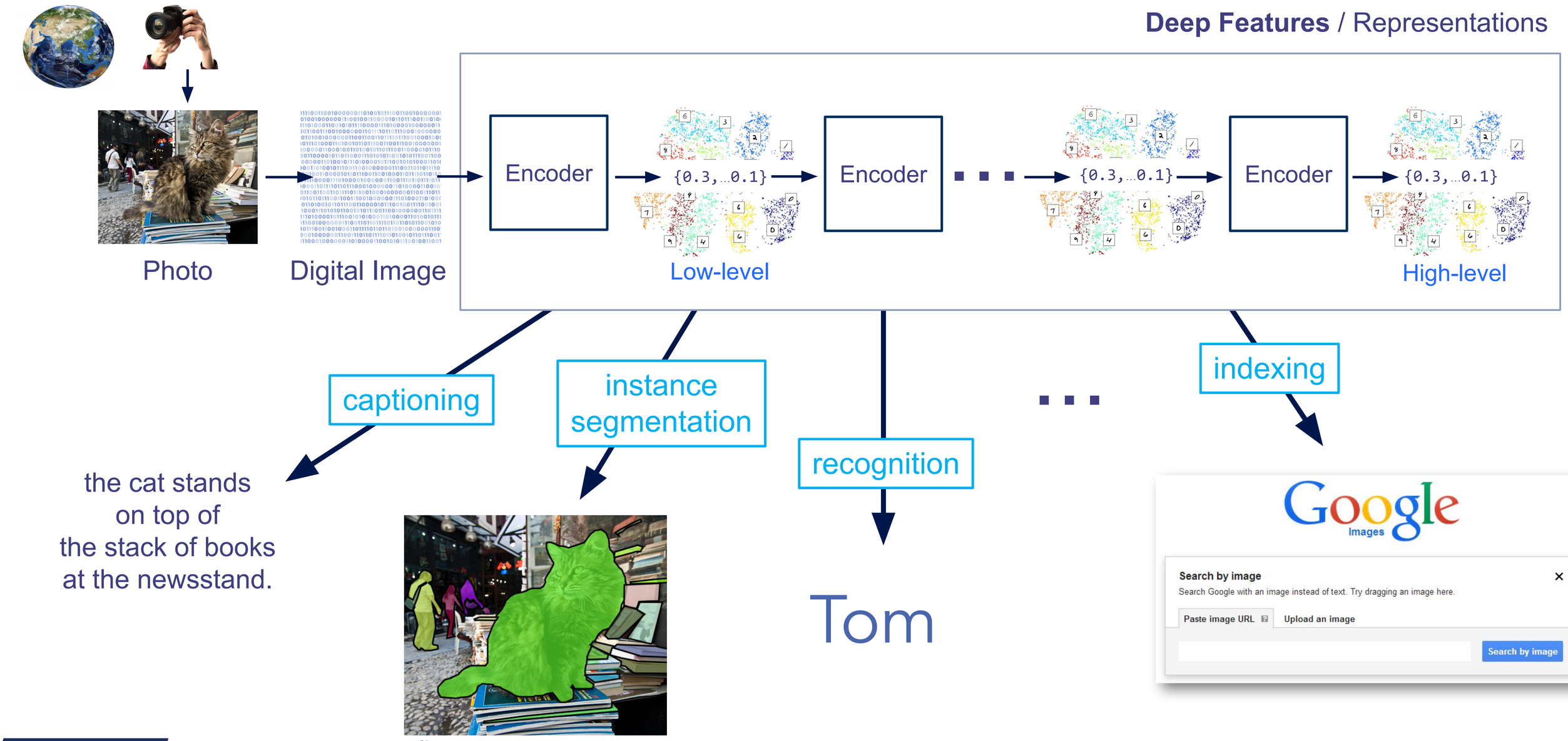
Image credits: https://medium.com/@MITIBMLab/estimating-information-flow-in-deep-neural-networks-b2a77bdda7a7

Multiple Levels Of Abstraction

AlexNet, 2012, Trained on a Classification task of 1,000 classes.

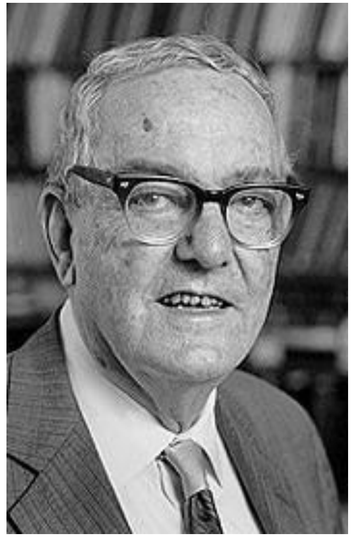


The Overall Picture



Intuition:
Why we use Deep Learning Features?

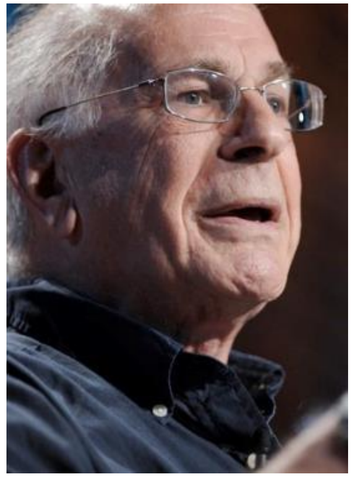
Intuition is Recognition



"Intuition is nothing more and nothing less than **recognition**"

Herbert Simon, Turing Award 1975 and the Nobel 1978

Simon defined intuition as the recognition of patterns stored in memory.



"There is really no difference between the physician recognising a particular disease from a facial expression and a little child learning, pointing to something and saying doggie. The little child has no idea what the **clues** are but he just said, **he just knows** this is a dog without knowing why he knows".

Daniel Kahneman, Nobel Prize 2002

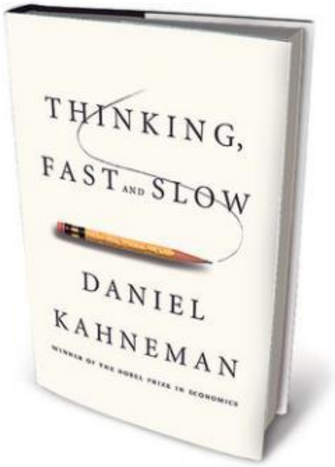
Thinking, Fast and Slow

Daniel Kahneman

- Psychologist
- Nobel Prize in Economic Sciences in 2002 (shared with Vernon L. Smith)

for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty

Thinking, Fast and Slow (2011)




HEAT
ENERGY

What did you see?

What is she going to do?

Thinking Fast

As surely and quickly as you saw that

- the young woman's hair is dark, you knew she is angry.

What you saw extended into the future.

- You <u>sensed</u> that this woman is about to say some very unkind words, probably in a loud and strident voice.

Thinking Fast

You did not intend

- to assess her mood, or
- to anticipate what she might do.

Your reaction to the picture did not have the feel of something you did.

It just happened to you.

It was an instance of fast thinking.

Thinking, Fast and Slow

$$17 \times 24 = ?$$

What came to your mind?

123

586

12.609

Thinking Slow

17x24=

- this is a multiplication problem
- you knew that you could solve it
- you would be quick to recognize that both 12,609 and 123 are implausible
- you would not be certain at first that answer is not 568

Thinking Slow

A precise solution
did not come to mind

You felt you could choose
whether or not
to engage in the computation

Thinking, Fast and Slow

If you engaged the computation, you proceeded through a sequence of steps.

you retrieved from memory
the cognitive program
then you implemented it

Carrying out the computation was a strain.

You felt the burden of:

- holding much material in memory,
- as you needed to keep track of
  where you were and of where you were going,
- while holding on to the intermediate result.

Thinking, Fast and Slow

Kahneman describes two different ways the brain forms thoughts:

- System 1: Fast, automatic, frequent, emotional, stereotypic, subconscious
- System 2: Slow, effortful, infrequent, logical, calculating, conscious

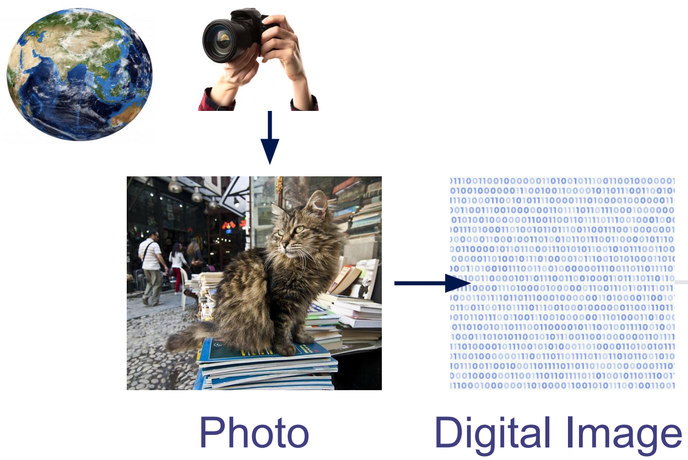
Deep Learning ≈ Thinking Fast

Thinking, Fast and Slow

$$17 \times 24 = 408$$

From Light to Bits

From Light to Bits



Deep Features / Representations

captioning

the cat stands
on top of
the stack of books
at the newsstand.

instance segmentation

recognition

Tom

…

indexing

From Light to Bits

photography

digitalization

(compression)

Acquisition Device
(Camera, Scanner)

Photo

Bits

Exposure Time / Shutter Speed

- The amount of time the image sensor is exposed to light
- Side effects with moving objects

Aperture

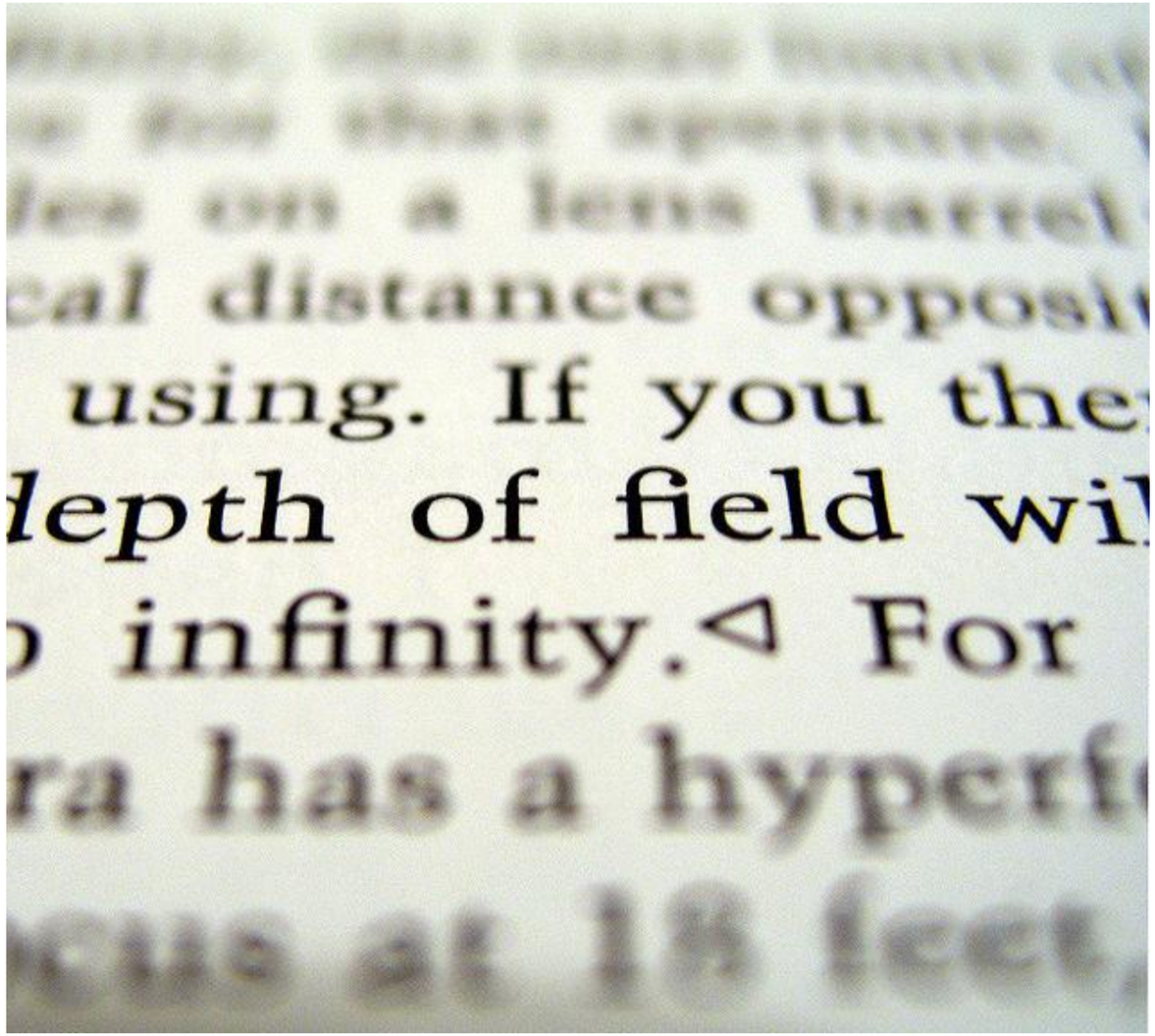
- Opening through which light travels
- Affects the depth of field

Aperture / Depth of fields

- Same flowers with different apertures resulting in distinct depth of fields

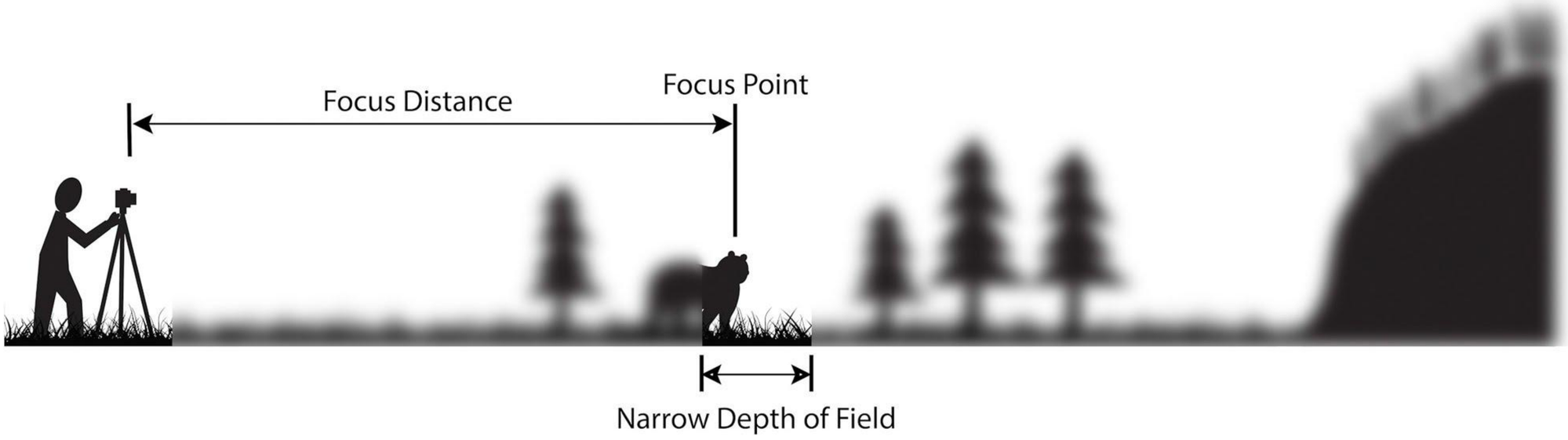
Depth of Field

- The distance between the nearest and farthest objects in a scene that appear acceptably sharp in an image
- Precise focus is possible at only one distance; at that distance, a point object will produce a point image.
- When this circular spot is sufficiently small, it is indistinguishable from a point, and appears to be in focus; it is rendered as "acceptably sharp".

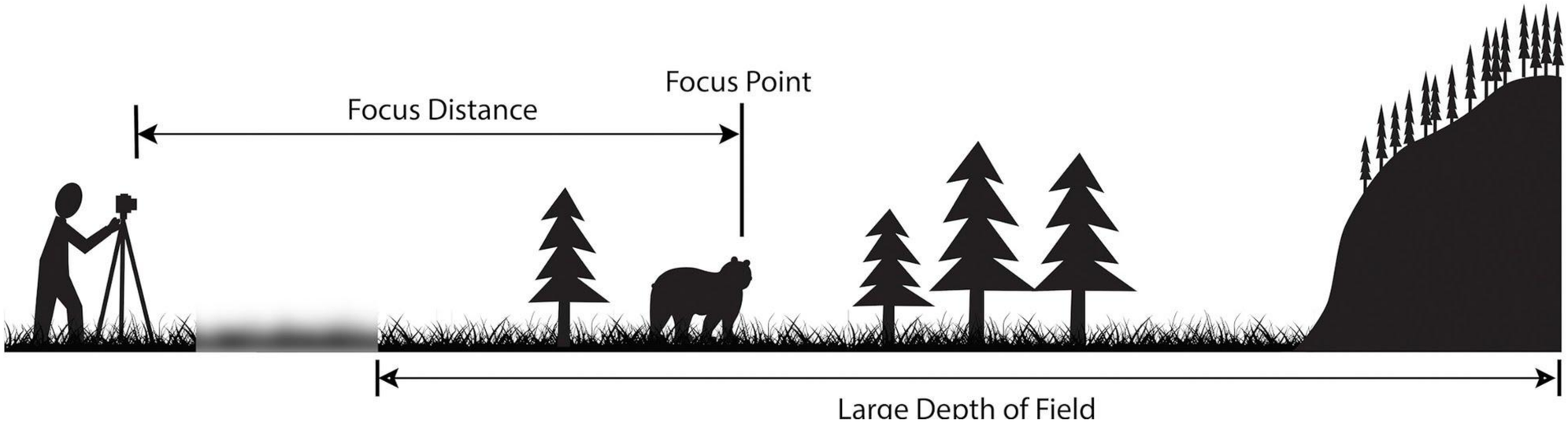


Depth of Field

Big Aperture, Narrow Depth of Field



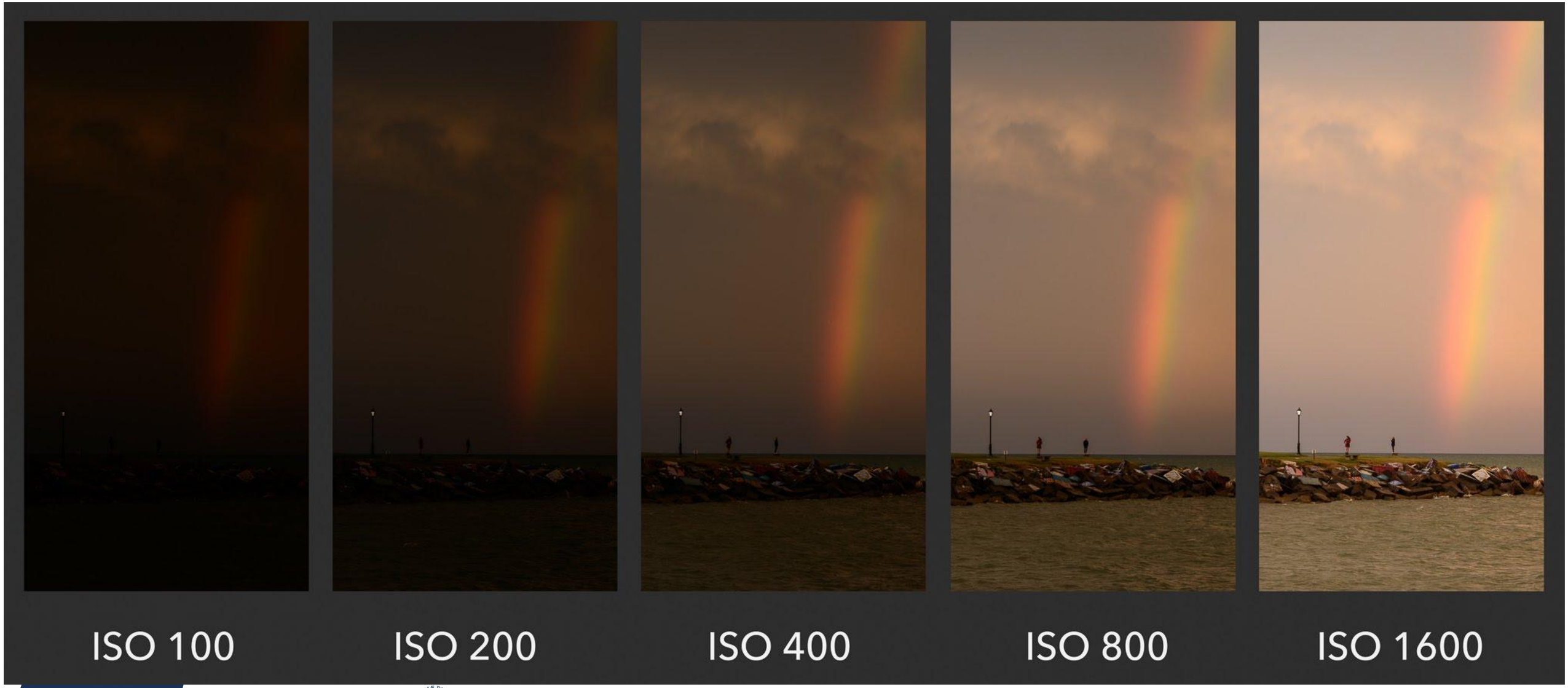
Small Aperture, Large Depth of Field



https://photographylife.com/what-is-depth-of-field

ISO

- In Digital Photography ISO measures the sensitivity of the image sensor.



https://photographylife.com/what-is-iso-in-photography

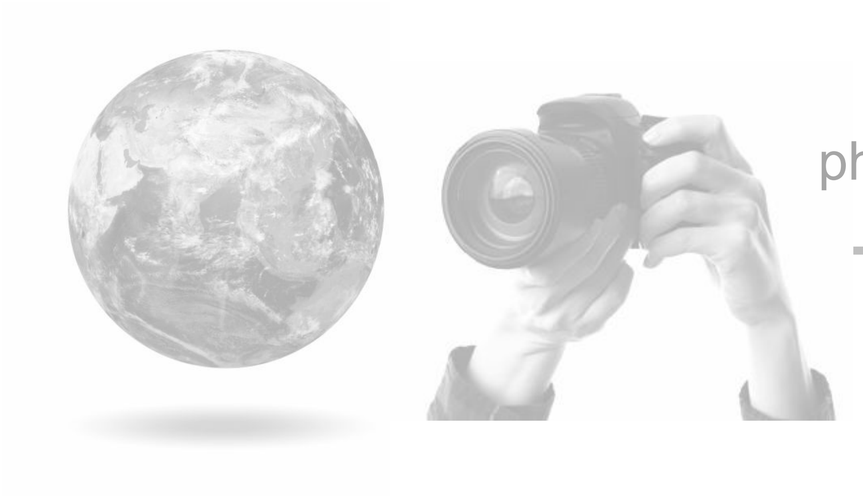
https://www.exposureguide.com/iso-sensitivity/

Low ISO sensitivity, slow shutter speed.

High ISO sensitivity, fast shutter speed.

https://imaging.nikon.com/lineup/dslr/basics/13/index.htm

From Light to Bits



photography

digitalization

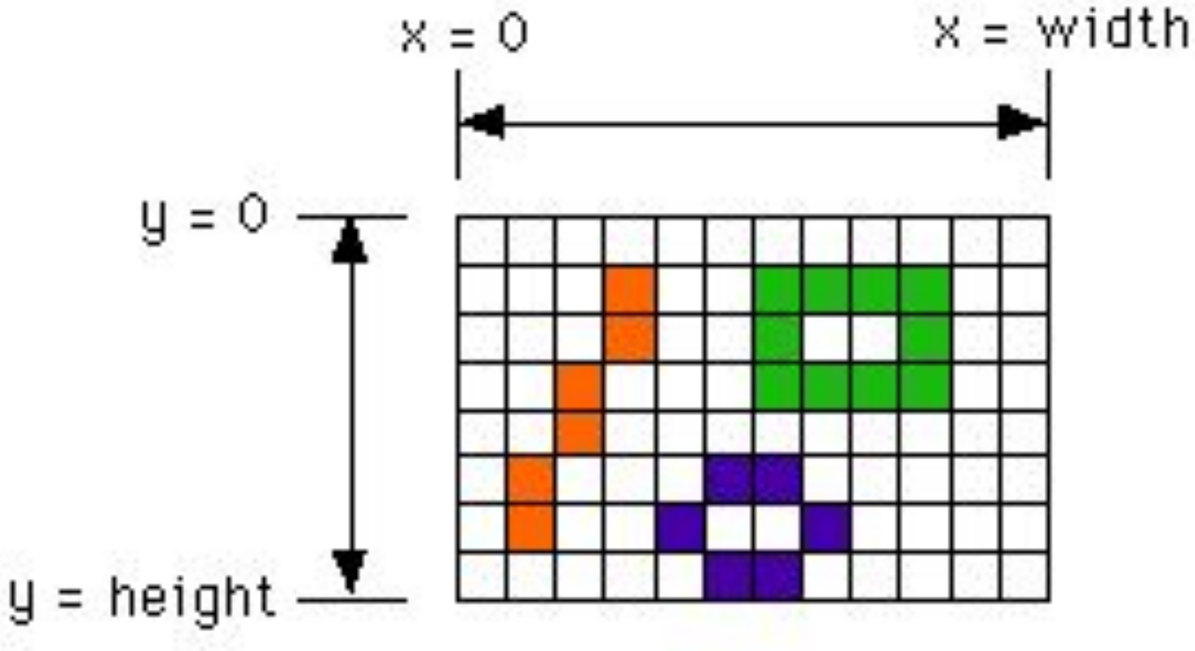
(compression)

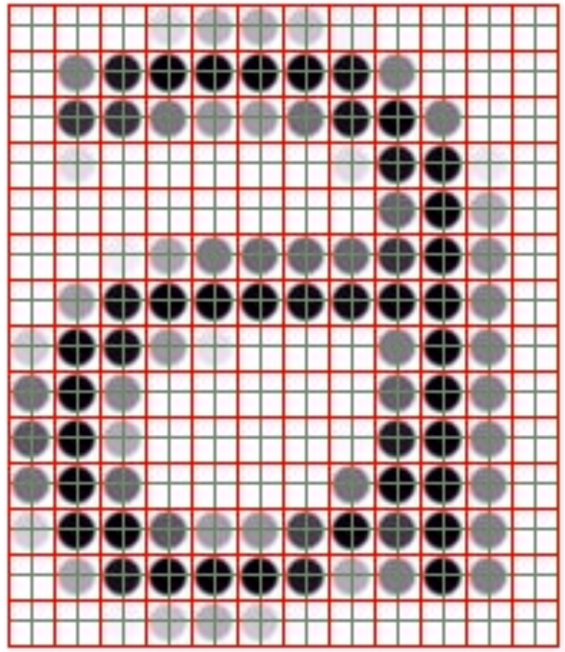
Camera

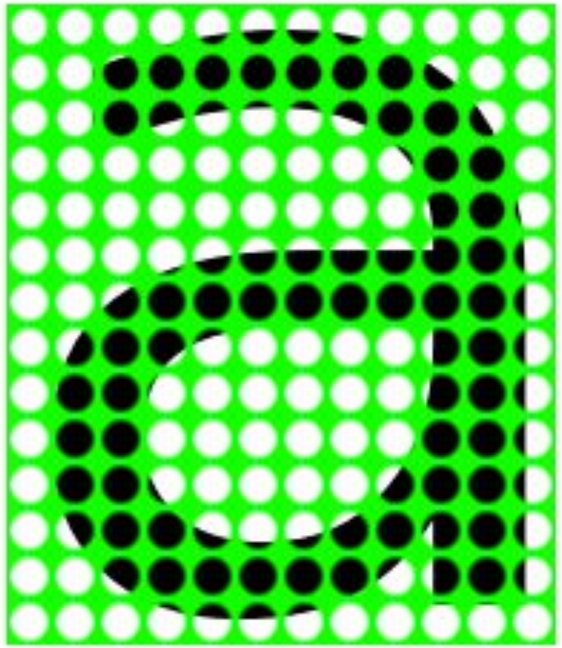
Photo

Bits

Still Images



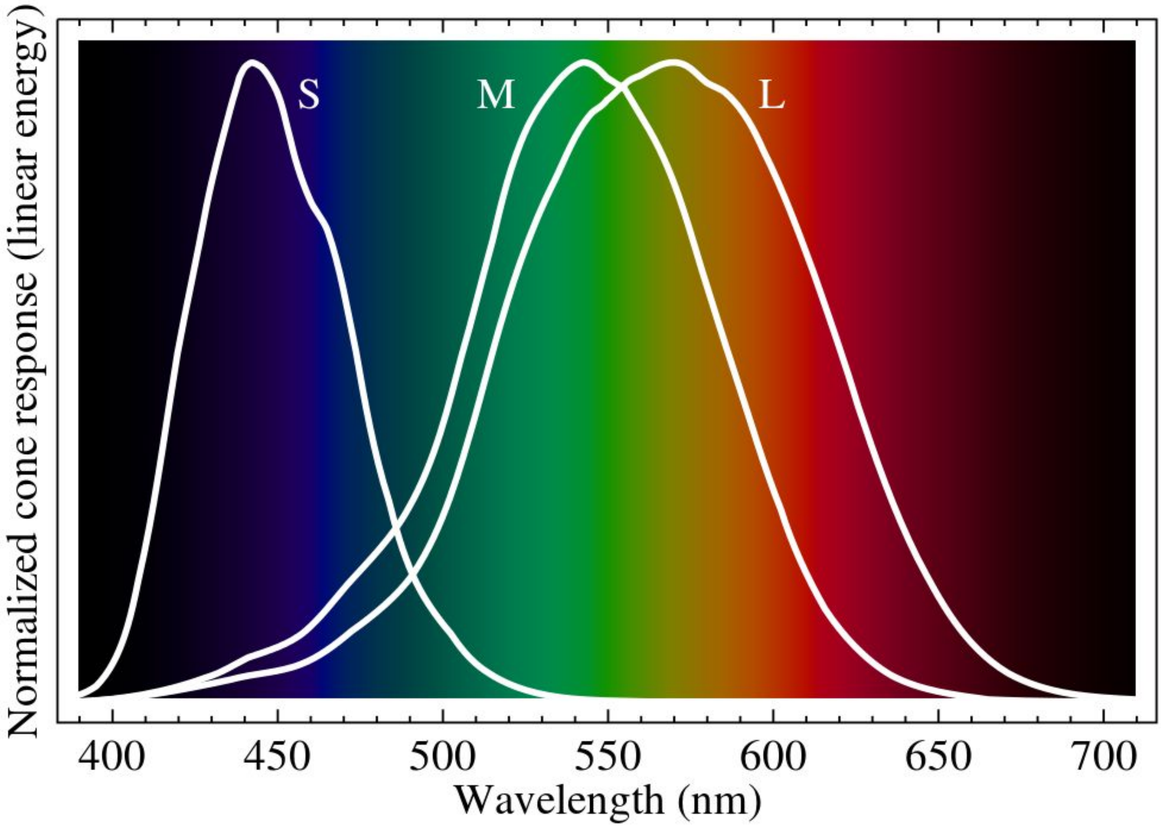




From the GIMP software documentation

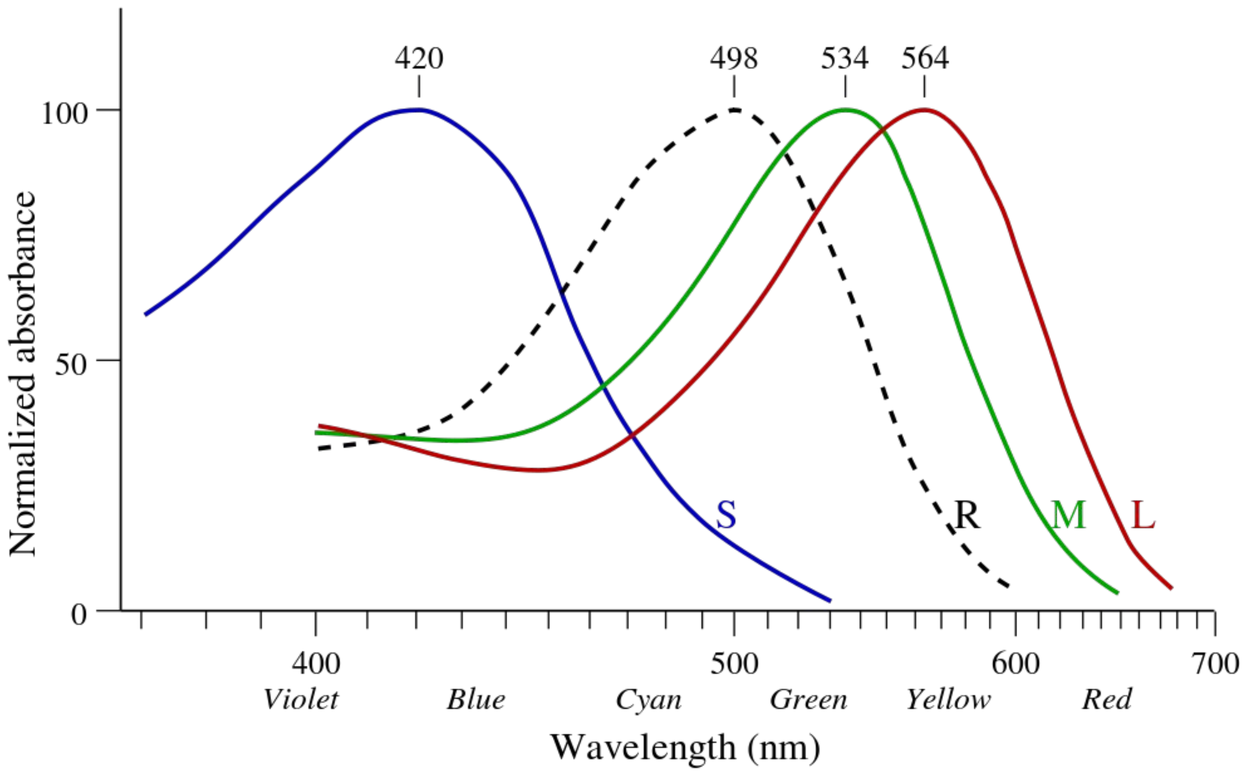
Trichromacy

- Humans have 3 types of sensors (cones): L, M and S.
- Colors are perceived by the interaction of at least 2 types of cones
- N. of perceived color between 1 and 10 mln



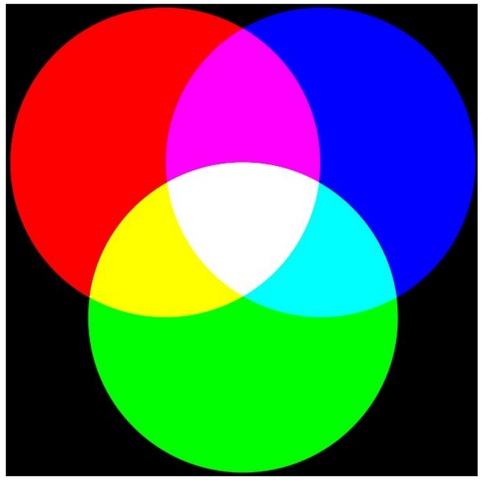
Rod Cells

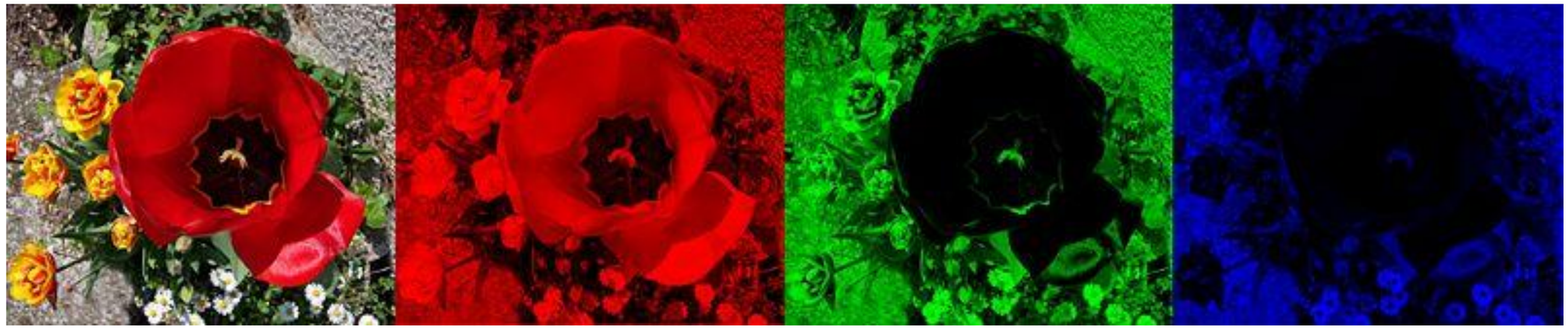
- Are in the retina of the eye
- Are sensitive to less intense light than the others (black dotted line)
- Concentrated on the outer edges
- Used in peripheral vision and in low light



RGB

- Additive color model
- Based on human perception
- R, G, and B levels vary between devices
- Color management is needed
- Color spaces are used to ensure consistency

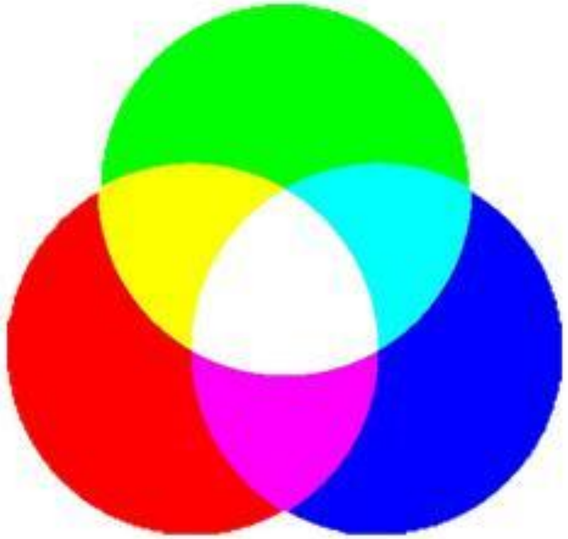




Additive / Subtractive

RGB
Additive
Color

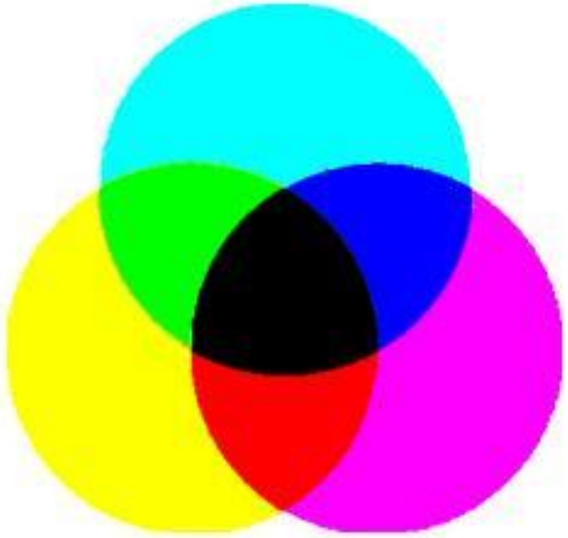
mixing light



RED GREEN BLUE

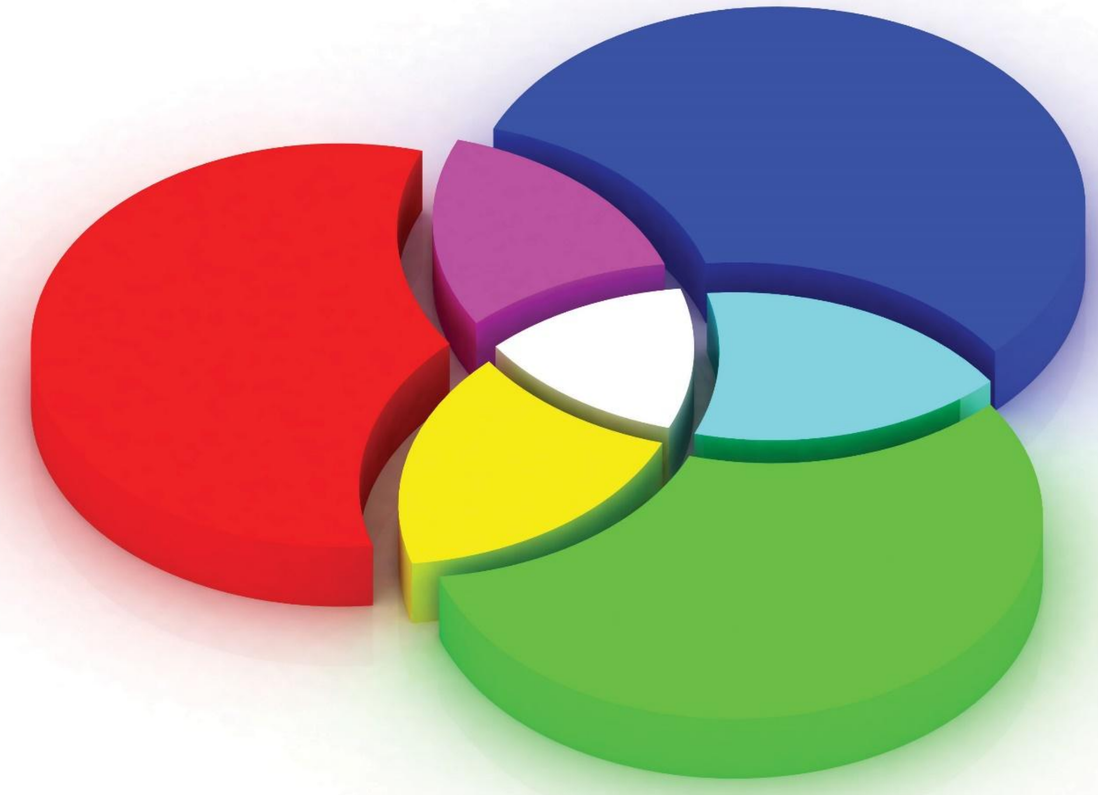
CMYK
Subtractive
Color

mixing ink

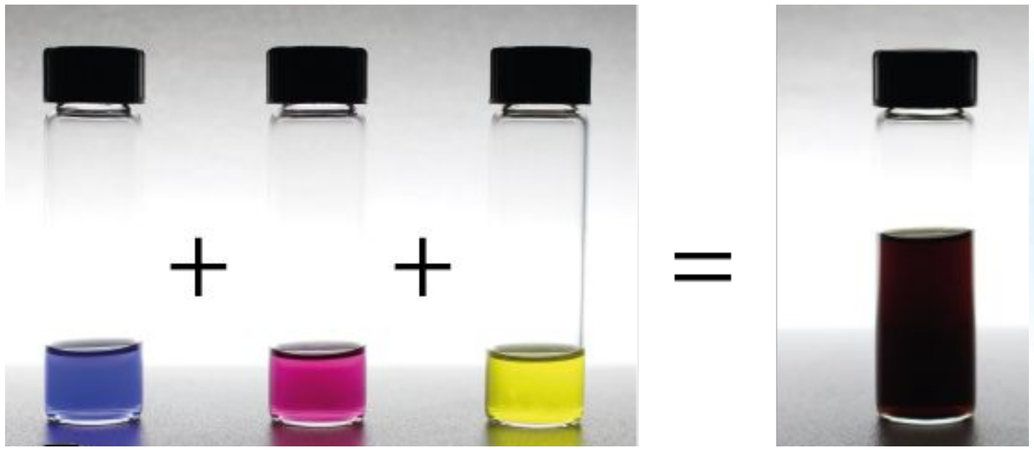


CYAN MAGENTA YELLOW

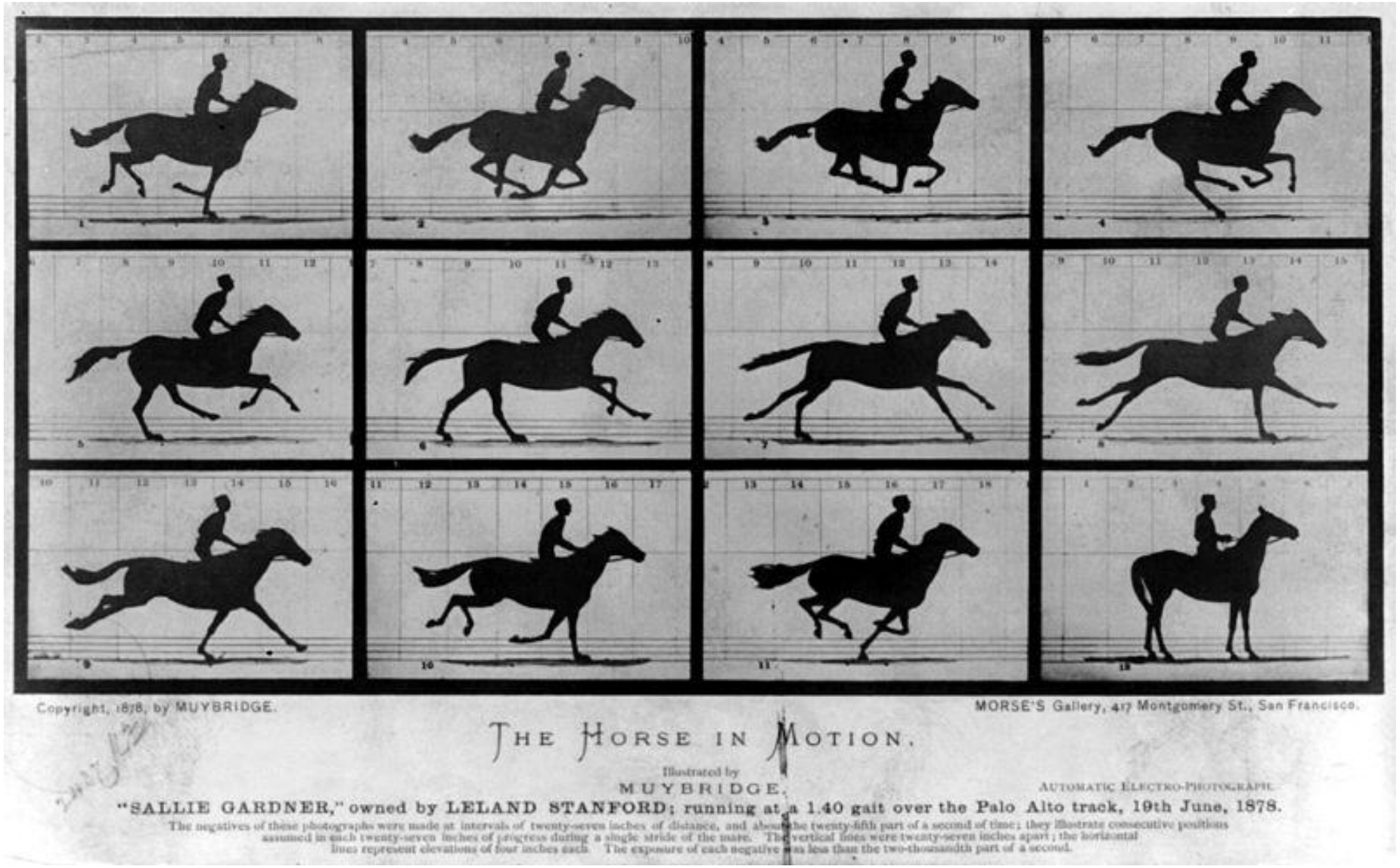
Additive Color Mode



Subtractive Color Mode



1878, Muybridge

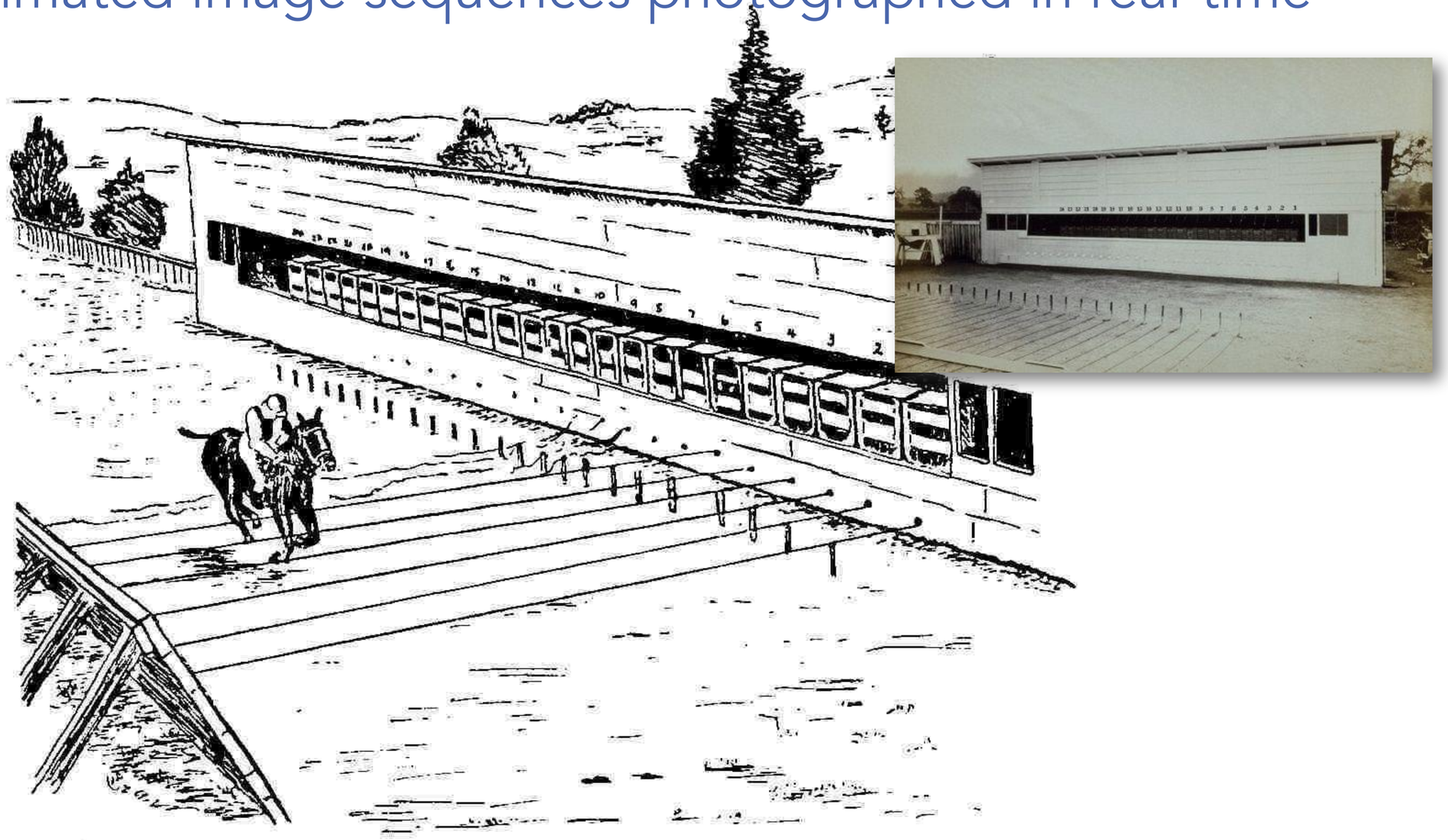


1878: The horse in motion

first animated image sequences photographed in real-time

Moving pictures

- first animated image sequences photographed in real-time

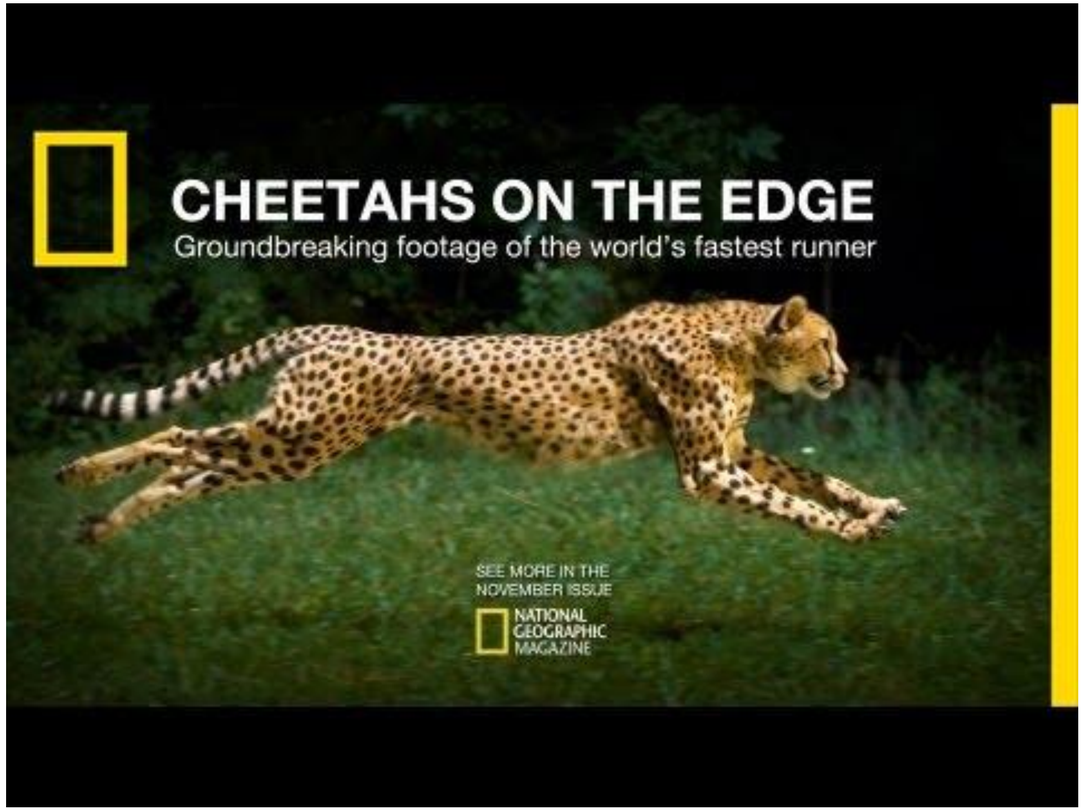


Reel

How many frames per seconds?

- Frame Per Seconds (FPS):
  24: traditional 35 mm sound film starting from 1930
  25: PAL (EU TV)
  29.97: NTSC
  48: The Hobbit: An Unexpected Journey
  (accused to breaks the suspension of disbelief)
  50/60: HDTV
  72: experimental
  90/100: GoPro
  120: UHDTV
  144/240: Gaming monitors
  300: Tested by BBC for sports broadcasts

FPS: recording vs display



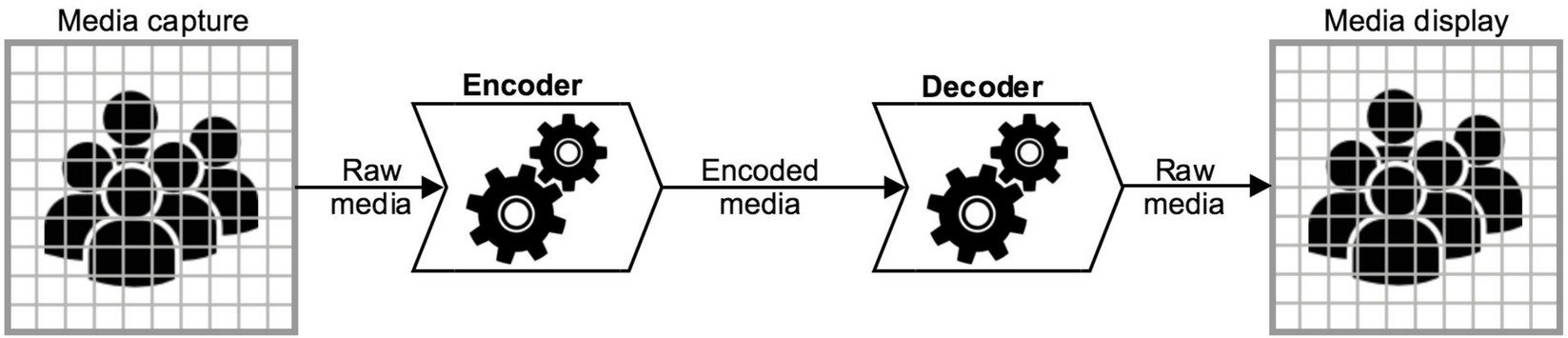
FPS Recording > FPS Play
slow motion video

FPS Play > FPS Recording
time lapse

Storing and Sharing Images

Codec

- Codec = coder-decoder (or compressor-decompressor)
- Software for encoding or decoding a data stream



- Lossy codecs (Quality vs Compression):
  - Mostly based on human perception, trying to achieve better human perceived quality at a given predefined compression
  - Examples: MPEG-2, MPEG-1 and 2 layer III (MP3), AAC, MPEG-2, H.264, DivX
- Lossless codecs (Original data can be perfectly reconstructed)
  - Mostly based on statistical modelling
  - Examples: LZW, FLAC, PNG, TIFF, etc.
  - Transcoding between lossy formats results in loss of data

Image File Formats

IMAGE FORMAT	AVAILABLE COLORS	COMPRESSION	FILE SIZE	BEST FOR
RAW	Billions	No	Very big (<10MB)	Editing
JPEG	16,1 million	Lossy	Small (<1MB)	Websites and storage
GIF	256	Lossless	Small (<1MB)	Animation
PNG	16,1 million + transparency	Lossless	Big (<3MB)	Websites, editing, storage
TIFF	Variable	Variable	Big (<3MB)	Editing and printing
BMP	Variable	Lossless	Big (<3MB)	-

Sharing Image Collections

- IIIF - International Image Interoperability Framework (https://iiif.io)

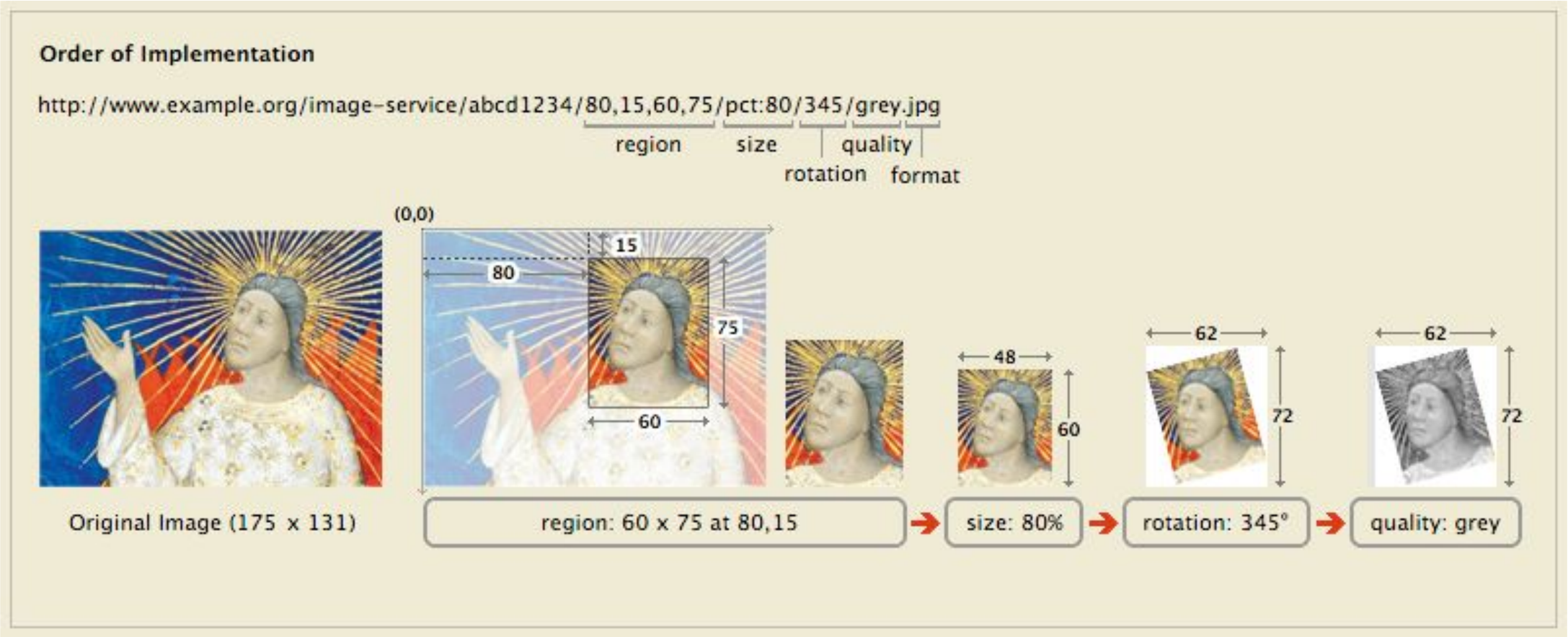
- Set of open standards for delivering high-quality, attributed digital objects (images, audio/visual) online at scale.

- Born to "facilitate systematic reuse of image resources in digital image repositories maintained by cultural heritage organizations."

International
Image
Interoperability
Framework

IIIF Image API

. Define how image data can be served and accessed (URL-based API)



IIIF Presentation API

- Provides the information necessary to allow a rich, online viewing environment for compound digital objects to be presented to a human user.
- Manifest file in JSON format containing:
  - Descriptive metadata like labels, rights and other information
  - Links to Images and AV resources
  - Ordering of Images in sequences and table of contents
- Demos

  https://uv-v3.netlify.app/

  https://demos.biblissima.fr/chateauroux/osd-demo/

More on IIIF

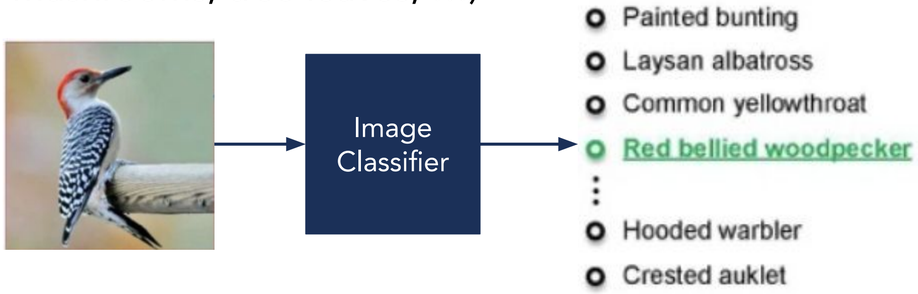
- https://iiif.io/get-started/
- https://training.iiif.io/
- IIIF-compliant image servers: https://iiif.io/get-started/image-servers/
- IIIF-compliant image viewers: https://iiif.io/get-started/iiif-viewers/

Image Classification

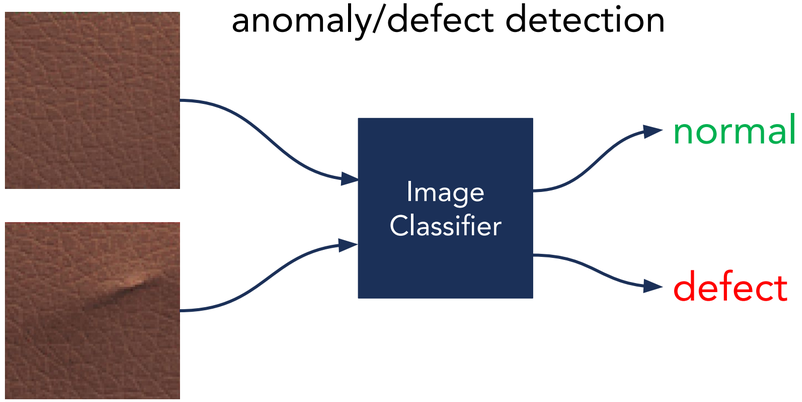
Image Classification

- Automatically assign an image to categories or classes of interest depending on its visual content only. (No metadata available).
- Several problems can be framed as image classification. E.g.:

fine-grained animal/object recognition (birds, mushrooms, tree leaves, …)

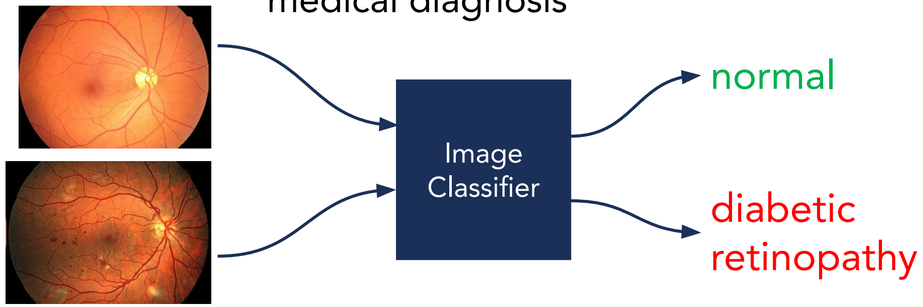


anomaly/defect detection

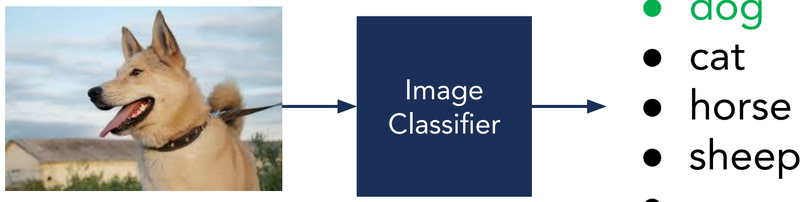


and many more….

medical diagnosis



generic object recognition

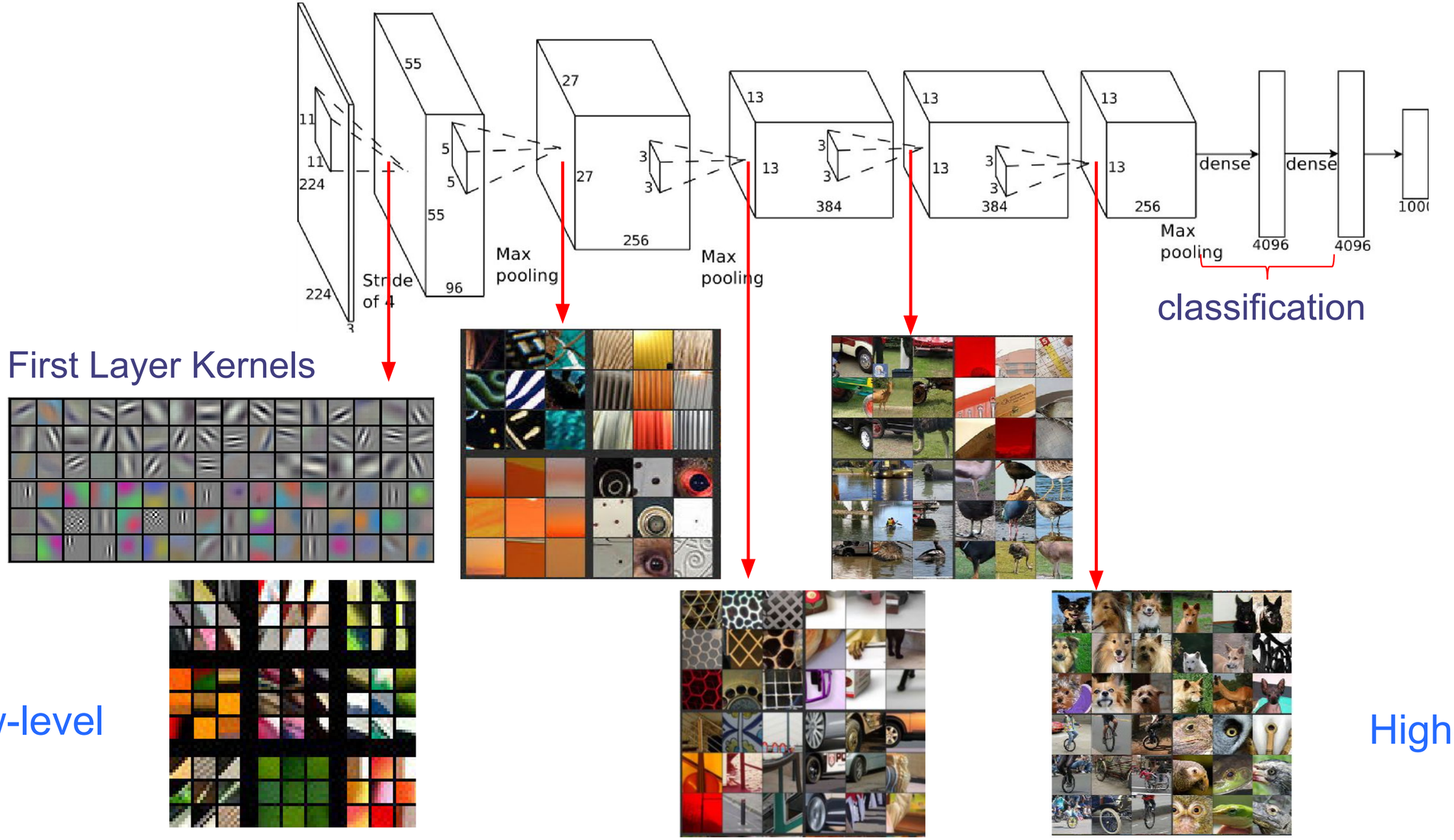


Deep Learning for Image Classification

- Understand relationships between input images and categories is often difficult to do manually (hand-crafted features).
- Most solutions use Deep Learning (specifically, artificial neural networks) to learn the mapping between image and category.
- Networks must be trained on a set of images to learn the specific mapping; neurons change connections to learn patterns during training.
- Once trained, the network is frozen and used to classify new images.

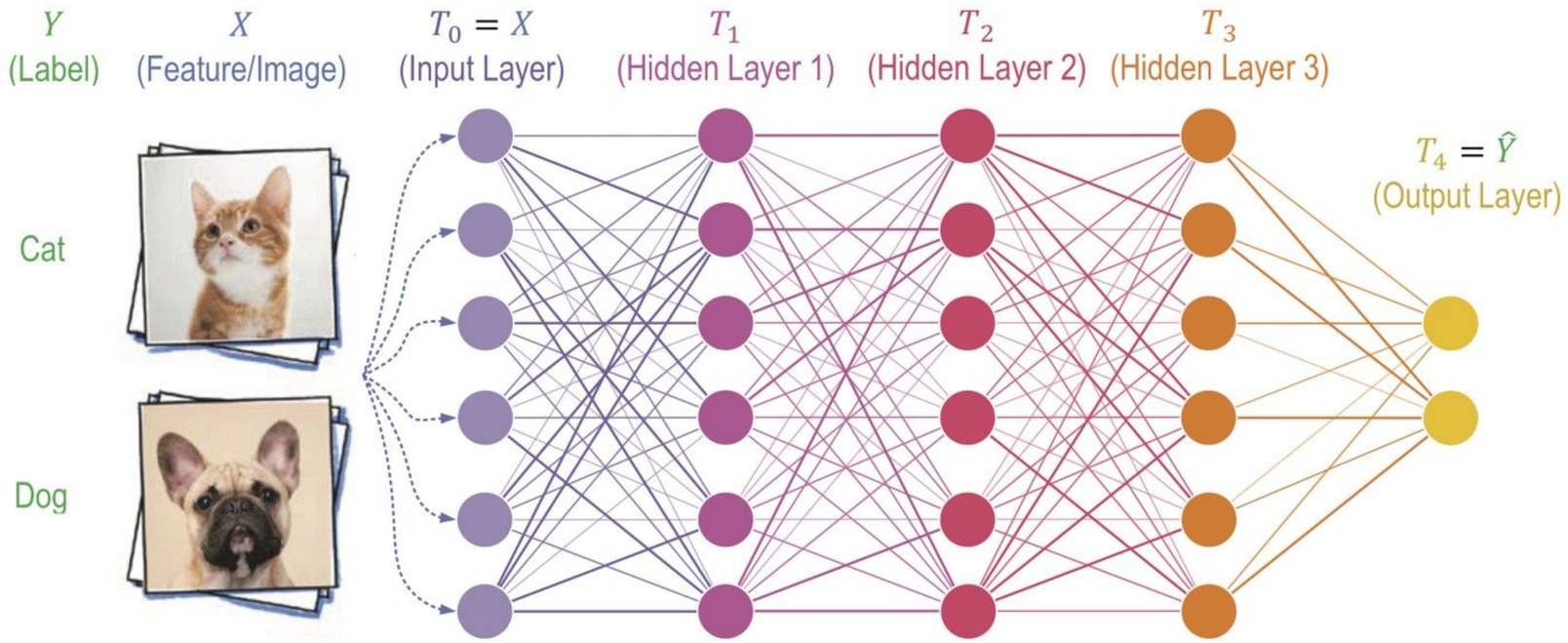
Multiple Levels Of Abstraction

AlexNet, 2012, Trained on a Classification task of 1,000 classes.



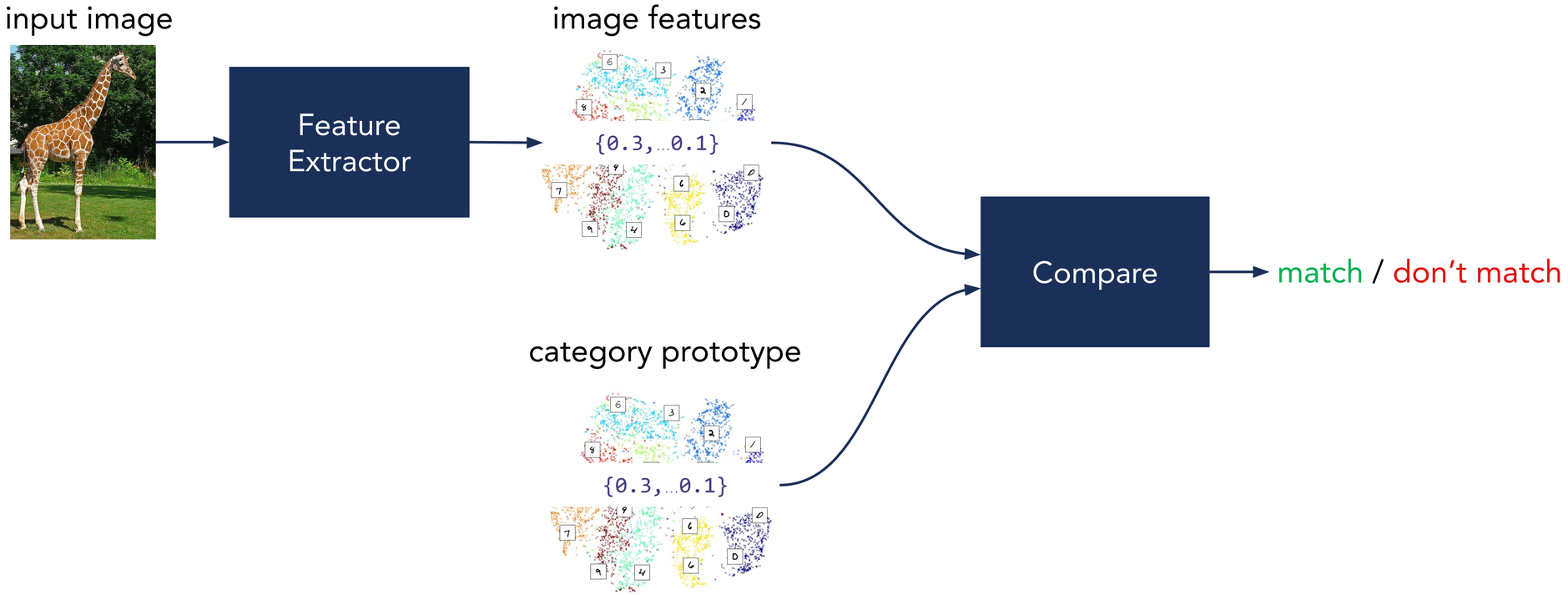
Closed-set vs Open-set Classification

- Closed-set classification
  - We want to classify the input image into one of N predefined categories.
  - The classifier guesses the best one out of the N categories.
  - The classifier is trained on examples belonging to the N categories.



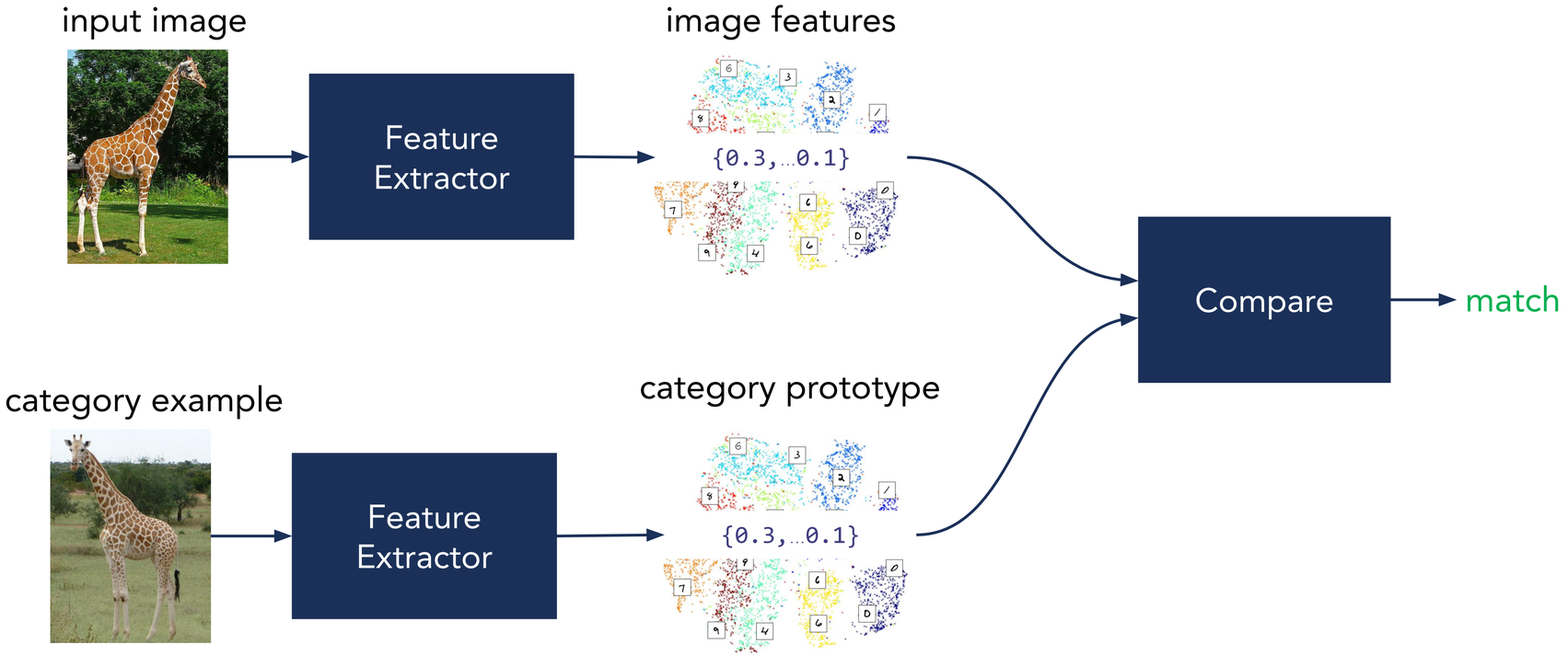
Closed-set vs Open-set Classification

- Open-set classification
  - We do not have a predefined set of categories
  - The classifier is trained to extract generic image features/representations (a string of numbers!)
  - Image features are compared with novel category "prototypes" to check if they match



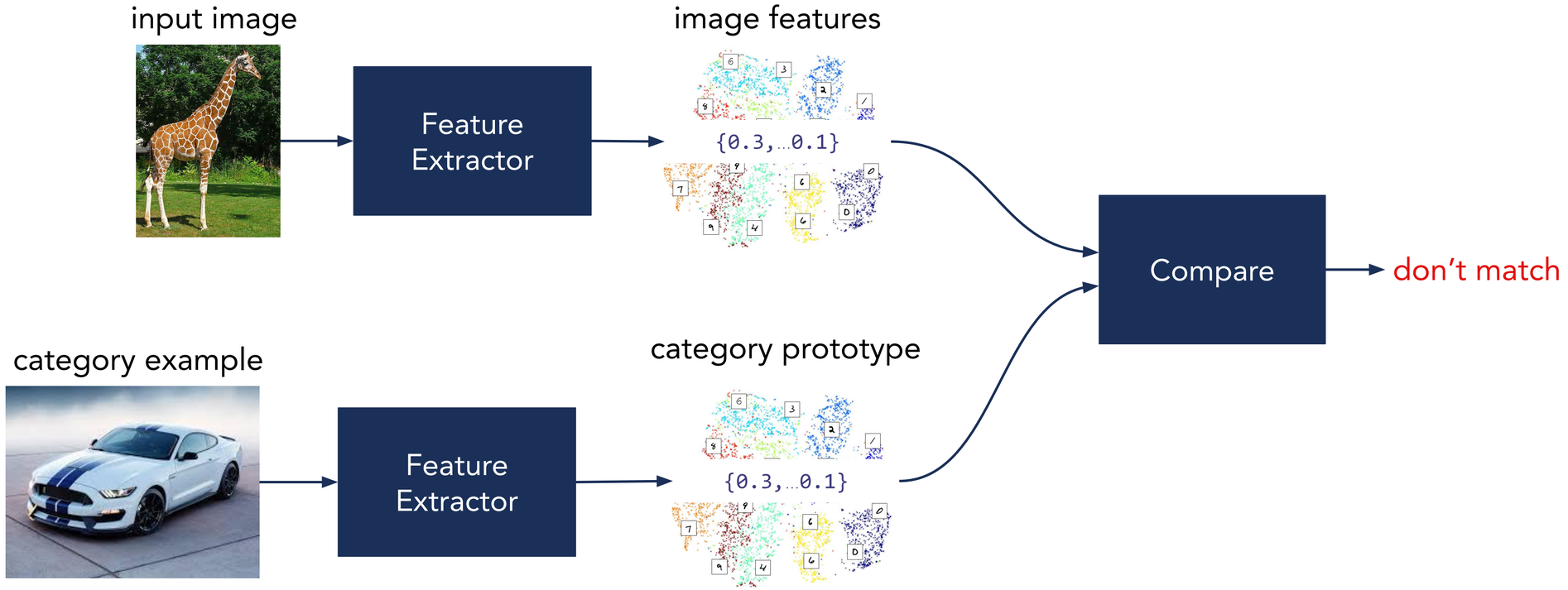
Closed-set vs Open-set Classification

- Open-set classification
  - We do not have a predefined set of categories
  - The classifier is trained to extract generic image features/representations (a string of numbers!)
  - Image features are compared with novel category "prototypes" to check if they match



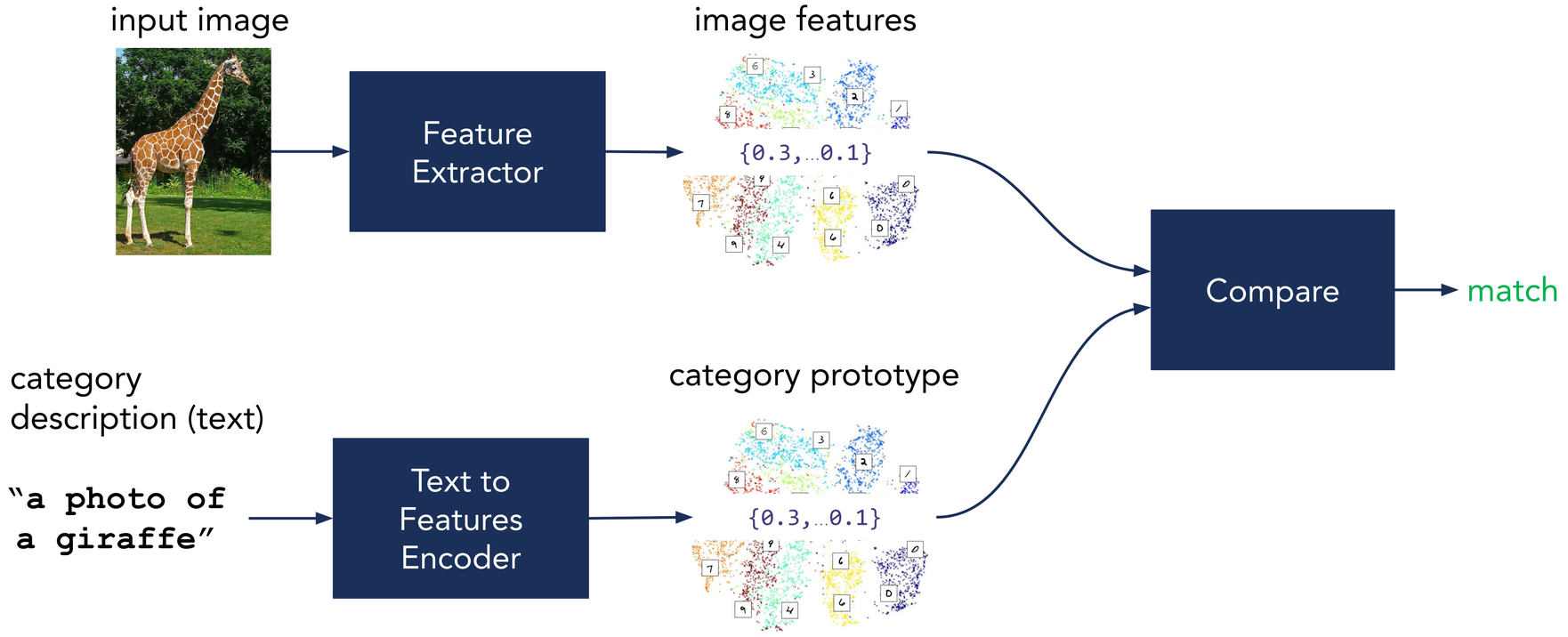
Closed-set vs Open-set Classification

- Open-set classification
  - We do not have a predefined set of categories
  - The classifier is trained to extract generic image features/representations (a string of numbers!)
  - Image features are compared with novel category "prototypes" to check if they match



Closed-set vs Open-set Classification

- Open-set classification
  - We do not have a predefined set of categories
  - The classifier is trained to extract generic image features/representations (a string of numbers!)
  - Image features are compared with novel category "prototypes" to check if they match



Consiglio Nazionale delle Ricerche

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

Image Retrieval

Nicola Messina

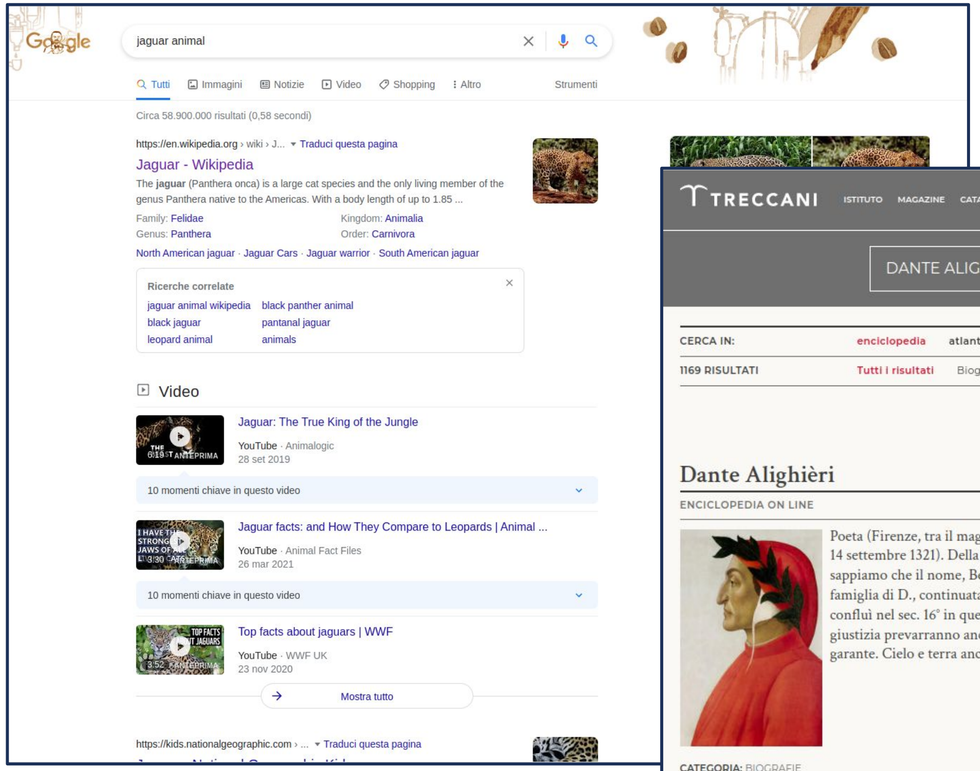
- M.Sc. Computer Engineering @ UniPi (Feb. 2018)
- PhD in Information Engineering @ UniPi (May 2021)
- PostDoc @ Istituto di Scienza e Tecnologie dell'Informazione (ISTI) - CNR
  - AIMH (Artificial Intelligence for Media and Humanities) Lab

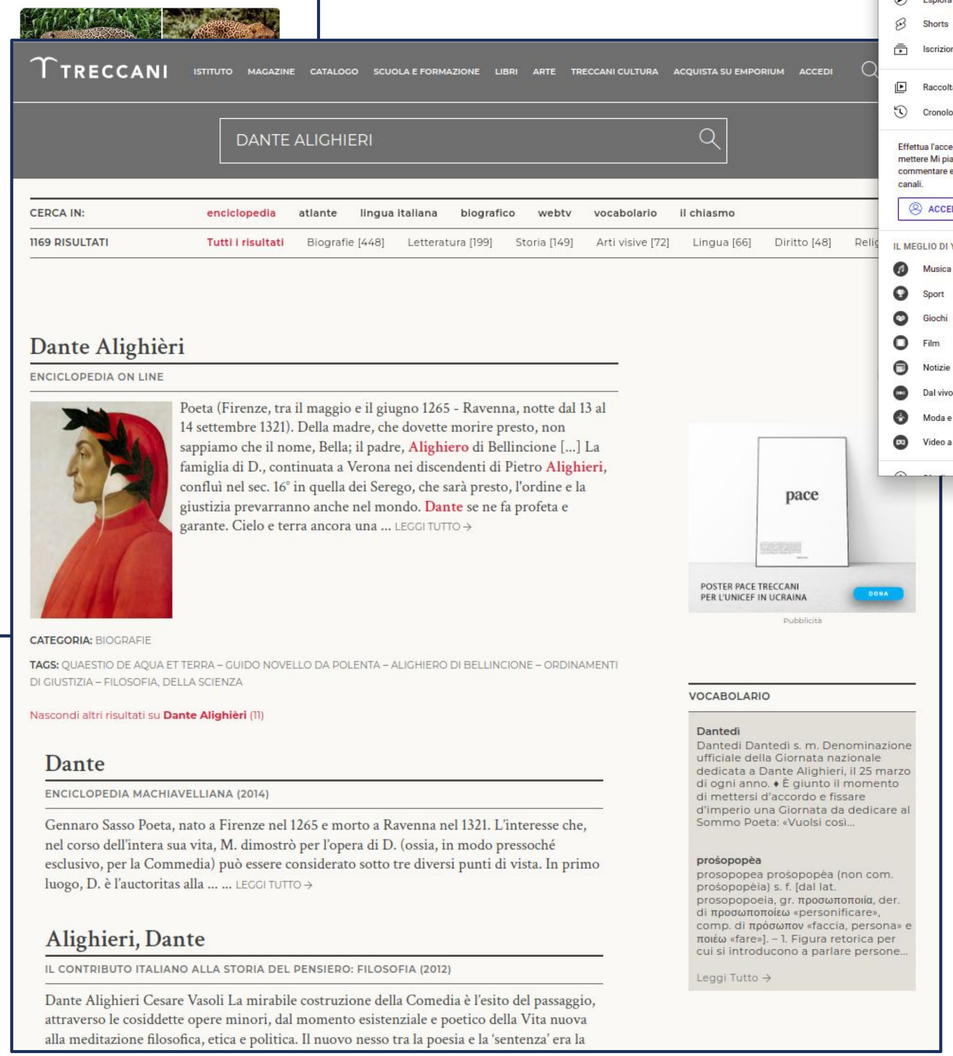
nicola.messina@isti.cnr.it

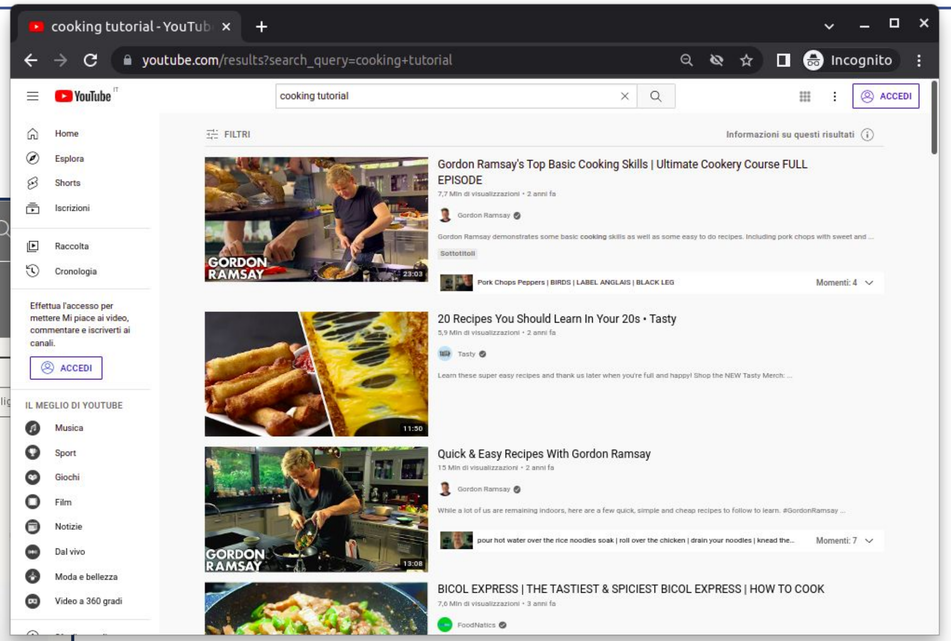
Outline

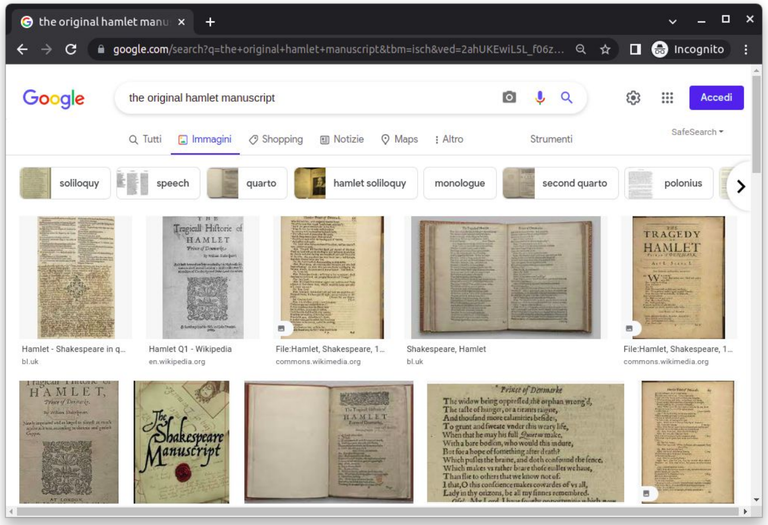
- The Multimedia Information Retrieval. Why do we care?
- A brief look into textual retrieval (and his limitations)
- Image Retrieval
- Image Representations
- Deep Learning to obtain powerful representations
- Text-to-image retrieval

Motivation


jaguar animal
Tutti
Immagini
Notizie
Video
Shopping
Altro
Strumenti
Circa 58.900.000 risultati (0,58 secondi)
https://en.wikipedia.org › wiki › J... Traduci questa pagina
Jaguar - Wikipedia
The jaguar (Panthera onca) is a large cat species and the only living member of the genus Panthera native to the Americas. With a body length of up to 1.85 ...
Family: Felidae
Kingdom: Animalia
Genus: Panthera
Order: Carnivora
North American jaguar · Jaguar Cars · Jaguar warrior · South American jaguar
Ricerche correlate
jaguar animal wikipedia
black panther animal
black jaguar
pantanal jaguar
leopard animal
animals
Video
Jaguar: The True King of the Jungle
YouTube · Animalogic
28 set 2019
10 momenti chiave in questo video
Jaguar facts: and How They Compare to Leopards | Animal ...
YouTube · Animal Fact Files
26 mar 2021
10 momenti chiave in questo video
Top facts about jaguars | WWF
YouTube · WWF UK
23 nov 2020
Mostra tutto
https://kids.nationalgeographic.com › ... Traduci questa pagina

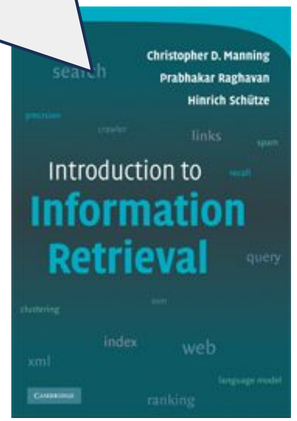

TRECCANI
ISTITUTO
MAGAZINE
CATALOGO
SCUOLA E FORMAZIONE
LIBRI
ARTE
TRECCANI CULTURA
ACQUISTA SU EMPORIUM
ACCEDI
DANTE ALIGHIERI
CERCA IN:
enciclopedia
atlante
lingua italiana
biografico
webtv
vocabolario
il chiasmo
1169 RISULTATI
Tutti i risultati
Biografie [448]
Letteratura [199]
Storia [149]
Arti visive [72]
Lingua [66]
Diritto [48]
Dante Alighièri
ENCICLOPEDIA ON LINE
Poeta (Firenze, tra il maggio e il giugno 1265 - Ravenna, notte dal 13 al 14 settembre 1321). Della madre, che dovette morire presto, non sappiamo che il nome, Bella; il padre, Alighiero di Bellincione [...] La famiglia di D., continuata a Verona nei discendenti di Pietro Alighieri, confluì nel sec. 16° in quella dei Serego, che sarà presto, l'ordine e la giustizia prevarranno anche nel mondo. Dante se ne fa profeta e garante. Cielo e terra ancora una ... LEGGI TUTTO →
CATEGORIA: BIOGRAFIE
TAGS: QUAESTIO DE AQUA ET TERRA – GUIDO NOVELLO DA POLENTA – ALIGHIERO DI BELLINCIONE – ORDINAMENTI DI GIUSTIZIA – FILOSOFIA, DELLA SCIENZA
Nascondi altri risultati su Dante Alighièri (11)
Dante
ENCICLOPEDIA MACHIAVELLIANA (2014)
Gennaro Sasso Poeta, nato a Firenze nel 1265 e morto a Ravenna nel 1321. L'interesse che, nel corso dell'intera sua vita, M. dimostrò per l'opera di D. (ossia, in modo pressoché esclusivo, per la Commedia) può essere considerato sotto tre diversi punti di vista. In primo luogo, D. è l'auctoritas alla ... ... LEGGI TUTTO →
Alighieri, Dante
IL CONTRIBUTO ITALIANO ALLA STORIA DEL PENSIERO: FILOSOFIA (2012)
Dante Alighieri Cesare Vasoli La mirabile costruzione della Comedia è l'esito del passaggio, attraverso le cosiddette opere minori, dal momento esistenziale e poetico della Vita nuova alla meditazione filosofica, etica e politica. Il nuovo nesso tra la poesia e la 'sentenza' era la
pace
POSTER PACE TRECCANI PER L'UNICEF IN UCRAINA
DONA
Pubblicità
VOCABOLARIO
Dantedì
Dantedì Dantedì s. m. Denominazione ufficiale della Giornata nazionale dedicata a Dante Alighieri, il 25 marzo di ogni anno. ♦ È giunto il momento di mettersi d'accordo e fissare d'imperio una Giornata da dedicare al Sommo Poeta: «Vuolsi così...
prosopopèa
prosopopea prosopopèa (non com. prosopopèia) s. f. [dal lat. prosopopoeia, gr. προσωποποιία, der. di προσωποποιέω «personificare», comp. di πρόσωπον «faccia, persona» e ποιέω «fare»]. – 1. Figura retorica per cui si introducono a parlare persone...
Leggi Tutto →


cooking tutorial - YouTub
youtube.com/results?search_query=cooking+tutorial
Incognito
YouTube
cooking tutorial
ACCEDI
Home
Esplora
Shorts
Iscrizioni
Raccolta
Cronologia
Effettua l'accesso per mettere Mi piace ai video, commentare e iscriverti ai canali.
ACCEDI
IL MEGLIO DI YOUTUBE
Musica
Sport
Giochi
Film
Notizie
Dal vivo
Moda e bellezza
Video a 360 gradi
FILTRI
Informazioni su questi risultati
Gordon Ramsay's Top Basic Cooking Skills | Ultimate Cookery Course FULL EPISODE
20 Recipes You Should Learn In Your 20s • Tasty
Quick & Easy Recipes With Gordon Ramsay
BICOL EXPRESS | THE TASTIEST & SPICIEST BICOL EXPRESS | HOW TO COOK


the original hamlet manu
google.com/search?q=the+original+hamlet+manuscript&tbm=isch&ved=2ahUKEwiL5L_f06z...
Incognito
Google
the original hamlet manuscript
Accedi
Tutti
Immagini
Shopping
Notizie
Maps
Altro
Strumenti
SafeSearch
soliloquy
speech
quarto
hamlet soliloquy
monologue
second quarto
polonius
Hamlet - Shakespeare in q...
bl.uk
Hamlet Q1 - Wikipedia
en.wikipedia.org
File:Hamlet, Shakespeare, 1...
commons.wikimedia.org
Shakespeare, Hamlet
bl.uk
File:Hamlet, Shakespeare, 1...
commons.wikimedia.org

Information Retrieval

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature that satisfies an information need from within large collections (usually stored on computers)."



https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Multimedia Information Retrieval

Multimedia Information Retrieval

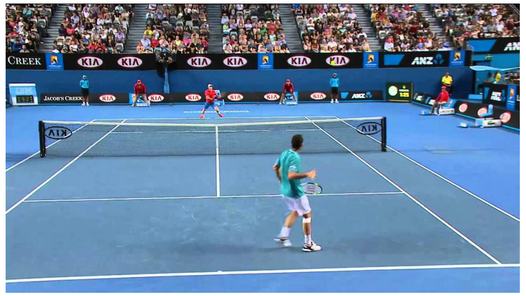
- *multimedia*: "two or more different media" and refers to different *modes* of information consumption
  - listening, seeing, reading, watching, smelling etc
- *multimedia information retrieval*: we want the query and the retrieved documents to have possibly different modalities

Milton Keynes's Peace Pagoda

Built by the monks and nuns of the Nipponzan Myohoji, this was the first Peace Pagoda…

Challenges in Multimedia Information Retrieval

- The content to retrieve must be encoded in some way
  - Exploit metadata
  - Directly exploit the content (Content-Based Information Retrieval)

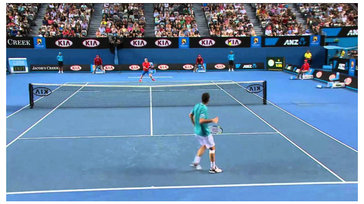


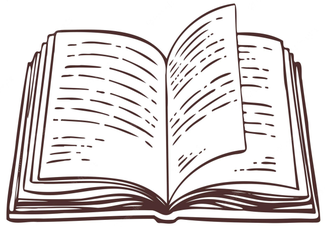
filename: “tennis.jpg”
size: 1280x720
timestamp: 22 gen 2012

- Large scale scenarios
  - Billions of items to be retrieved in few milliseconds
  - E.g.: Google

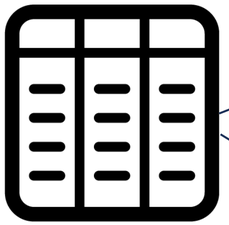
Challenges in Multimedia Information Retrieval

- Multimedia data (e.g. images, videos) are <u>not structured data</u>
  - Not possible to use standard relational databases (e.g. MySQL)
  - An image cannot be directly stored as nice tabular data





unstructured

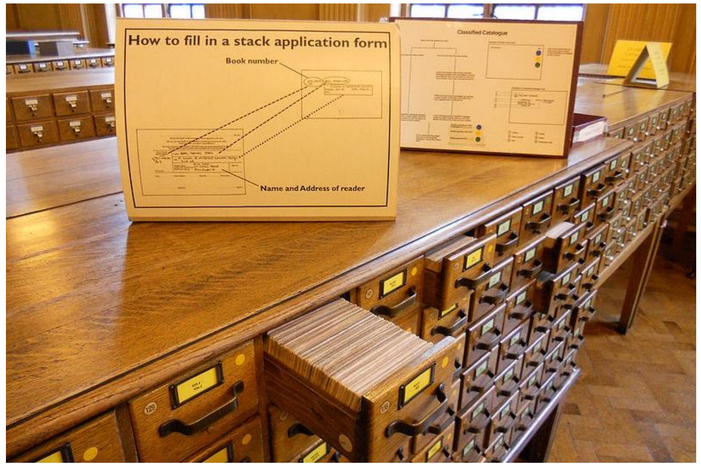


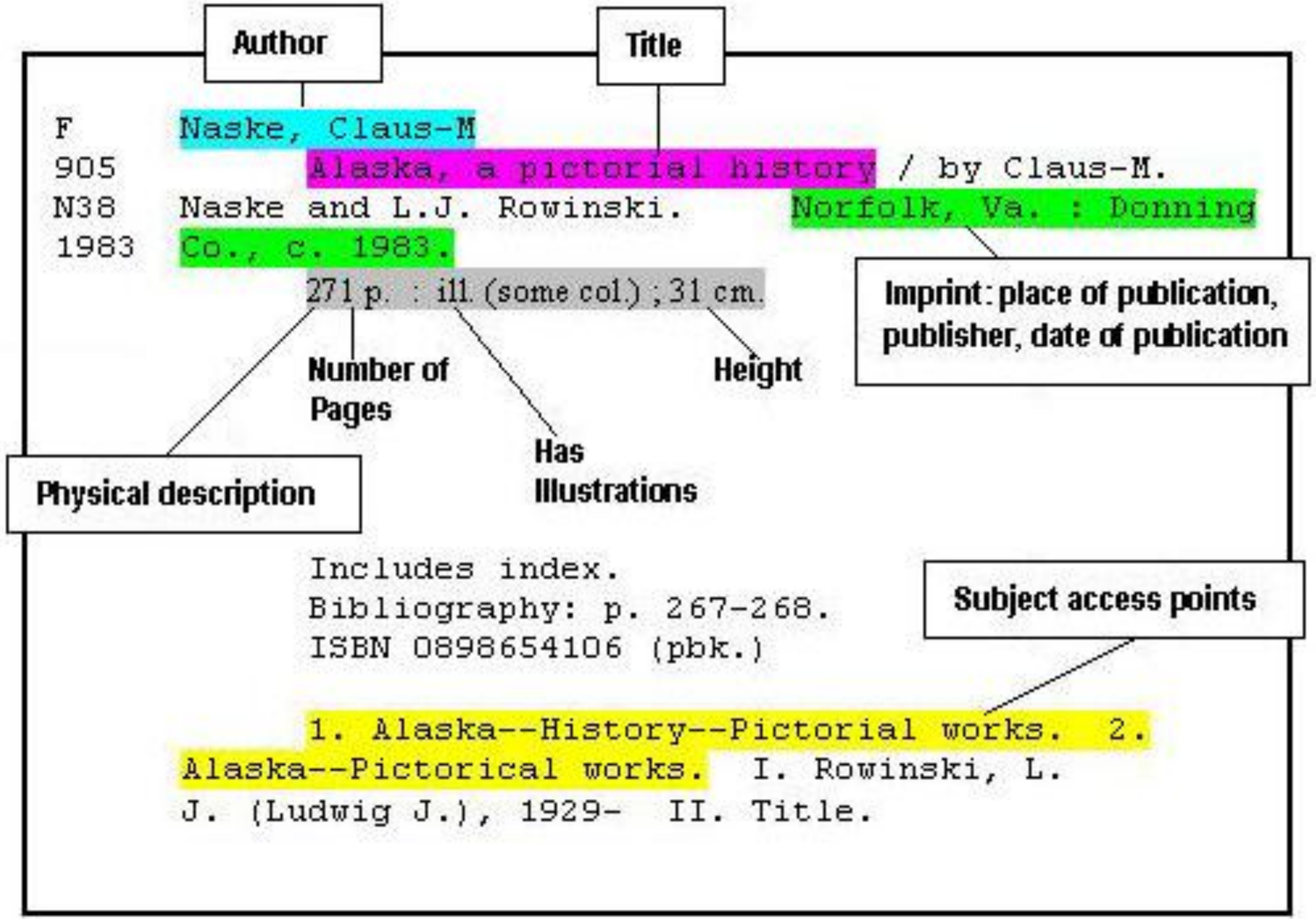
structured

Image	date	size	pixels(?)
…			
cat.jpg	15/06/22	300x200	???
…			

The analog way: Catalog card

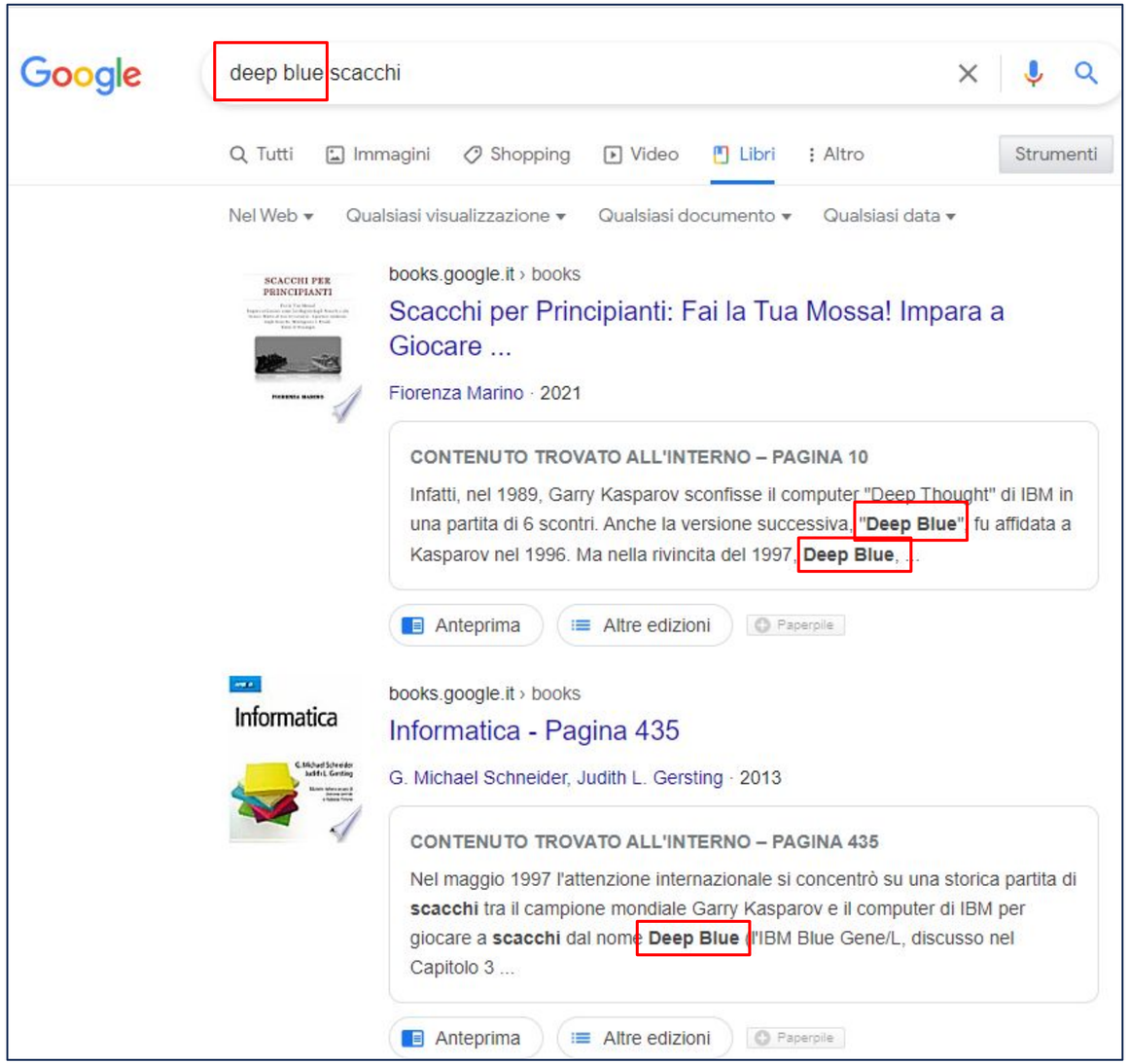
A catalog card is an individual entry in a library catalog containing bibliographic information, including author's name, book title, etc…





The importance of the Content

- In general, we need to search <u>also the content</u>
- Google indexes the **content**, so that we can search for something even in the content, not only in metadata
- For books (and textual documents in general) this is not too much difficult
  - Each document is a set of words
  - Find out how many words match with the query
  - We will see in few slides…

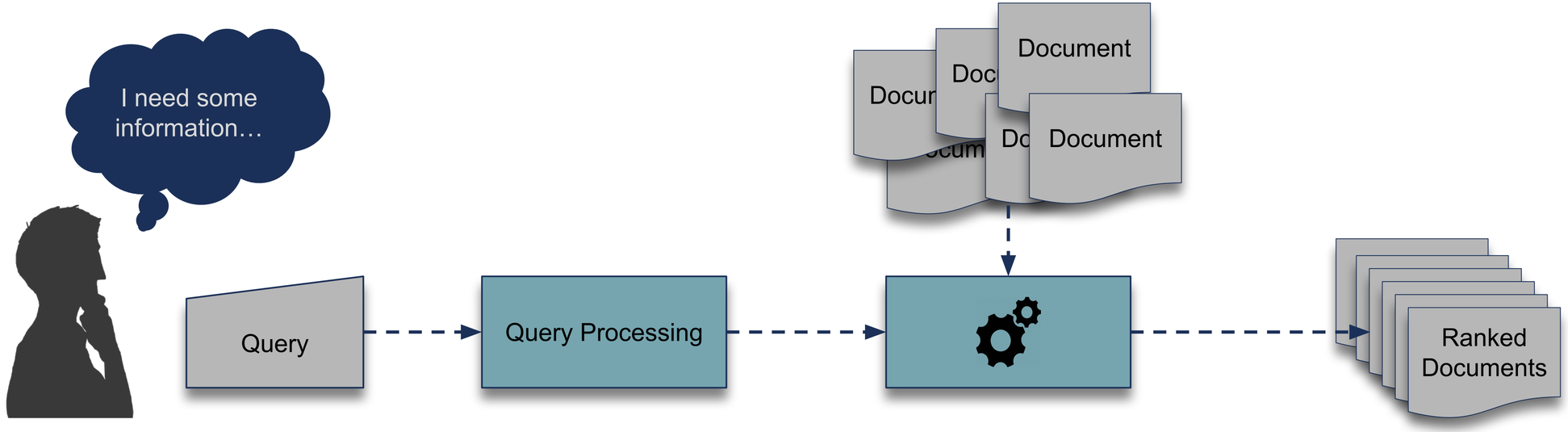


Information Retrieval System

In the more general terms, the main components are

. A query
. A database of documents among which we want to search

The <u>output is a subset of documents</u>, the ones **relevant** to the given query, ranked by decreasing relevance



Textual retrieval

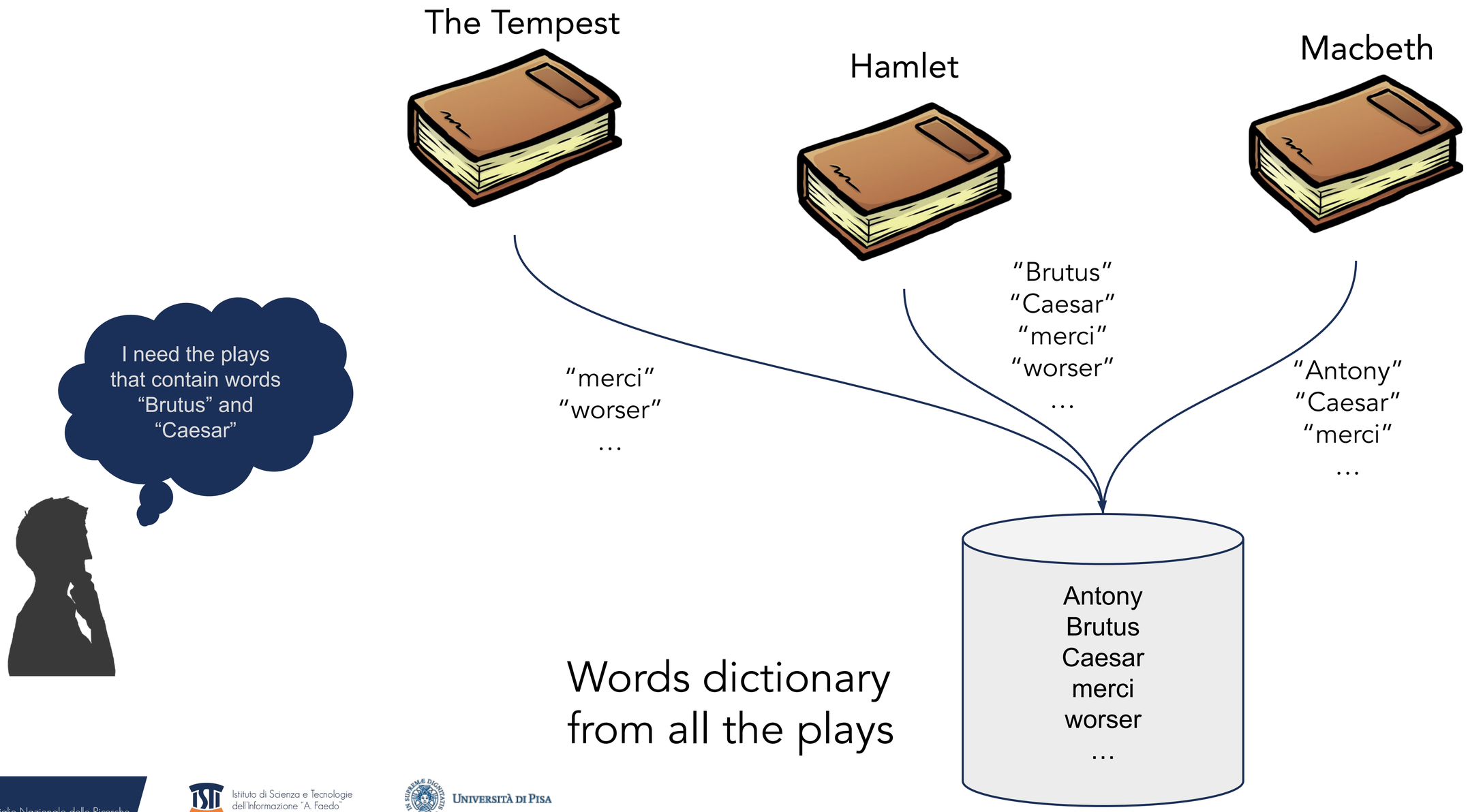
Historically, documents were only textual

- 1950s: textual document retrieval
- 1980s: multimedia documents acquired interest
  - Difficulty of processing "non-textual documents"
  - Medium mismatch problem, or semantic gap
    - E.g.: How to match an image with a text that describes it?

There is ... a machine called the Univac ... whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape. By this means the text of a document, preceded by its subject code symbol, can be recorded ... the machine ... automatically selects and types out those references which have been coded in any desired way at a rate of 120 words a minute

— J. E. Holmstrom, 1948

A simple retrieval model for texts



The simple boolean model

- We create an <u>incidence matrix</u> that tells us which word appears in each of the documents

Documents (Shakespeare Plays)

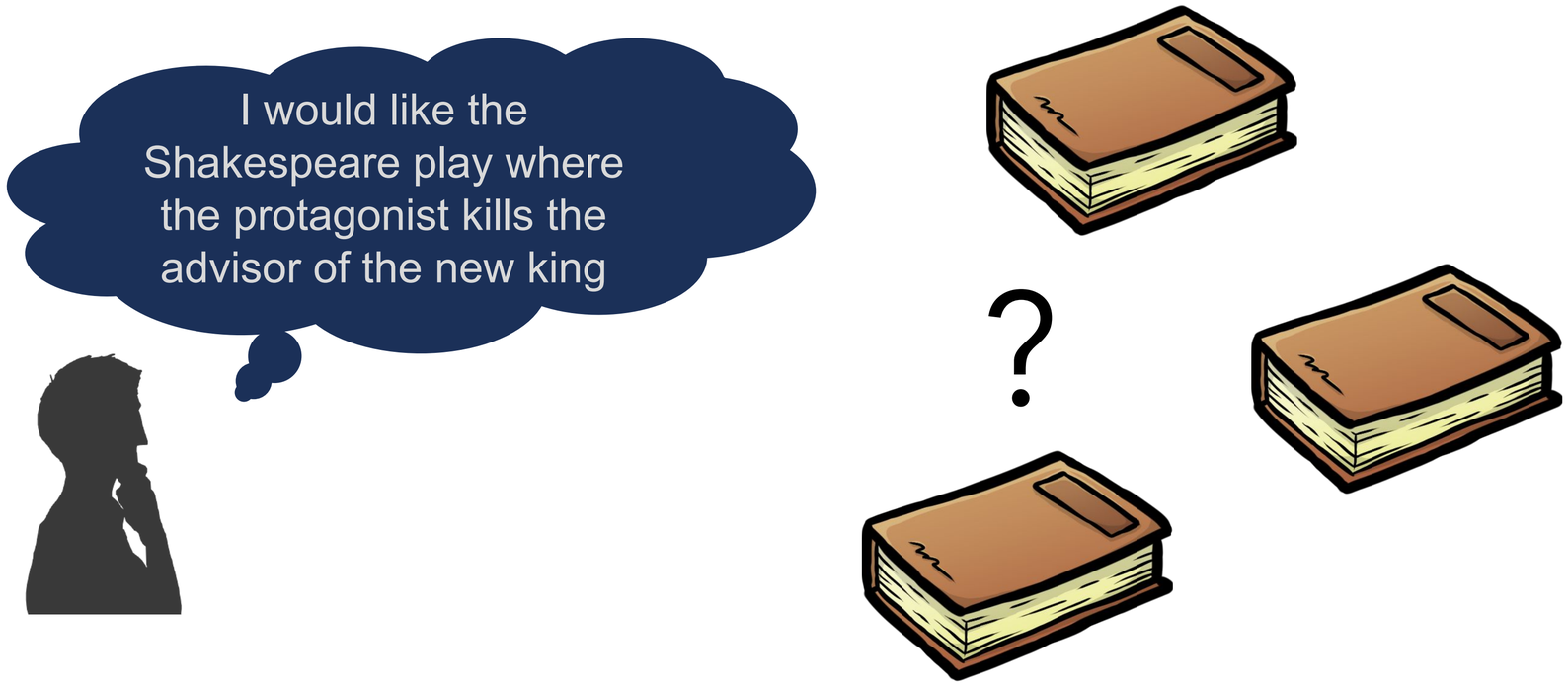
Words dictionary

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

- To answer the query Brutus AND Caesar AND NOT Calpurnia
  - We take the vectors for Brutus, Caesar and Calpurnia, complement the last, and then do a bitwise AND: 110100 AND 110111 AND 101111 = 100100
  - The answers for this query are thus Antony and Cleopatra and Hamlet

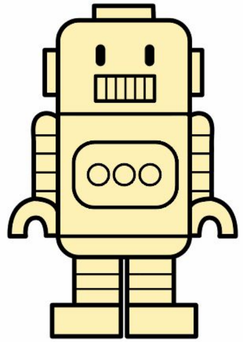
Can be improved but…

- This model could be improved by leveraging word occurrences
  - the more frequent a word, the more important it is
  - a word appearing too much in a corpus is not too significative (e.g, “the”)
- However, what about a query like:



Problems

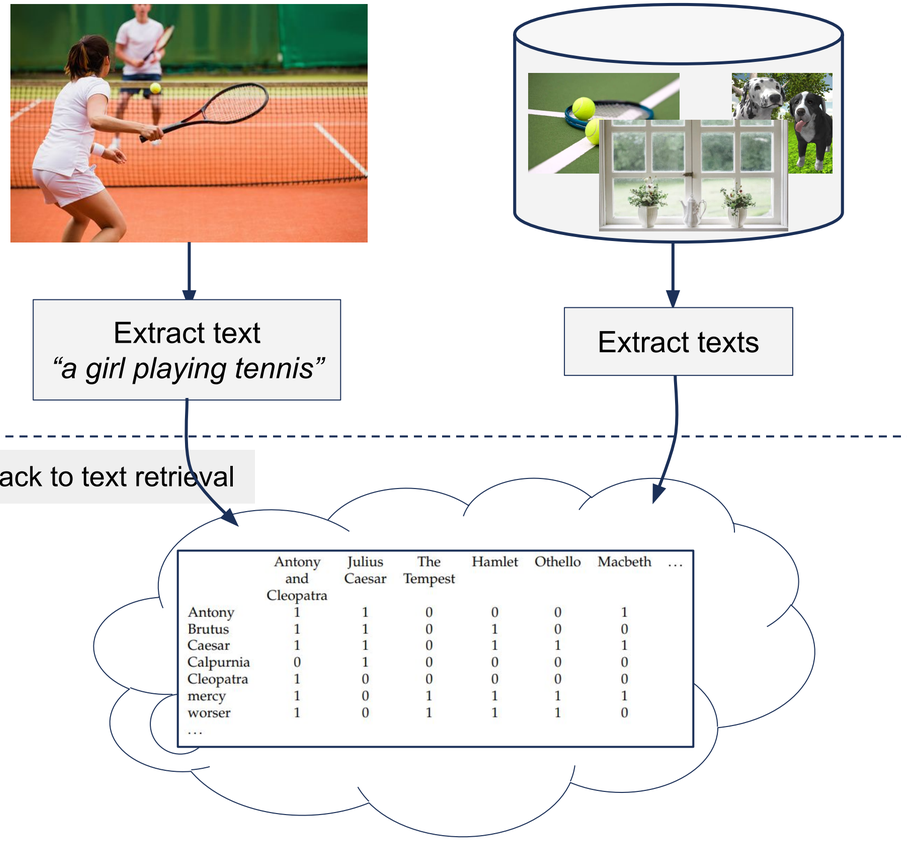
- The system works at a very low level, very "robotic"
  - Only exact word matching
  - What about synonyms? And what about the context?
- This method does not scale to other media objects
  - Internet is full of less structured media objects that need to be stored and efficiently retrieved
    - Images, videos, audio
    - We cannot search these unstructured objects using exact match



How to handle images, videos, …?

- Frame the problem <u>as a text retrieval problem</u>
  - For images, in principle we could rely on some metadata, e.g. the **alternative text** associated to the image HTML "alt" tag
  - For videos, use to the video description provided by the user that uploaded it


```
  <span id="imageBlockEDPOverlay"></span>
▼<li class="image                item itemNo0 maintain-height
selected" style="cursor: pointer;">
  ▼<span class="a-list-item">
    ▼<span class="a-declarative" data-action="main-image-click" data-
    main-image-click="{}">
      ▼<div id="imgTagWrapperId" class="imgTagWrapper" style="height:
      500px;">
          <img alt="Doritos Tortilla Chips, Nacho Cheese, 1.75-Ounce
          Large Single Serve Bags (Pack of 64)" src="https://images-
          na.ssl-images-amazon.com/images/I/
          71Br1LeeJGL._SY679SX…,0,0,486,679_PIbundle-
          64,TopRight,0,0_SX486_SY679_CR,0,0,486,679_SH20_.jpg" data-
          old-hires="https://images-na.ssl-images-amazon.com/images/I/
          71Br1LeeJGL._SL1366_.jpg" class="a-dynamic-image  a-stretch-
```



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

How to handle images, videos, …?

- But…
  - Often these data is not available, or contain errors
  - Not representative of the whole multimedia element
- We often need to rely only on the content

Looking at the <u>content</u> →high level understanding
→certain degree of <u>intelligence</u>

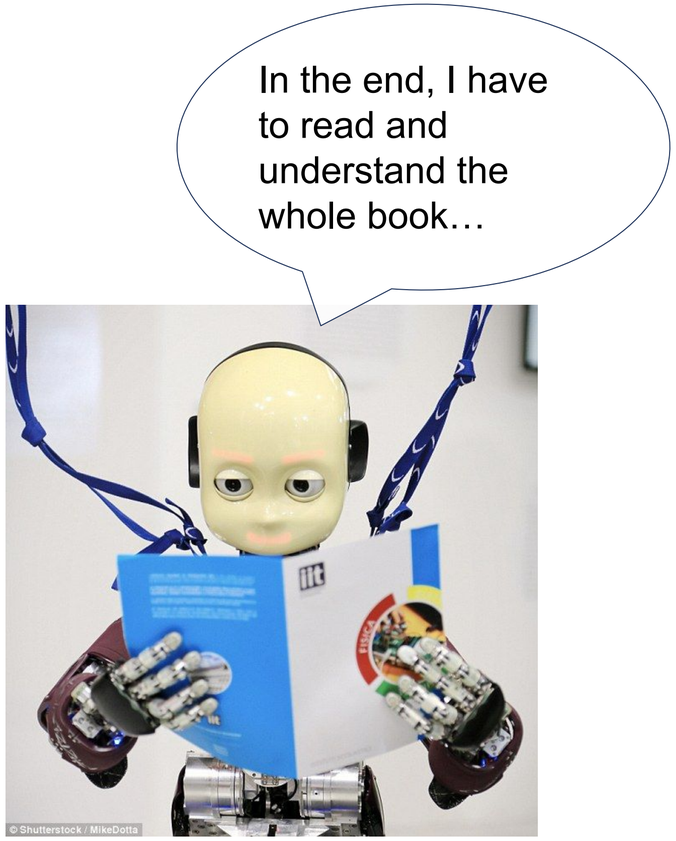


Image Retrieval

- Find images similar to the one given as a query
- Query by example

- We <u>cannot define a precise matching criterion</u> between images, but…
- We can quantify in some way their <u>similarities</u>

Paradigm shift: similarity search

- Encode a media object (a text, an image, or a video) into some <u>numerical representation (or feature)</u>
- Measure the similarity between the representation of the query and the representation of the documents in the database

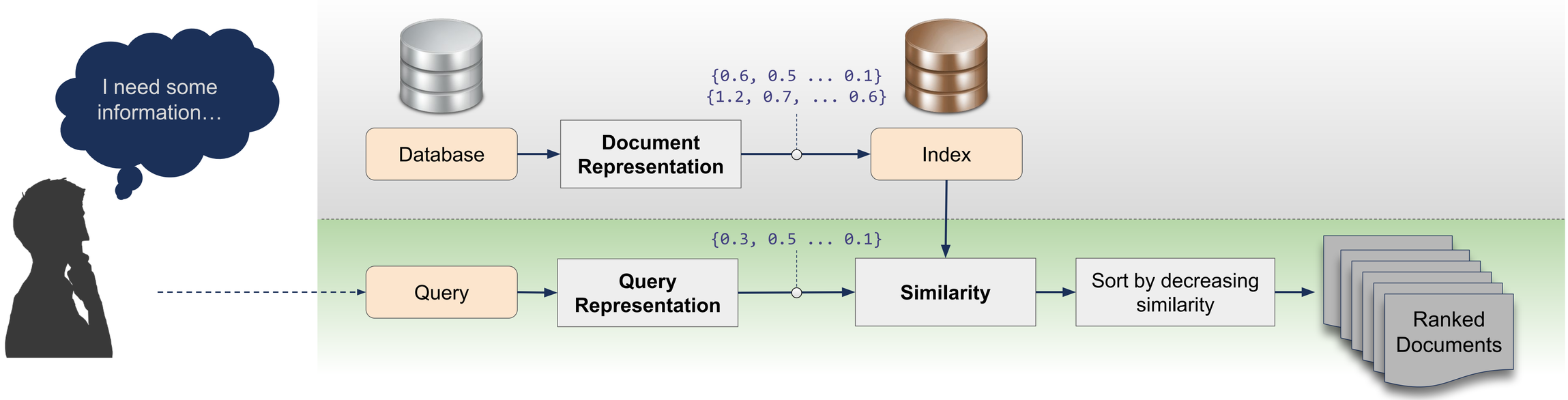
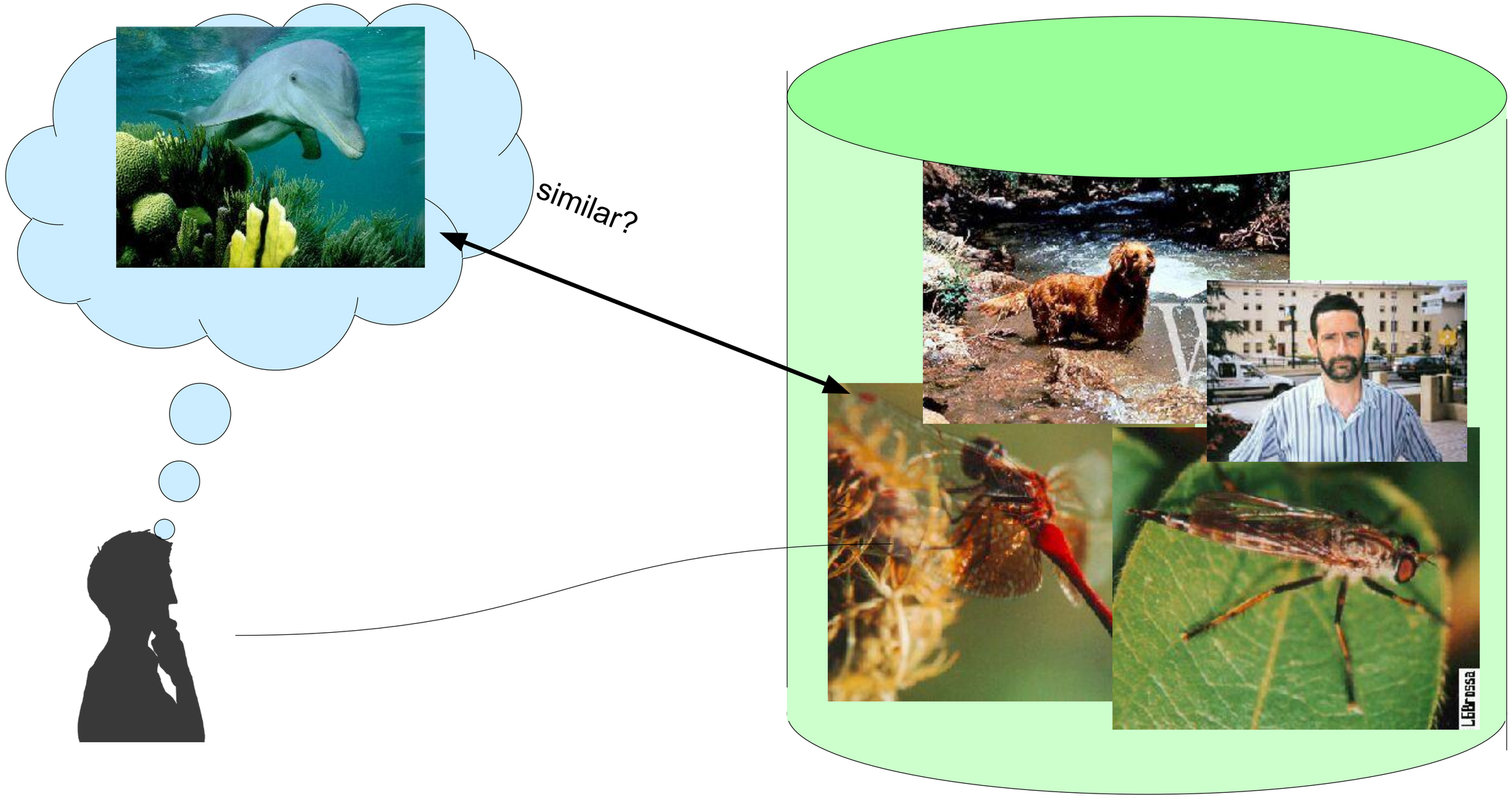
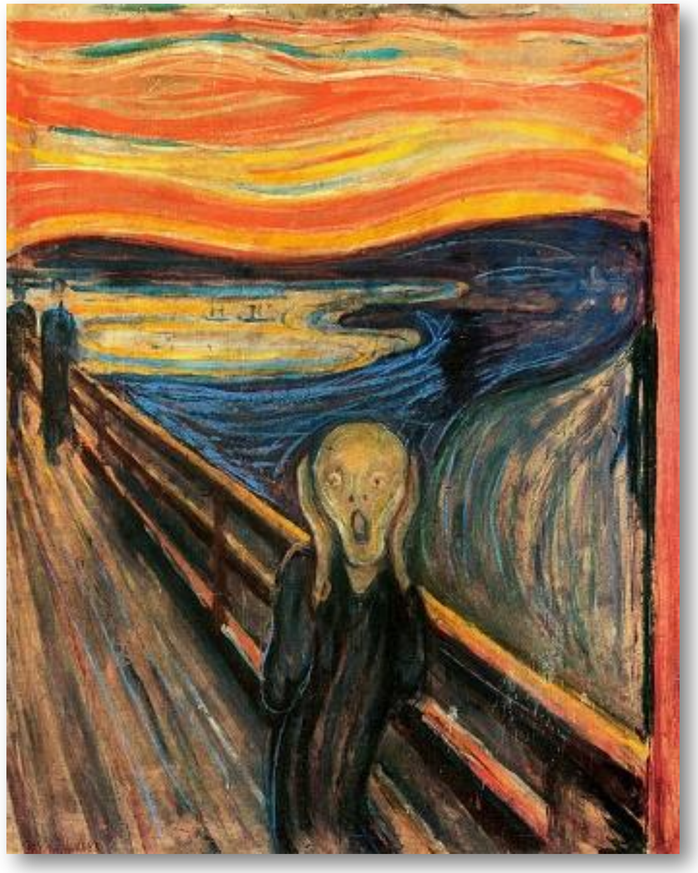


Image Retrieval: what does “similar” mean?

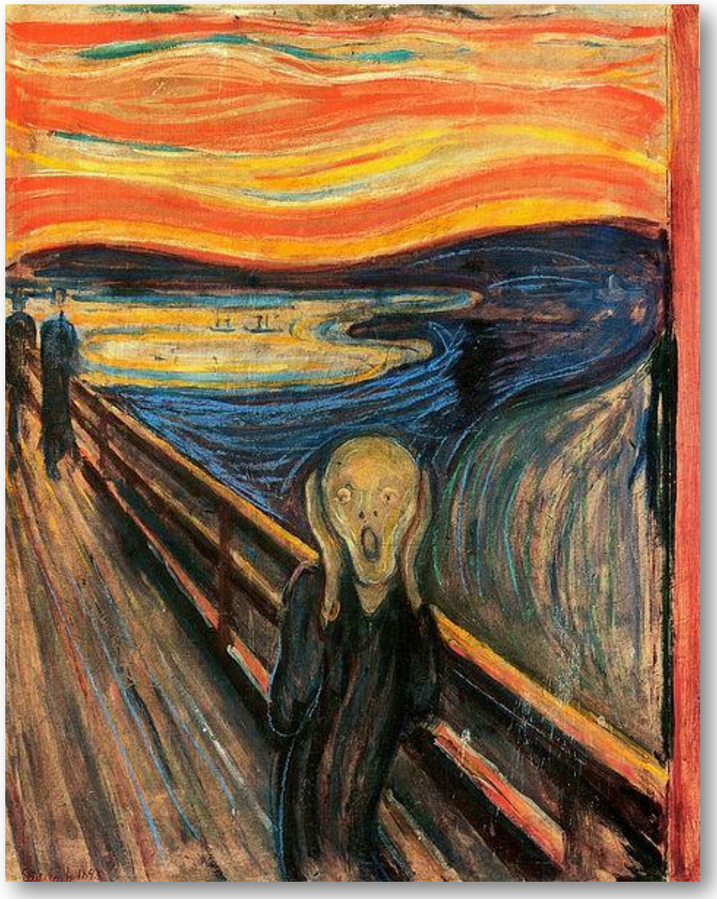


What’s This?



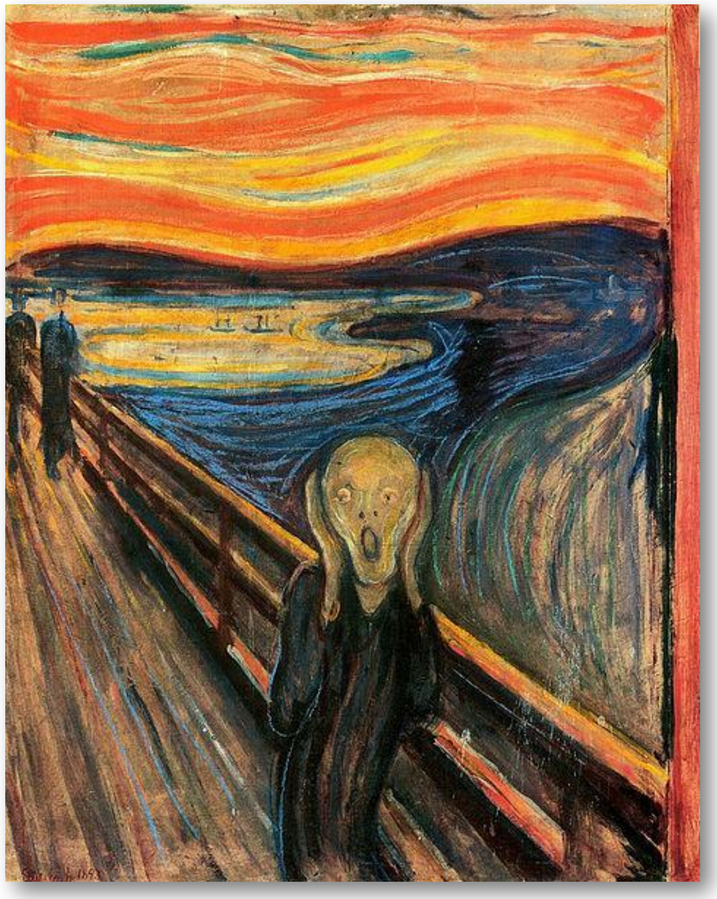
What's this?

The Scream, Edvard Munch

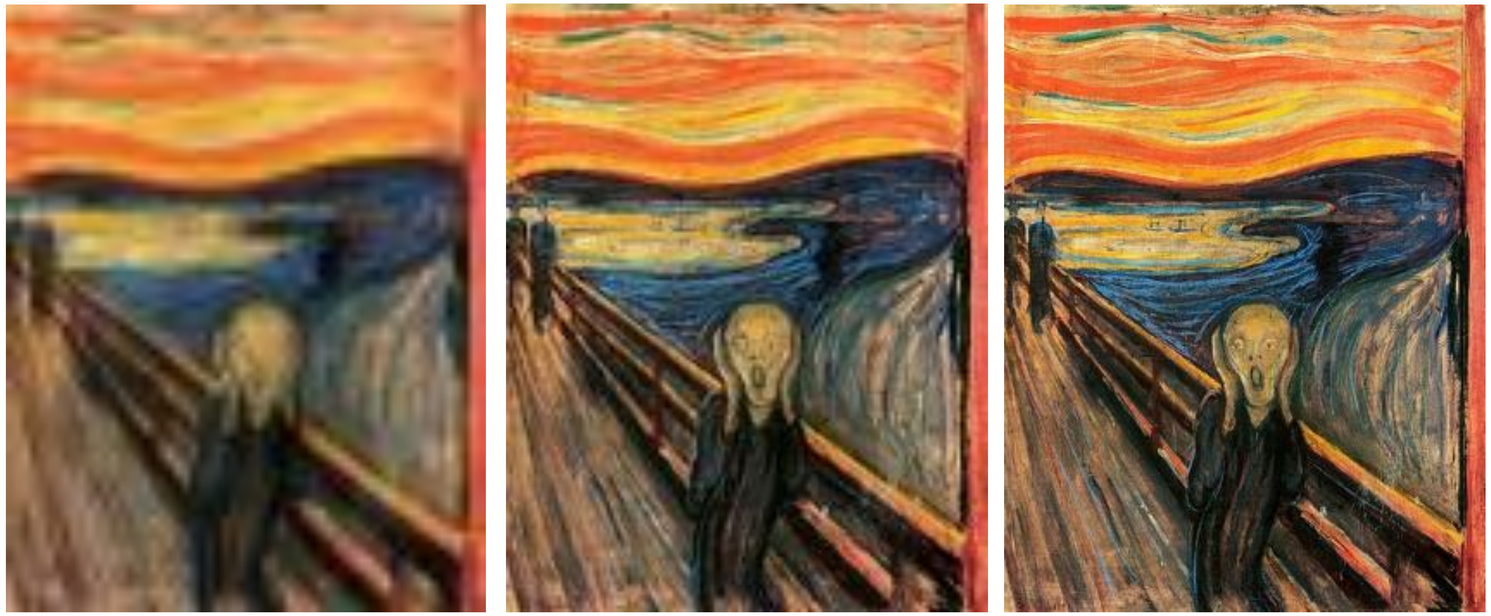


- The file at
  http://upload.wikimedia.org/.../475px-The_Scream.jpg

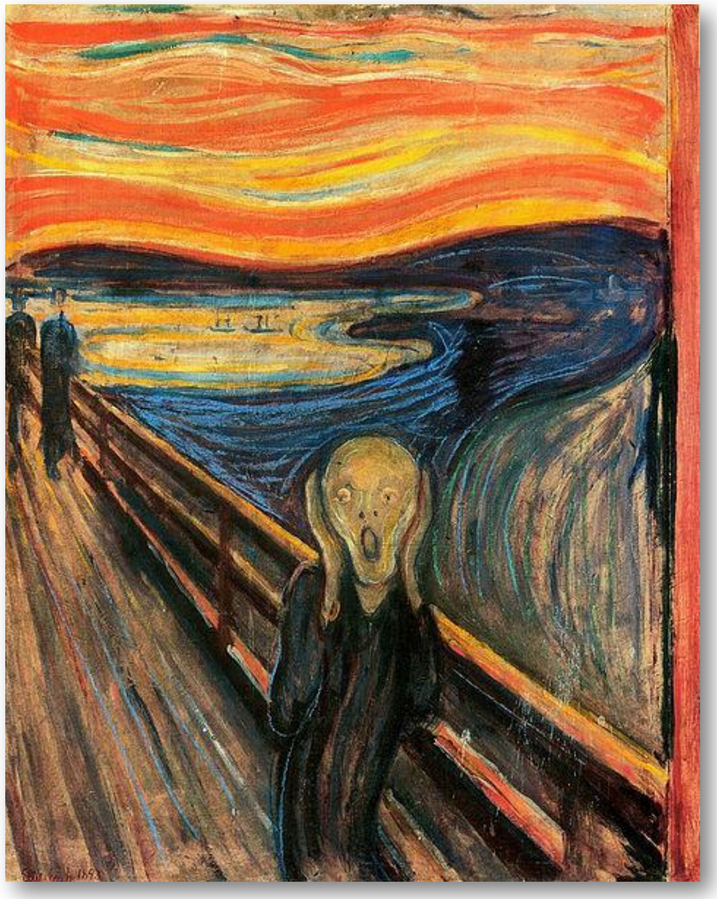
The Scream, Edvard Munch



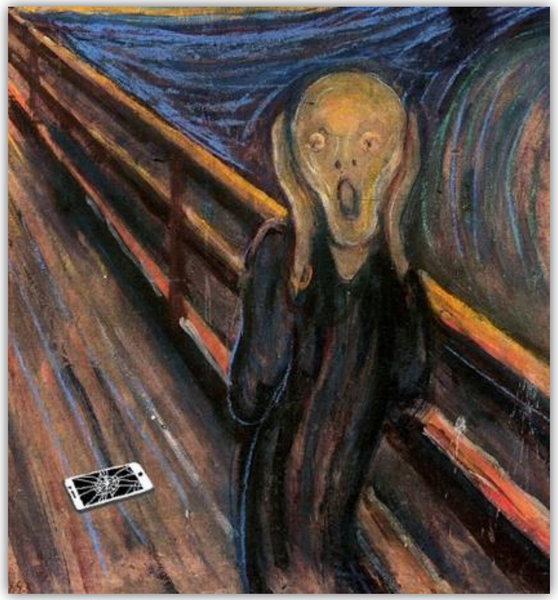
- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture as

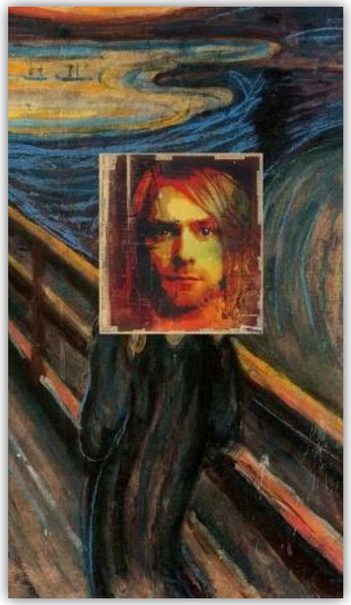


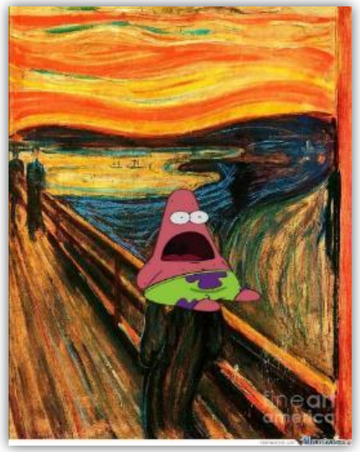
The Scream, Edvard Munch

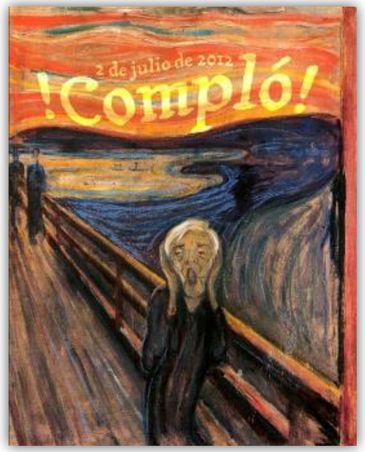


- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same as

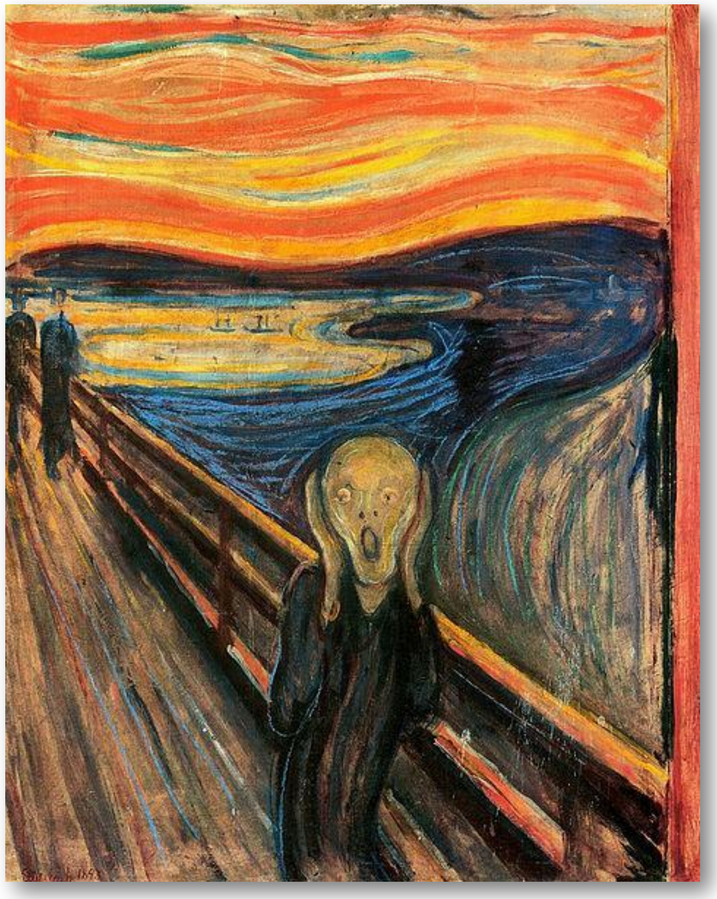




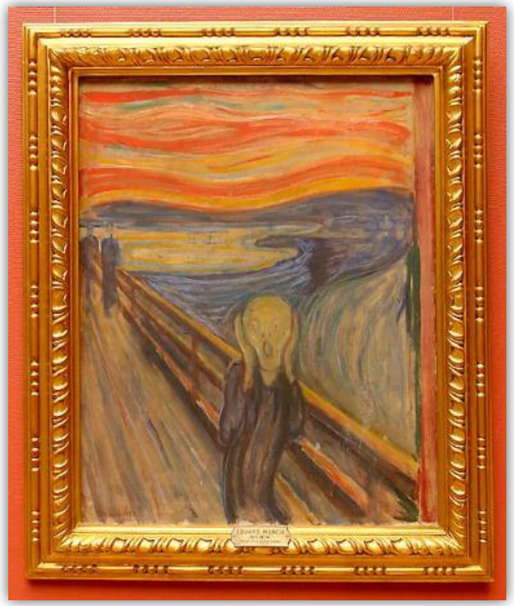




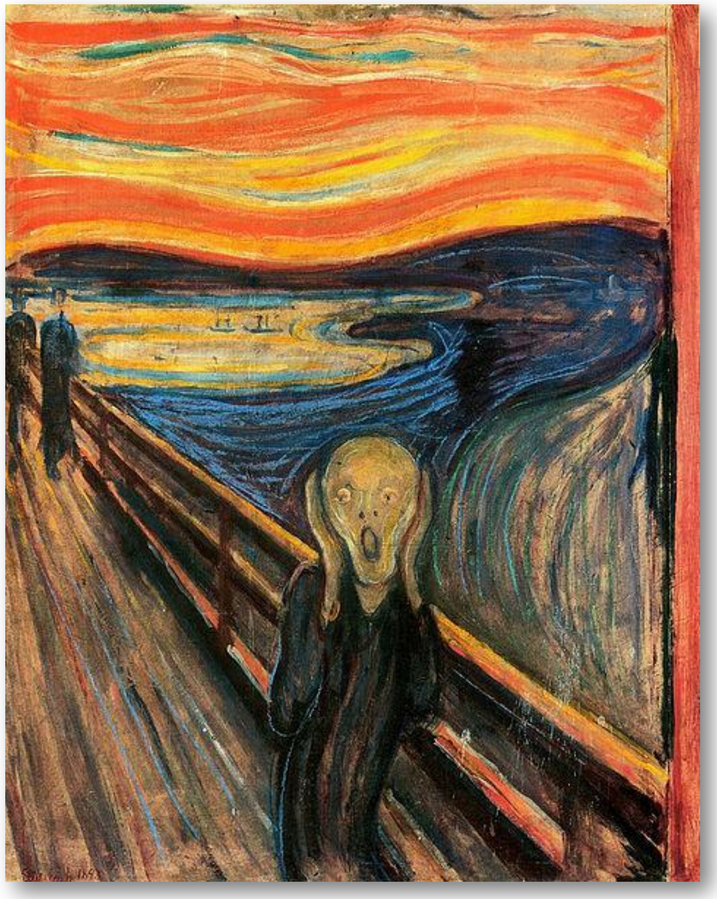
The Scream, Edvard Munch



- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo as

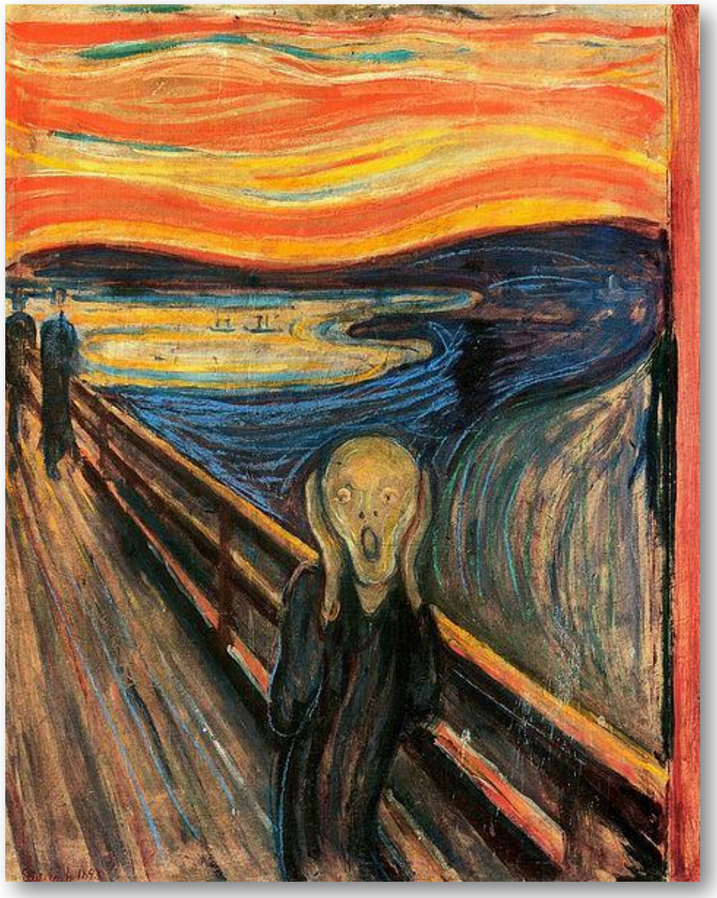


The Scream, Edvard Munch



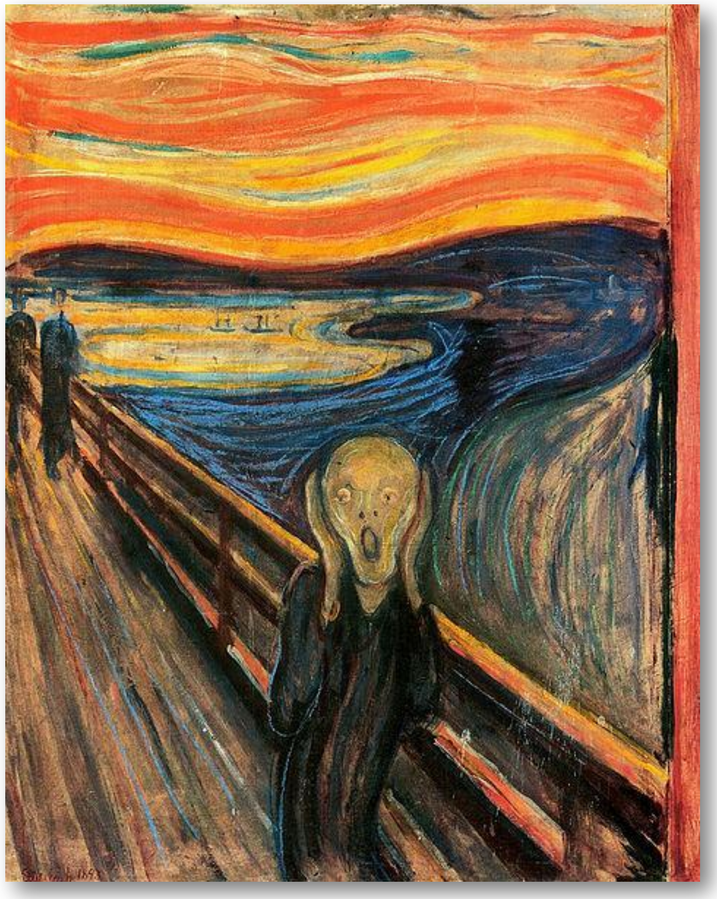
- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream" by Edvard Munch as

The Scream, Edvard Munch



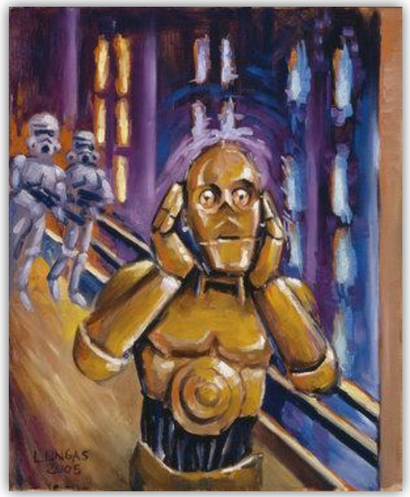
- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch as

The Scream, Edvard Munch

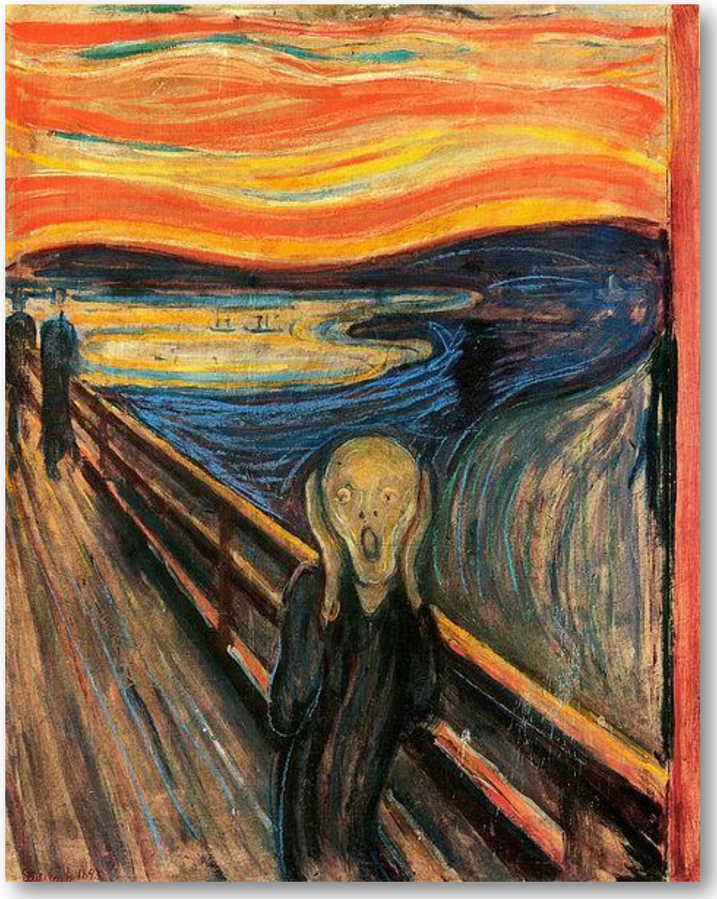


- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists

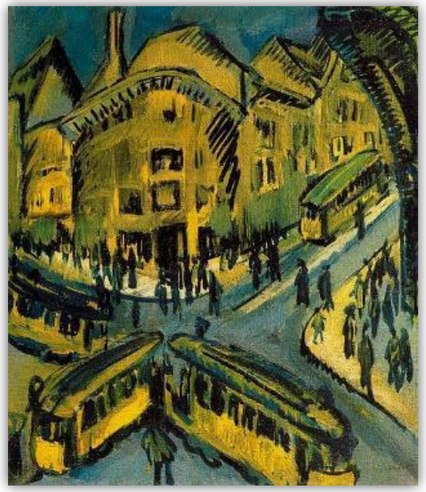
as

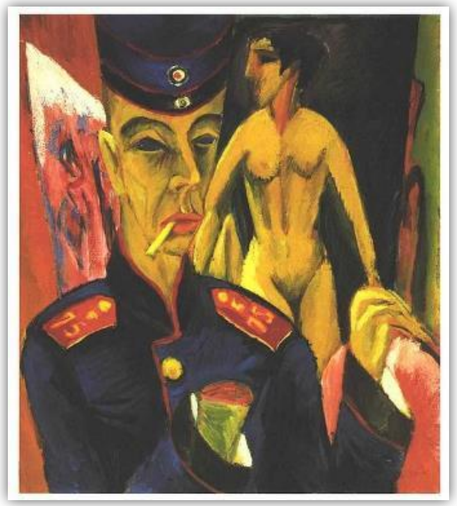


The Scream, Edvard Munch

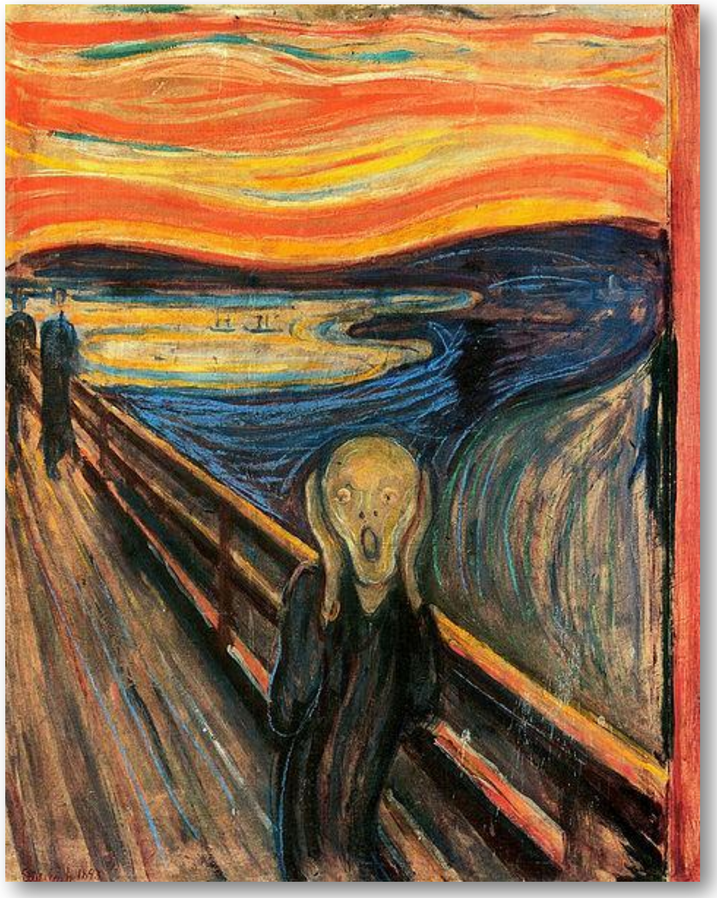


- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists
- An expressionist painting as



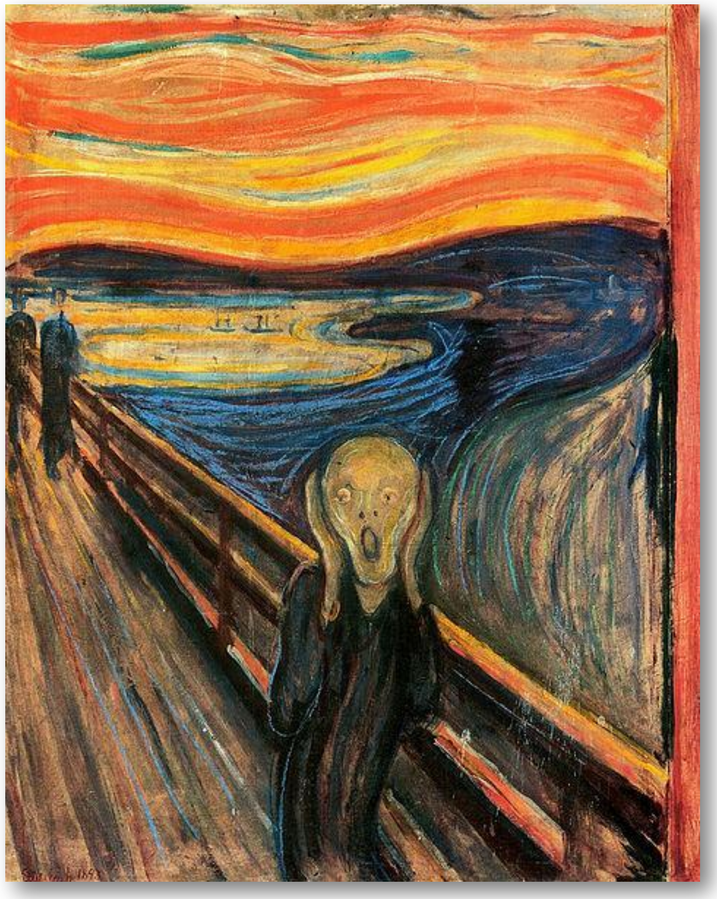


The Scream, Edvard Munch



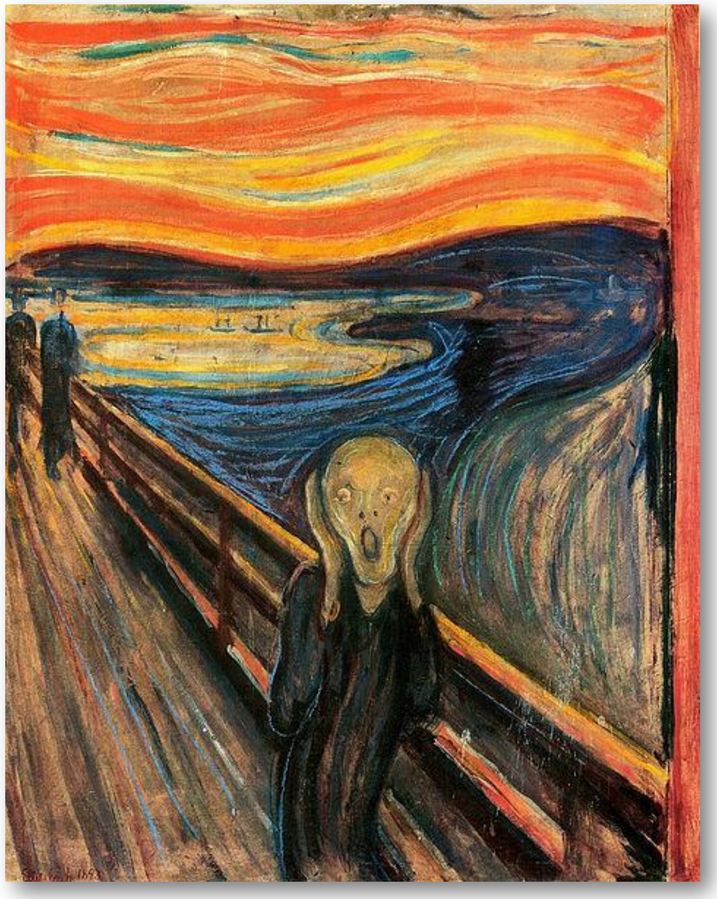
- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists
- An expressionist painting
- A painting as

The Scream, Edvard Munch



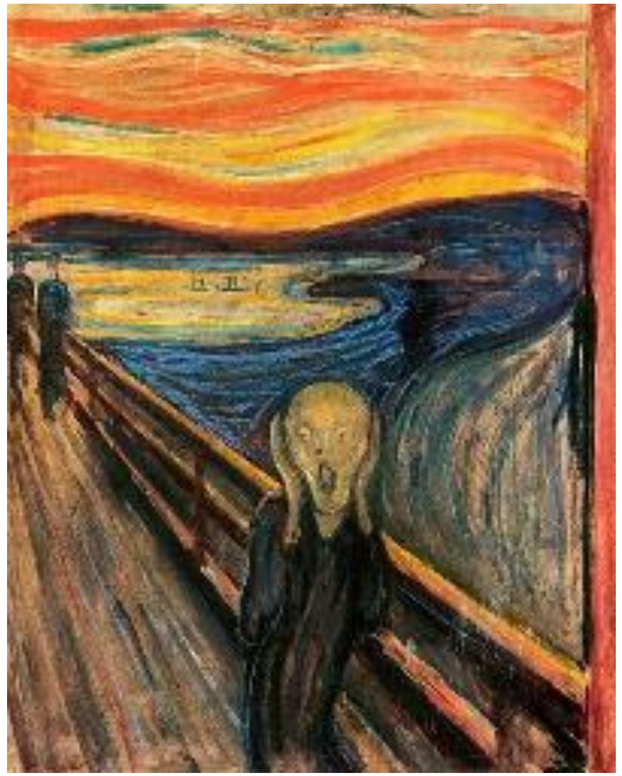
- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists
- An expressionist painting
- A painting
- An hand made object as

The Scream, Edvard Munch



- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists
- An expressionist painting
- A painting
- An hand made object
- An artificial object
  being the product of intentional human manufacture

The Scream, Edvard Munch

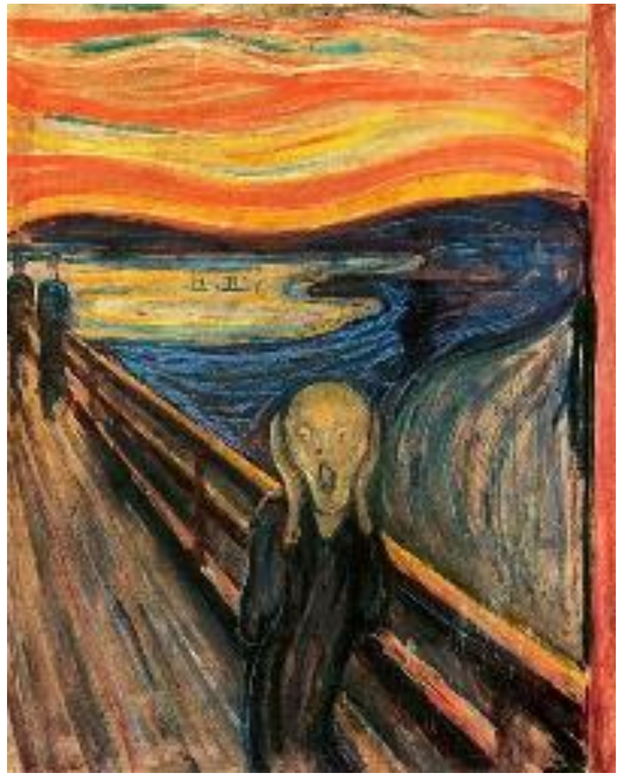


Low-level

- **The file** at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- **One** of the files of the same picture
- Almost the **same**
- A picture of **the object** at National Gallery, Oslo
- **One of** “The Scream”s by Edvard Munch
- **A** painting by Edvard Munch
- **One** of “The Scream”s by various artists
- **An** expressionist painting
- **A** painting
- **An** hand made object
- **An** artificial object
  being the product of intentional human manufacture

High-level

The Scream, Edvard Munch



Matching

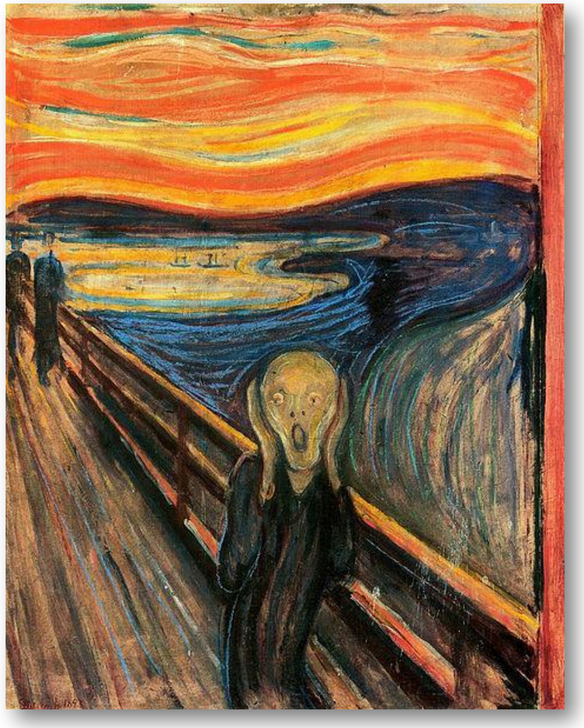
Recognition

Classification

- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the same
- A picture of the object at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists
- An expressionist painting
- A painting
- An hand made object
- An artificial object
  being the product of intentional human manufacture

Similarity between different representations

It is possible to define a concept of similarity for these different levels of abstraction



instance similarity
(very low abstraction)

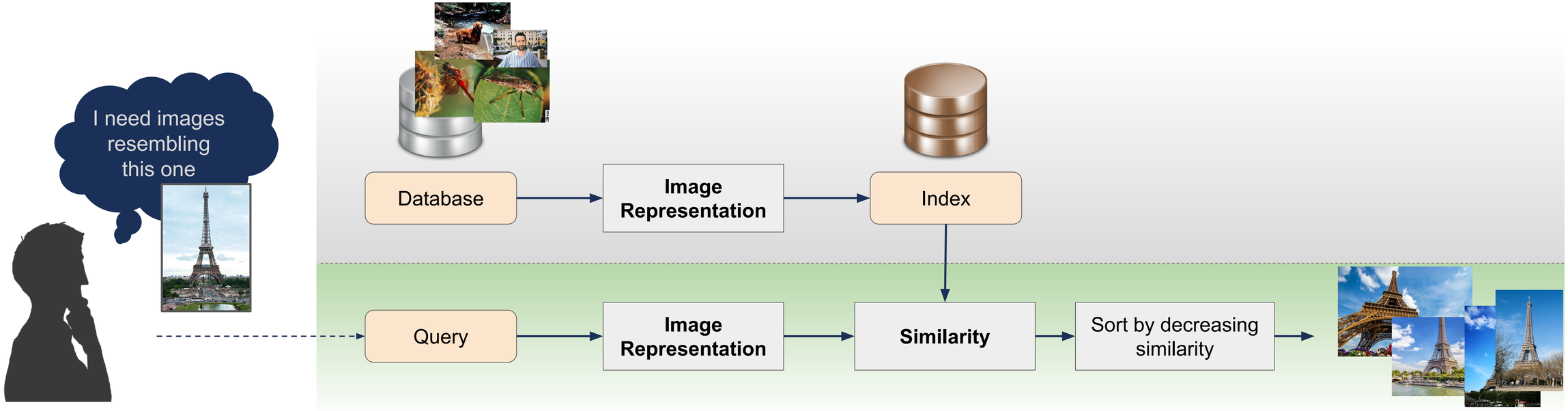
semantic similarity
(very high abstraction)

- The file at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- One of the files of the same picture
- Almost the **same**
- A picture of **the object** at National Gallery, Oslo
- One of "The Scream"s by Edvard Munch
- A painting by Edvard Munch
- One of "The Scream"s by various artists
- An expressionist painting
- A painting
- An hand made object
- An artificial object
  being the product of intentional human manufacture

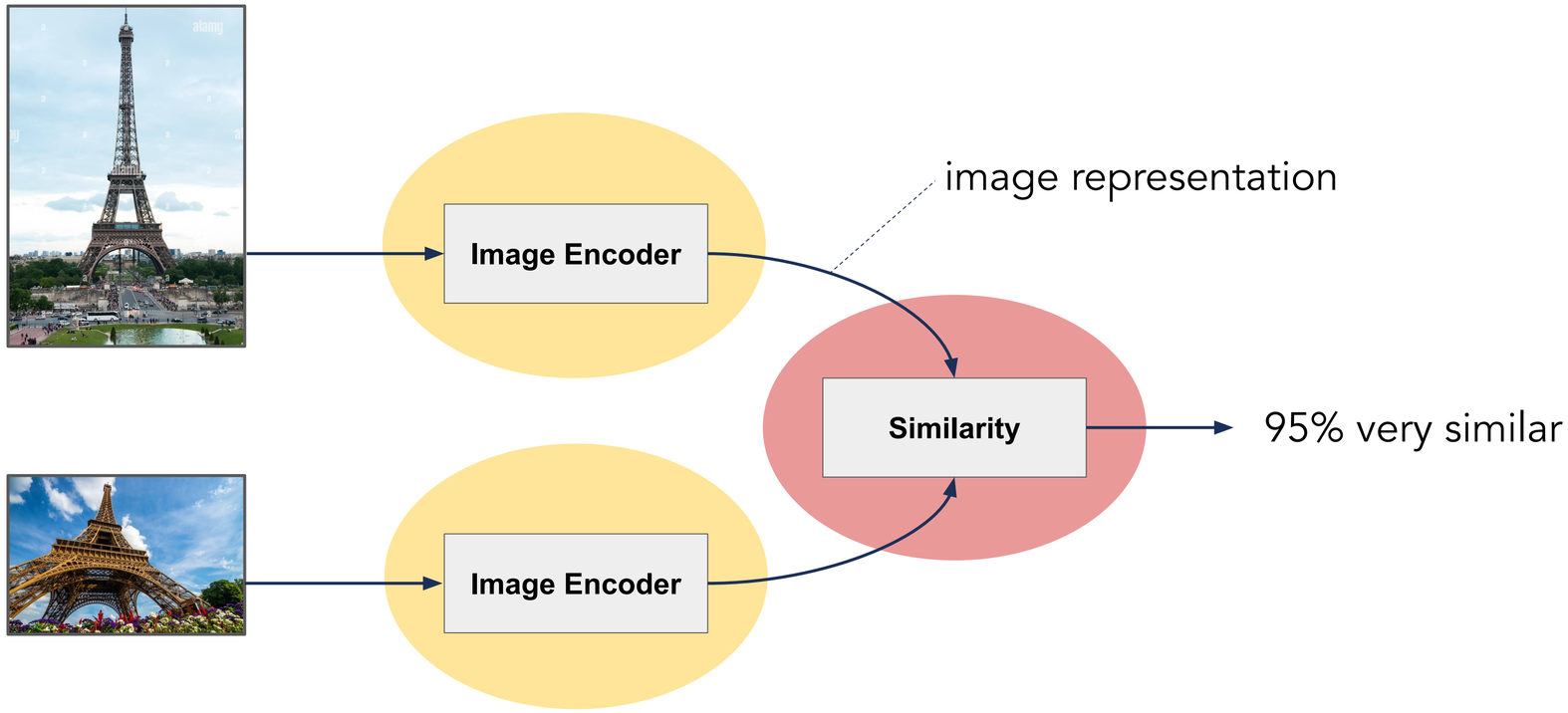
How can we associate a **representation** to each one of these abstraction levels?

Image Representations

Image retrieval setup



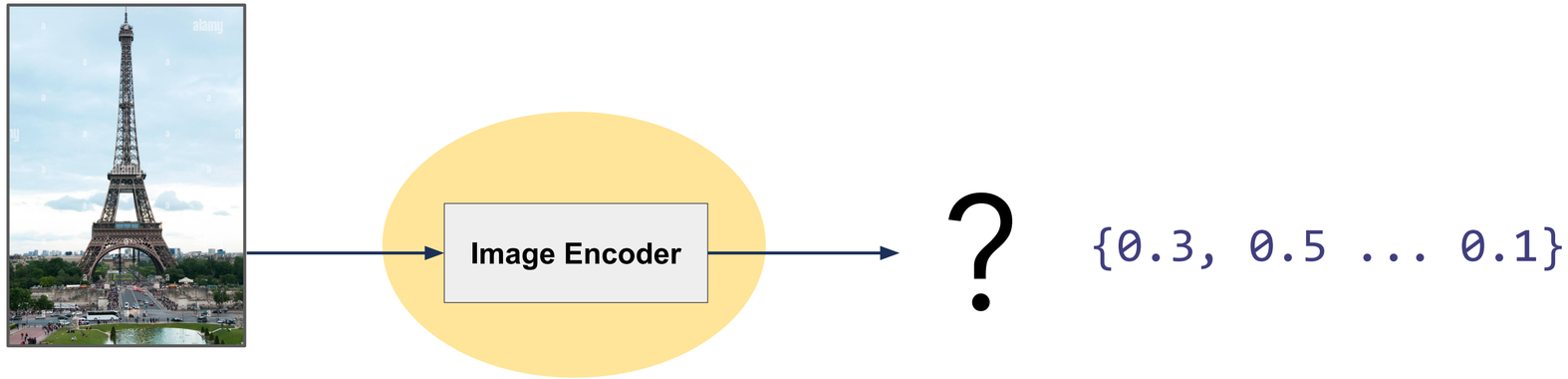
Representations and similarities



All boils down to

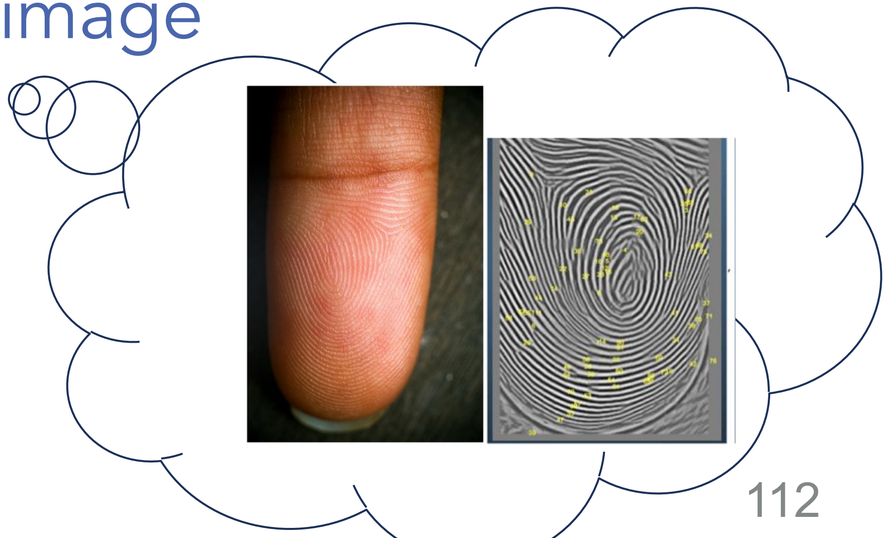
- Finding an image encoder that outputs good representations
- Measuring the similarity between these two representations

Representation / Features

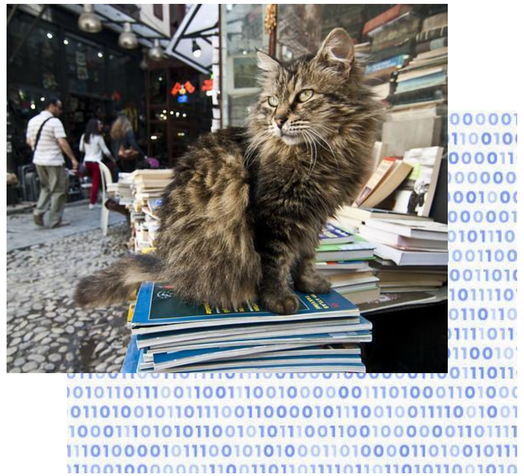


An image is converted into a set of numbers, called **vector**

It can be considered the "fingerprint" of that image



Representation / Features



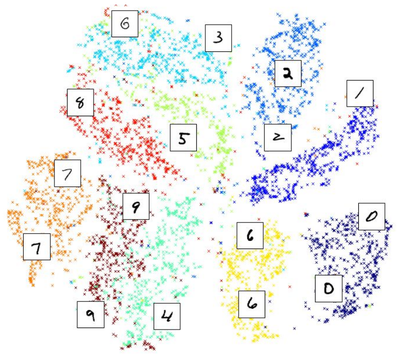
Digital Image

Encoder

feature extractor

A representation (a feature)

{0.3, 0.5 ... 0.1}



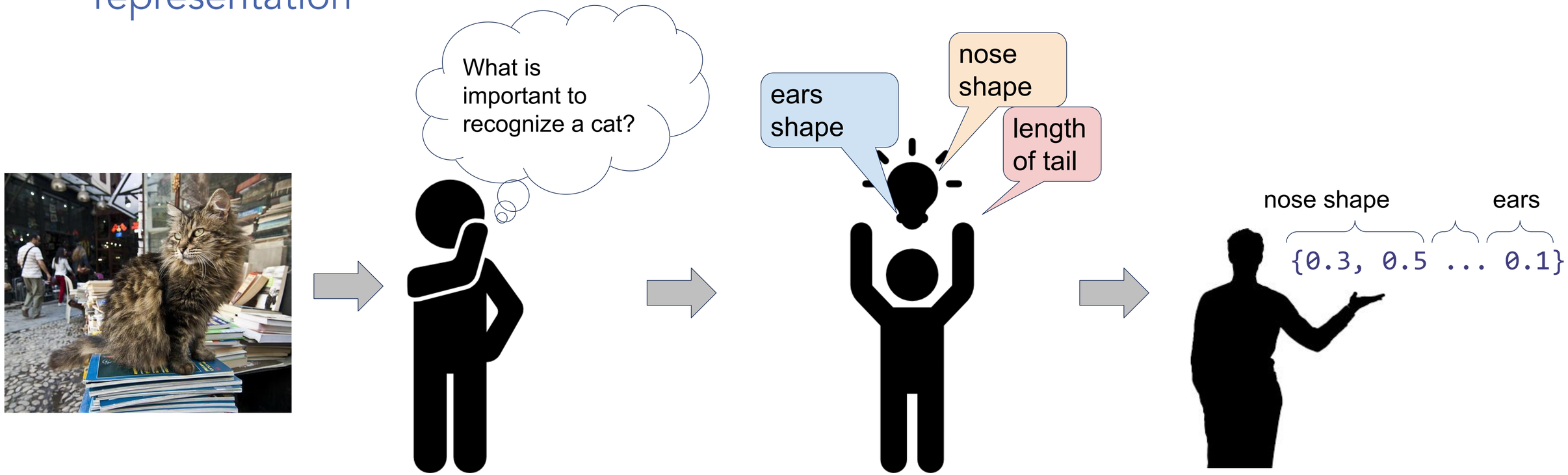
in a **latent space**

classification
“cat”

An encoder (or feature extractor)
takes an input (e.g., an image)
and produce a representation (or descriptor)
that is used (in place of the image) for the specific task.

Handcrafted features

- Before Deep Learning, handcrafted features/representations:
  - Human is always in the loop
  - <u>He chooses what is important</u> in the picture for creating a numerical representation



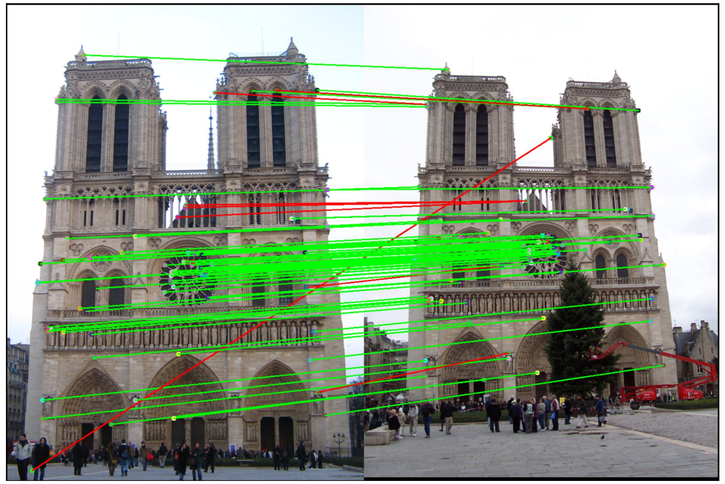
Representation / Features

Global Features:

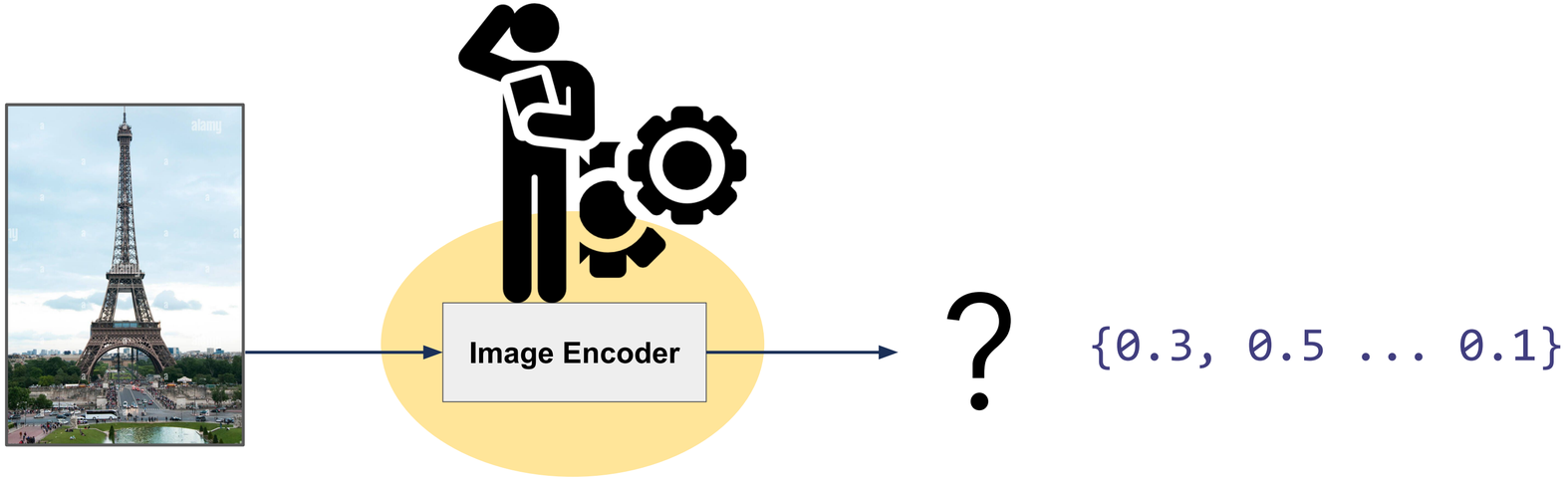
- color, edge, texture etc...

Local Features:

- representation of interest points/regions
- for image stitching or object recognition



Handcrafted features



- Good for low level features
  - Colors, shapes, important keypoints
- But how can we define a priori the characteristics of the image that enable us to recognize complex entities (e.g., a “cat”, or a “tower”)?

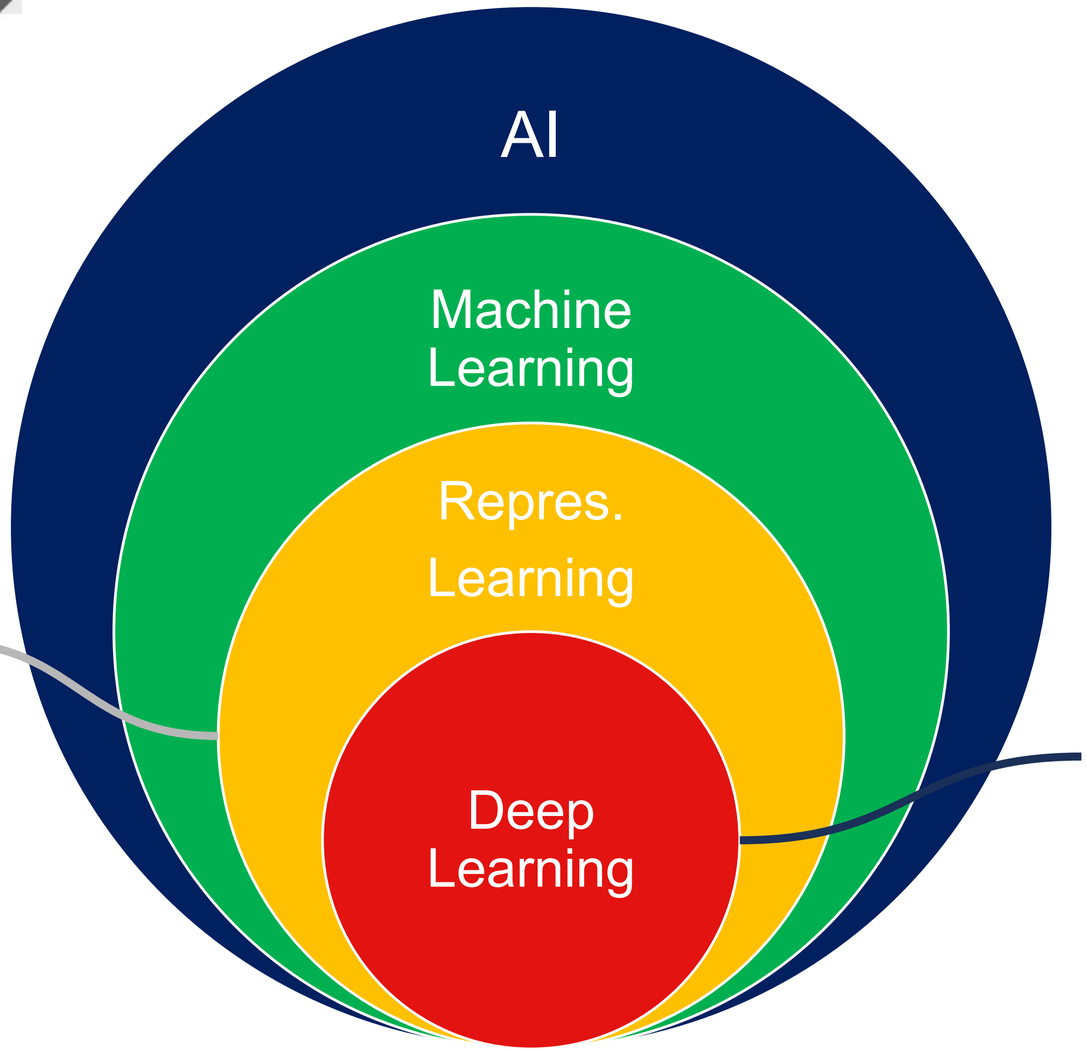
Deep learning (from Nature)

nature International weekly journal of science

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

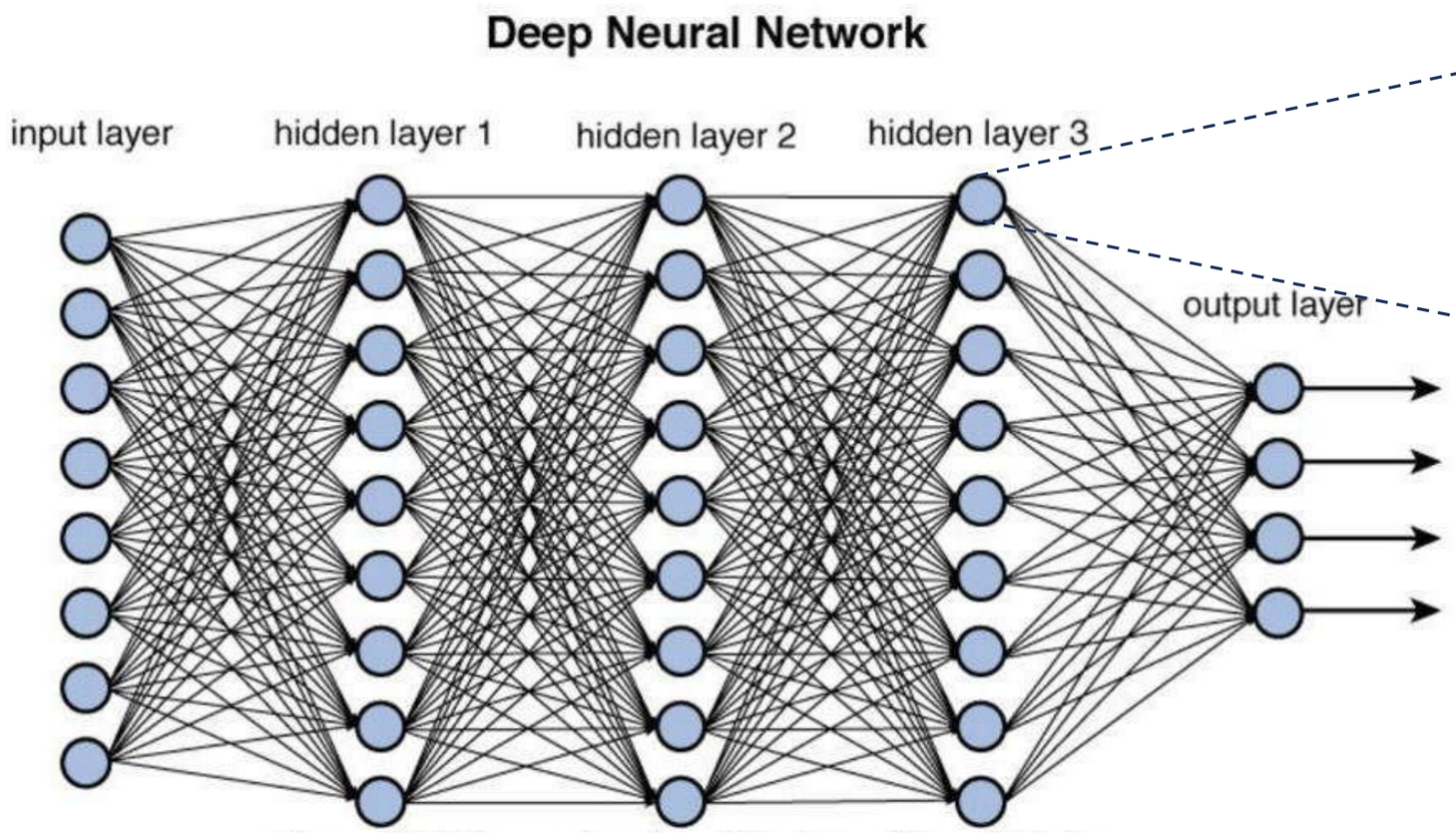
Yann LeCun, Yoshua Bengio & Geoffrey Hinton

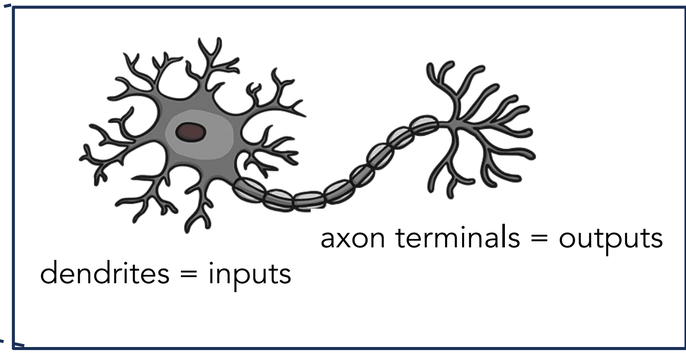
Representations are learned from data!
No more handcrafted!



Representations are learned inside multiple layers of a deep neural network

A Deep Neural Network

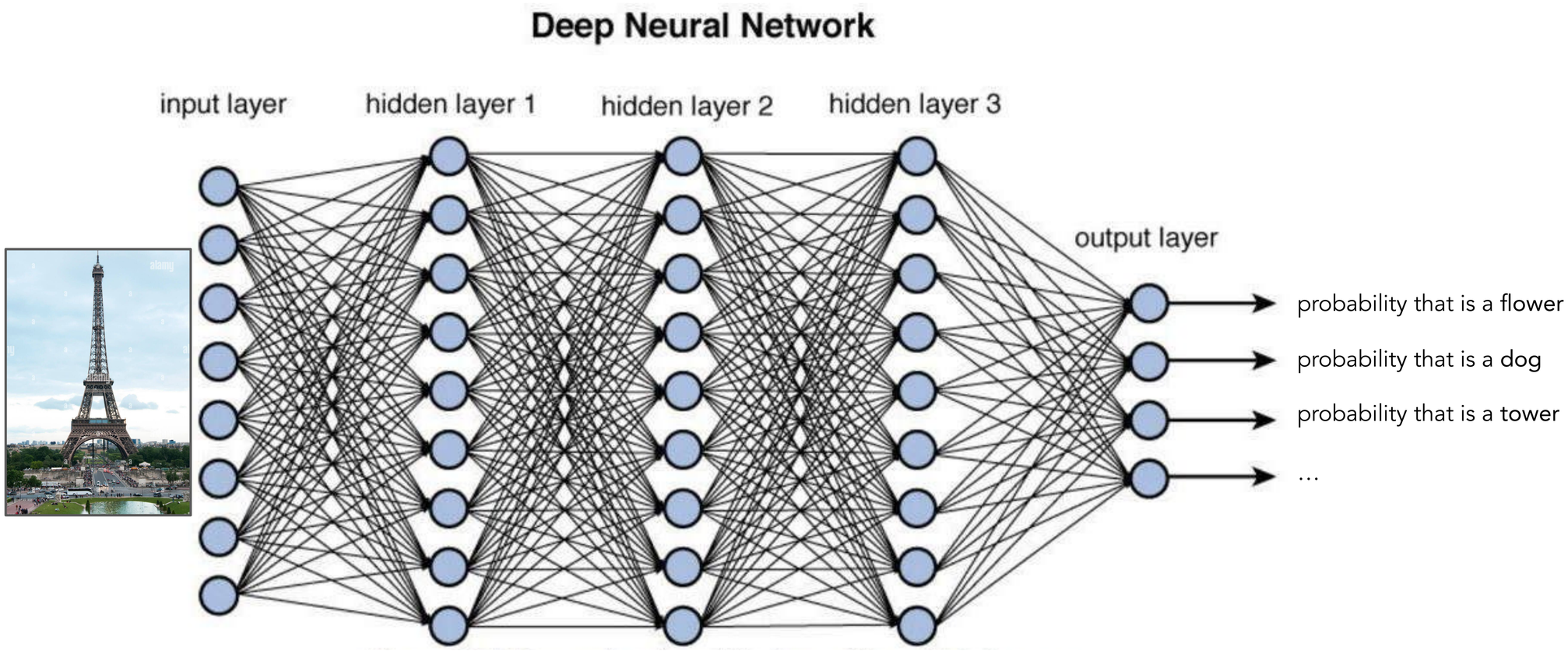




Every neuron in each layer is connected to all (or some) of the neurons of the previous layer

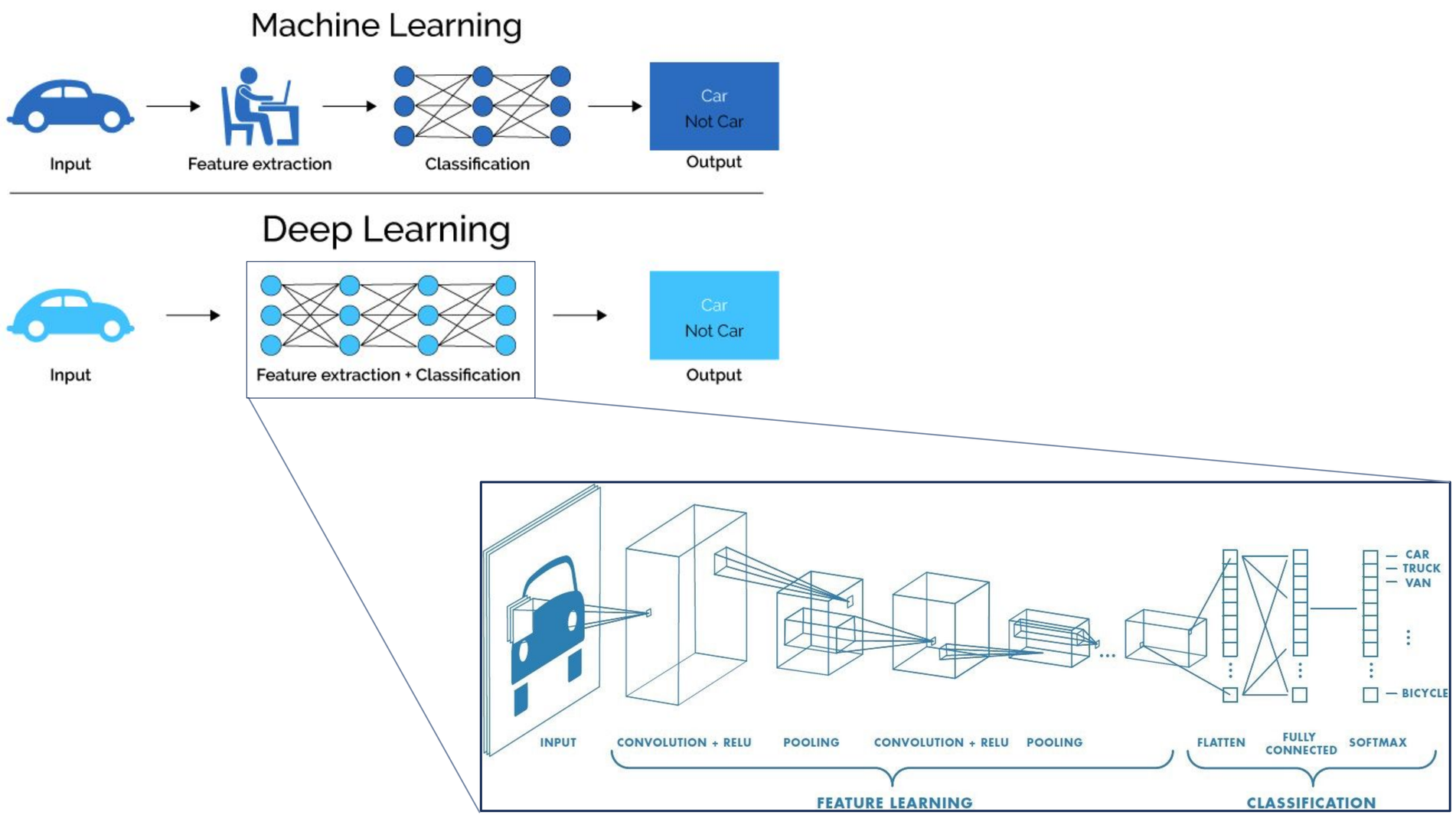
- Feed Forward Neural Networks

A Deep Neural Network



- The input could be an image
- The output the class of the object contained in the image (e.g., a "cat")

Deep Learning vs standard Machine Learning



Deep learning (from Nature)

nature International weekly journal of science

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

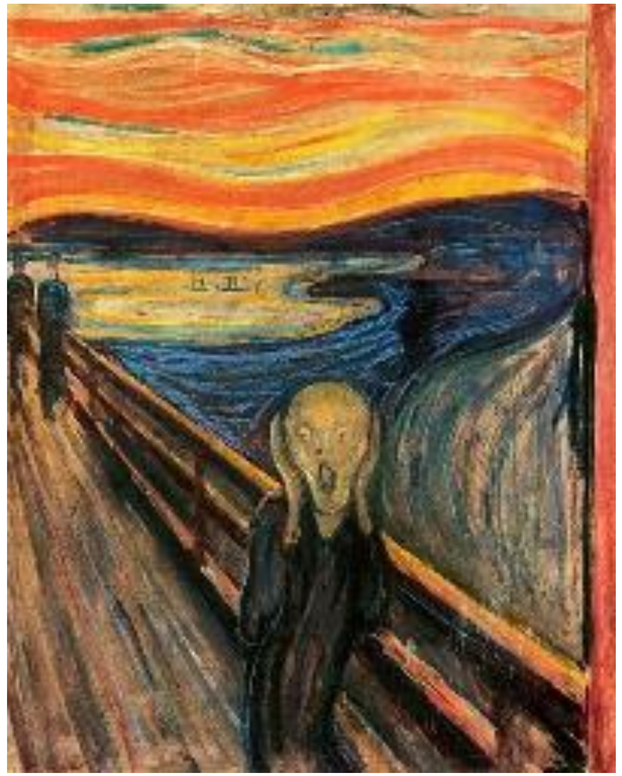
Representation learning methods:

allow a machine to be fed with raw data and to automatically discover the representations needed for classification.

Deep-learning are representation learning methods

- with **multiple levels** of representation, obtained by
- composing simple **modules** that
- transform the representation at one level into a representation at a higher, slightly more abstract level.

The Scream, Edvard Munch



Low-level

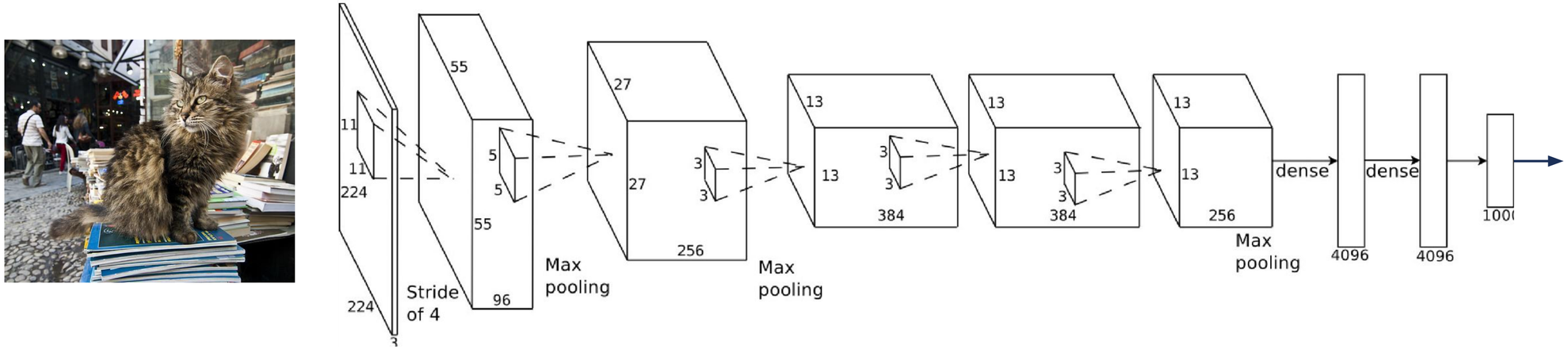
- **The file** at http://upload.wikimedia.org/.../475px-The_Scream.jpg
- **One** of the files of the same picture
- Almost the **same**
- A picture of **the object** at National Gallery, Oslo
- **One of** “The Scream”s by Edvard Munch
- **A** painting by Edvard Munch
- **One** of “The Scream”s by various artists
- **An** expressionist painting
- **A** painting
- **An** hand made object
- **An** artificial object
  being the product of intentional human manufacture

High-level

The training procedure

A deep network should be trained before being used

How is training performed?



The network in this case learns like small childs (direct supervision)

Supervised learning… Sometime it fails

Direct supervision <u>does not always work</u>

- works generally well but it is often overused
- does not work easily with logical or mathematical reasoning
- intuition vs reasoning

A nice side-effect

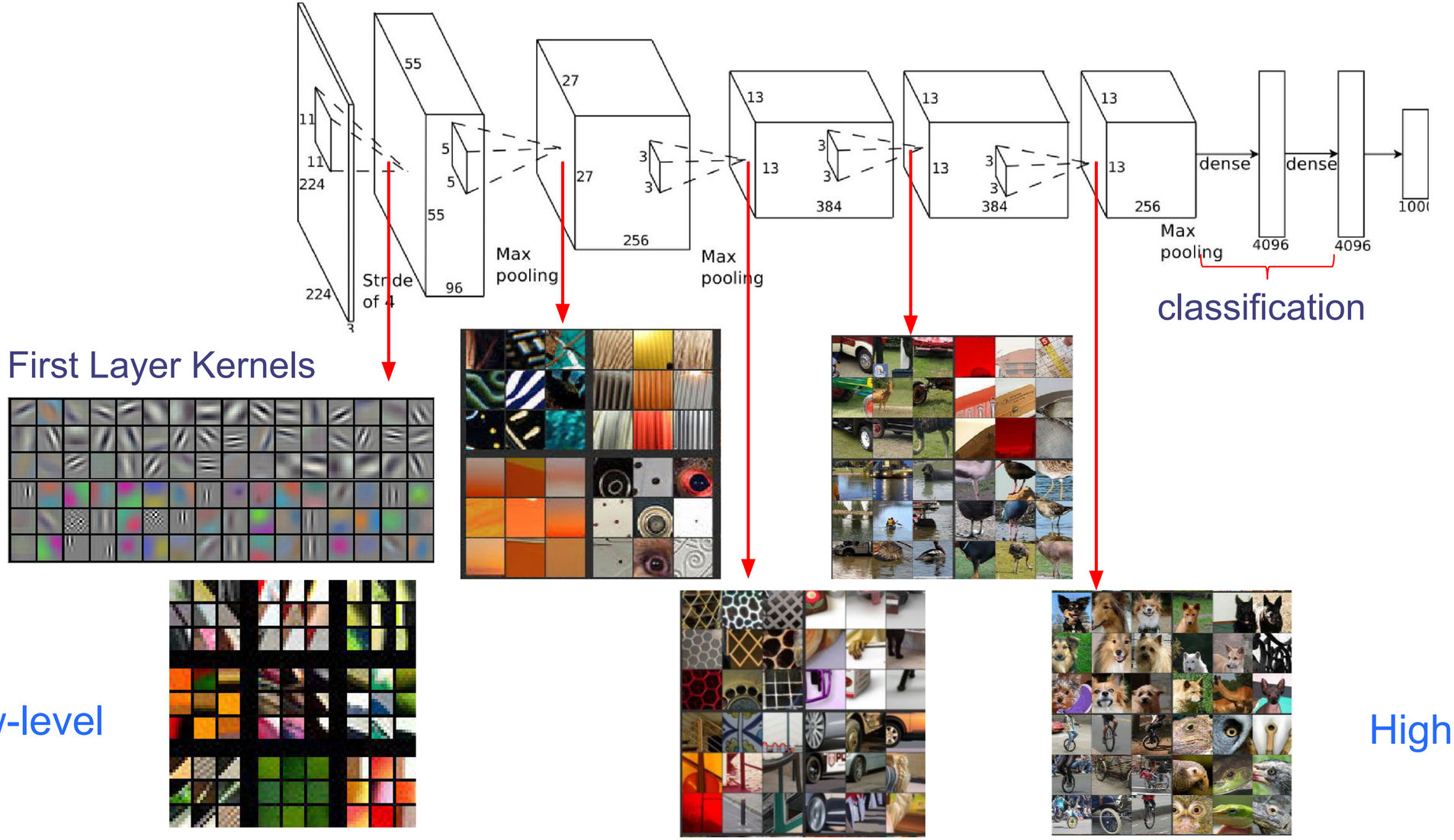
After the training is completed, we can observe the rise of nice representations from the intermediate layers!

The first layers learnt low level details (textures)

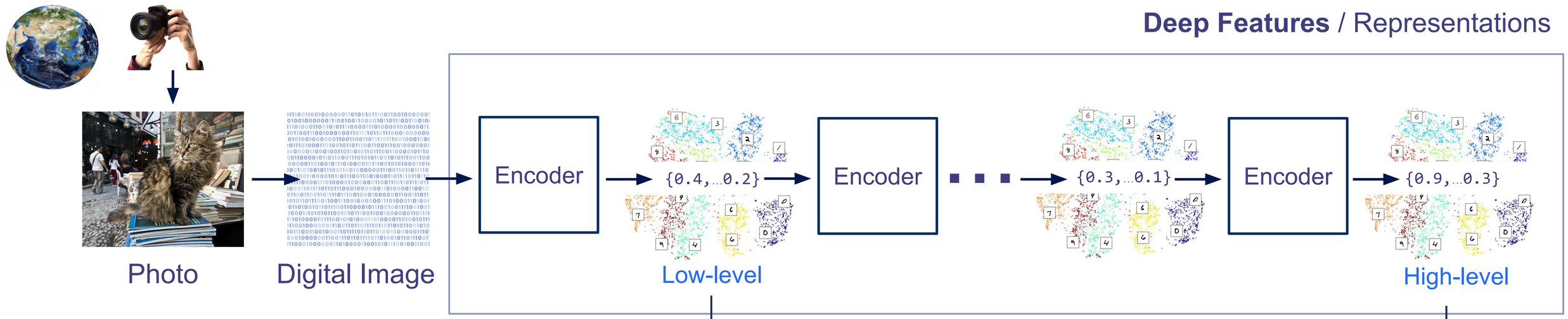
The higher layers learnt very semantic details (faces, objects)

Multiple Levels Of Abstraction

AlexNet, 2012, Trained on a Classification task of 1,000 classes.



The Overall Picture



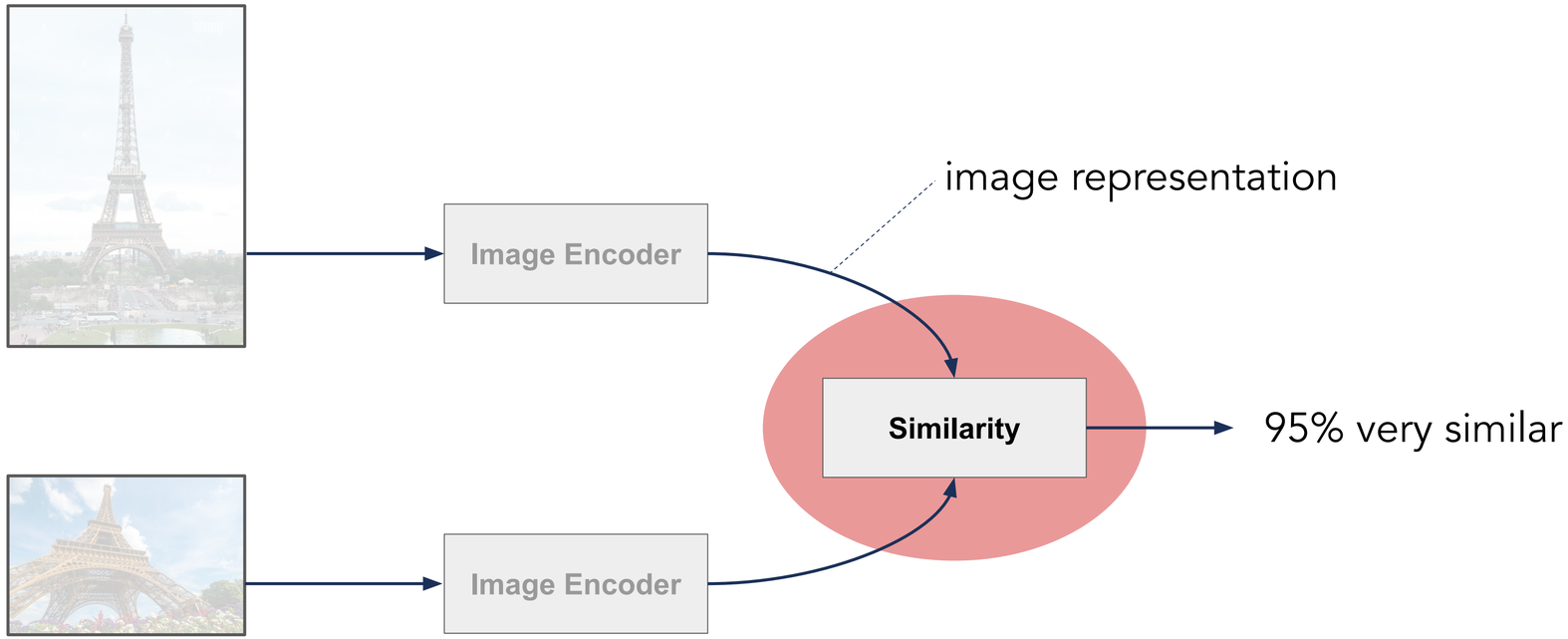
capable of recognizing
- fur texture
- body parts (ears, nose, mouth)

capable of recognizing
- the cat
- a person in the background

- Define a **similarity** between representations
- Similarity between representations as a way to measure the similarity between different images
  - of course, using representations from different levels bring to a different idea of relevance
    - low-level (instance) similarity
    - high-level (semantic) similarity

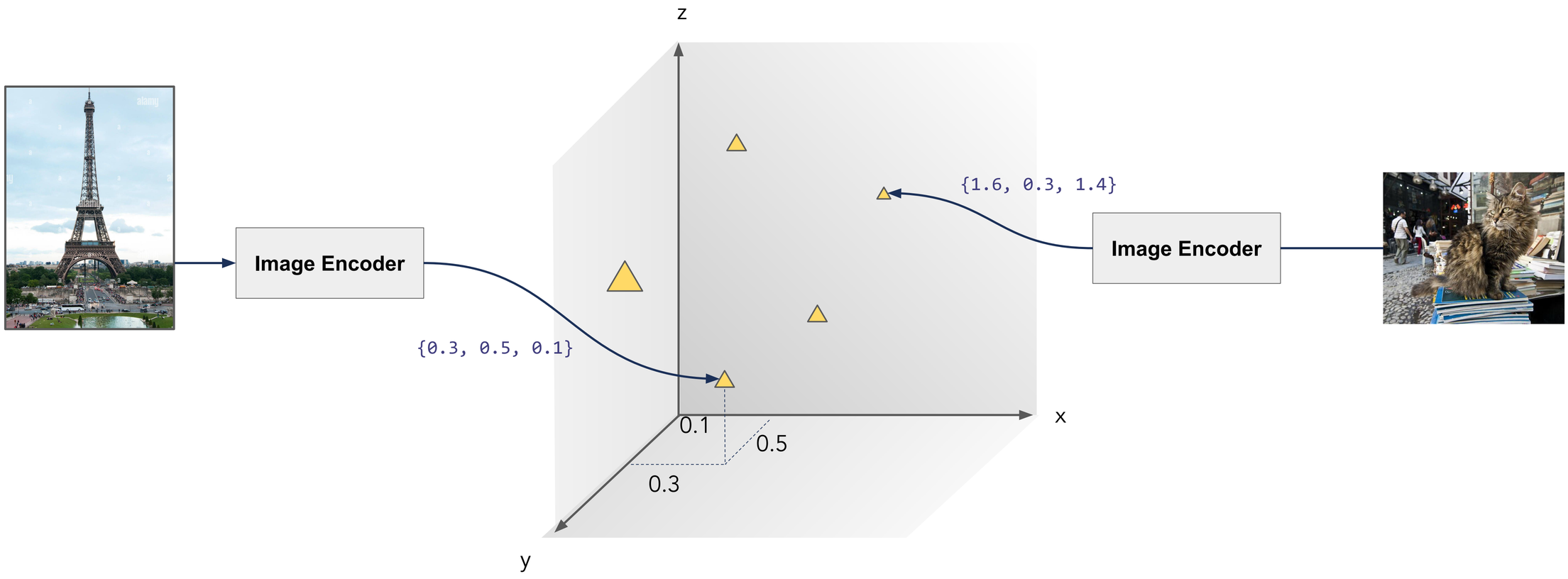
Similarity between Representations

Representations and similarities



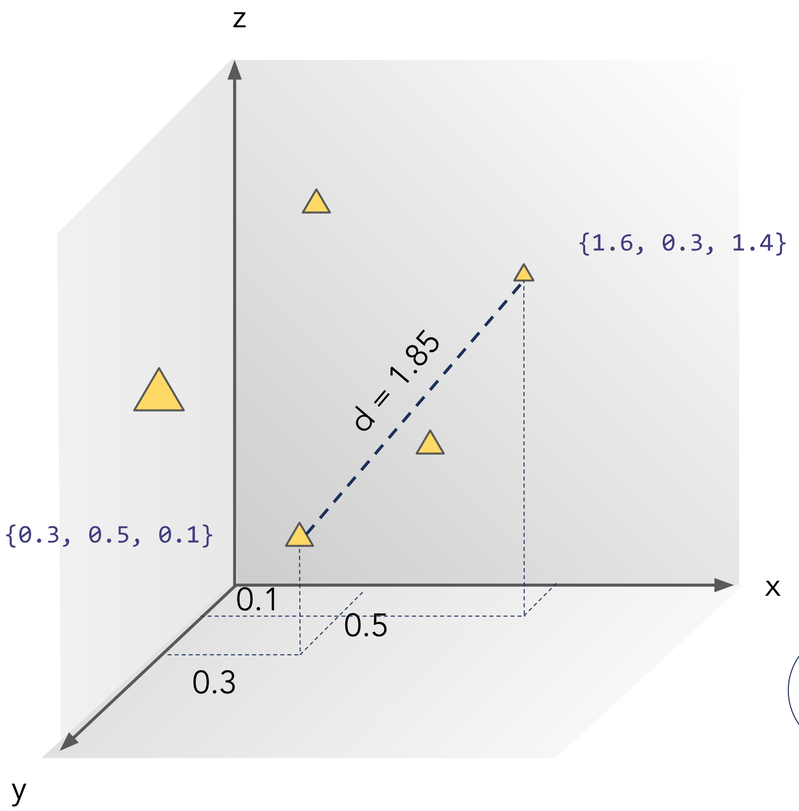
Representations in the space

- The representations are list of numbers `{0.3, 0.5, 0.1}`
- They can be represented in a cartesian space



Distance between representations

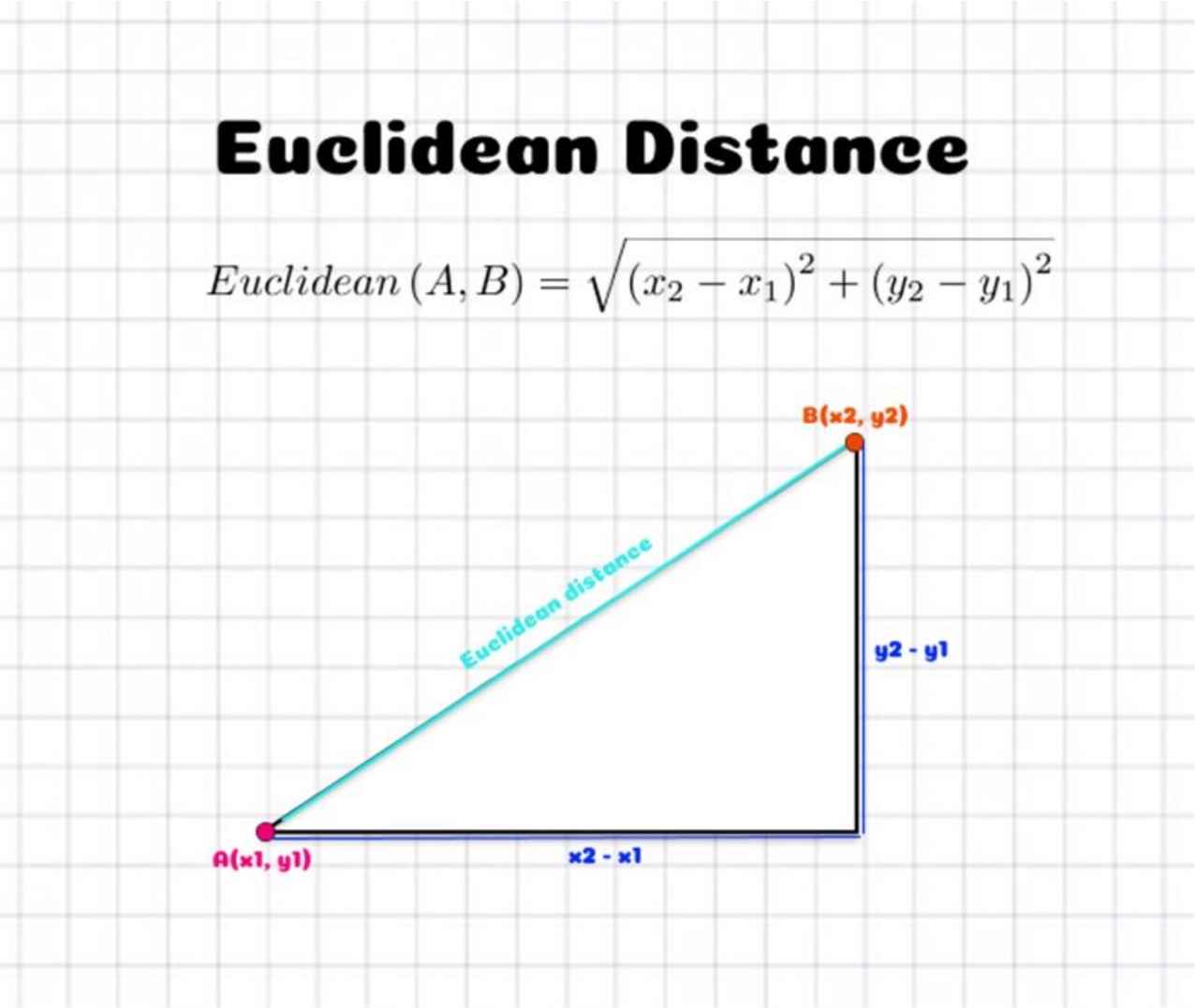
- We can define a distance between representations
- Usually, Euclidean distance, a.k.a. Pythagorean Theorem



$$d = \sqrt{(1.6-0.3)^2 + (0.3-0.5)^2 + (1.4-0.1)^2} \simeq \boxed{1.85}$$

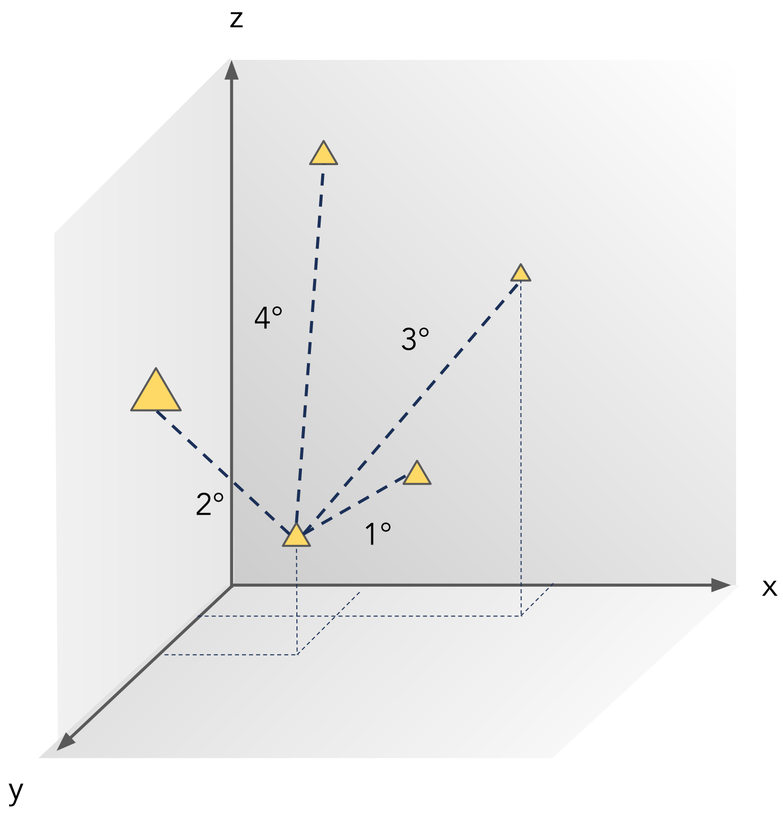
This is how in the game of boules we measure distances between the balls.

Euclidean distance (in 2D)



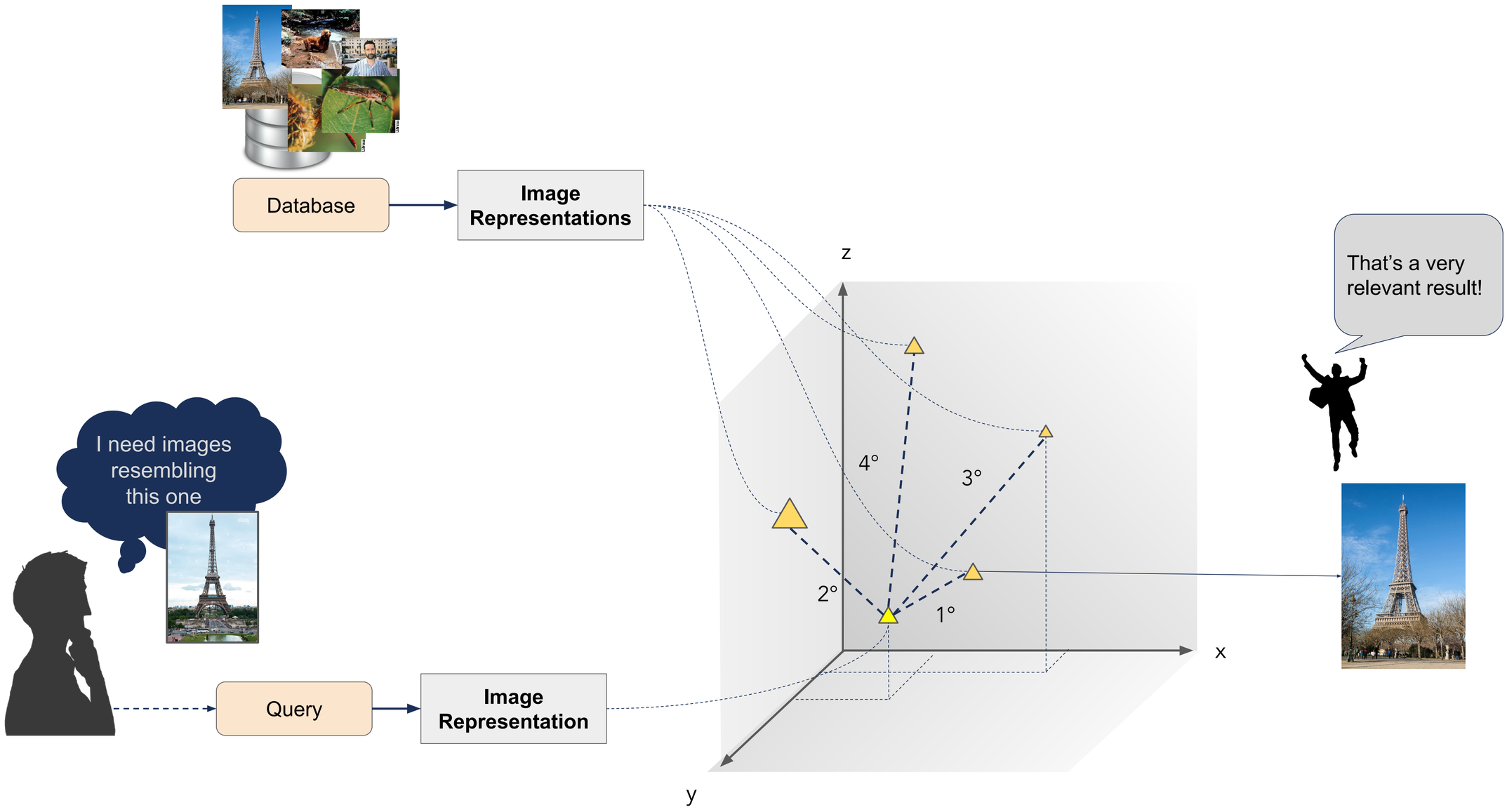
Similarity

- We can define the similarity as the opposite of the distance
- The more distant, the less similar they are

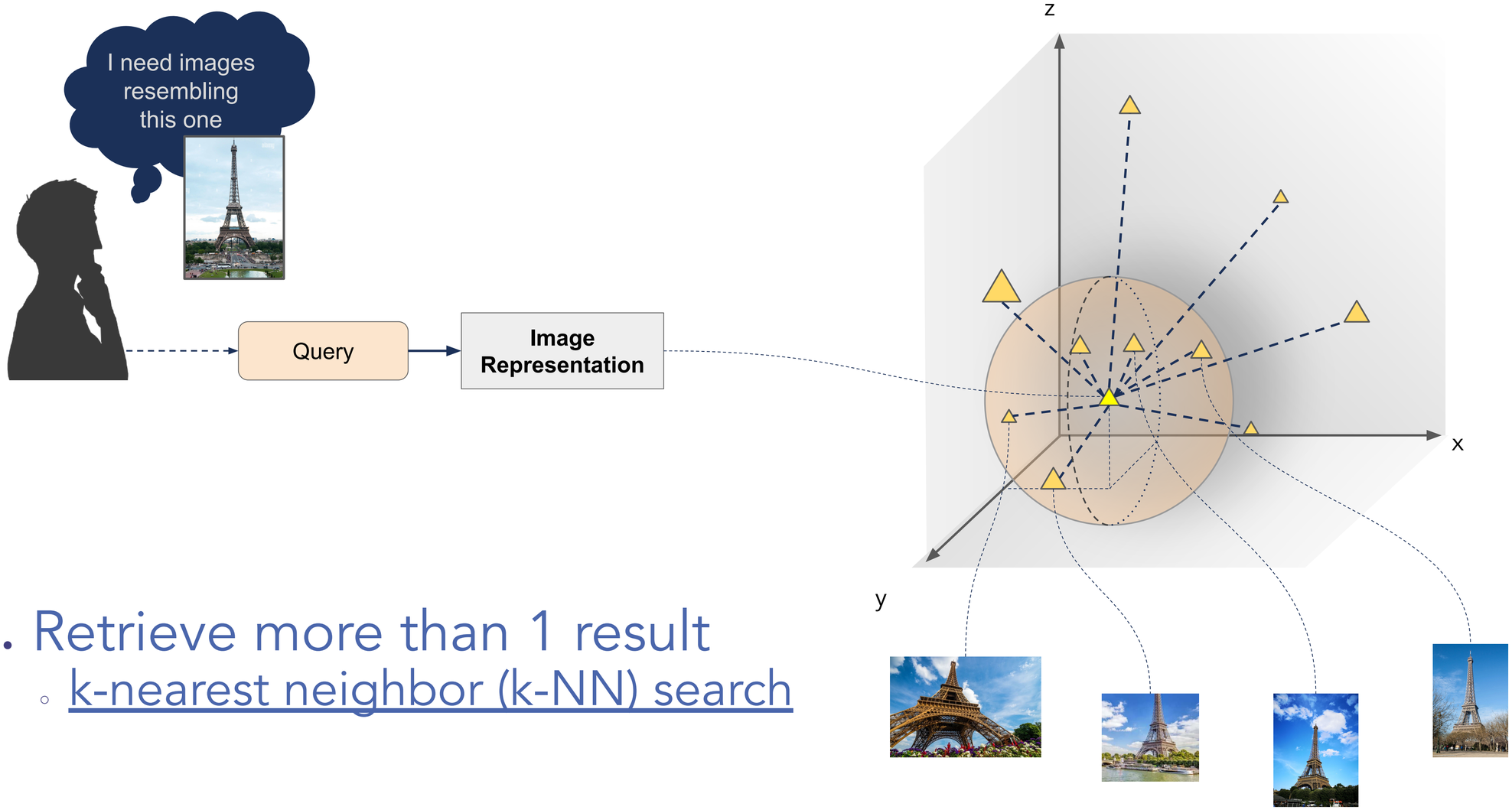


Again, this is actually the method used for assigning points in the game of boules! You have to find out which bowl is the nearest to the target ball

Back to image retrieval



Scaling up this idea

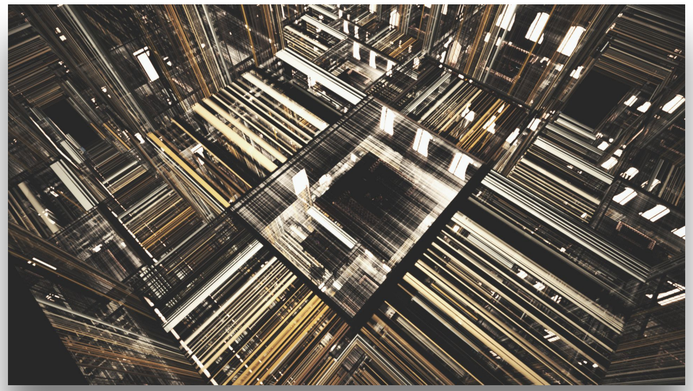


- Retrieve more than 1 result
  - k-nearest neighbor (k-NN) search

Scaling up this idea

Usually, the features have 512 or more dimensions, not just 3

- Difficult to visualize for us, that we live only in a 3-dimensional world
- Mathematically, this is possible without loss of generality
  - The Euclidean distance is still defined
- High-dimensional representations carry more information



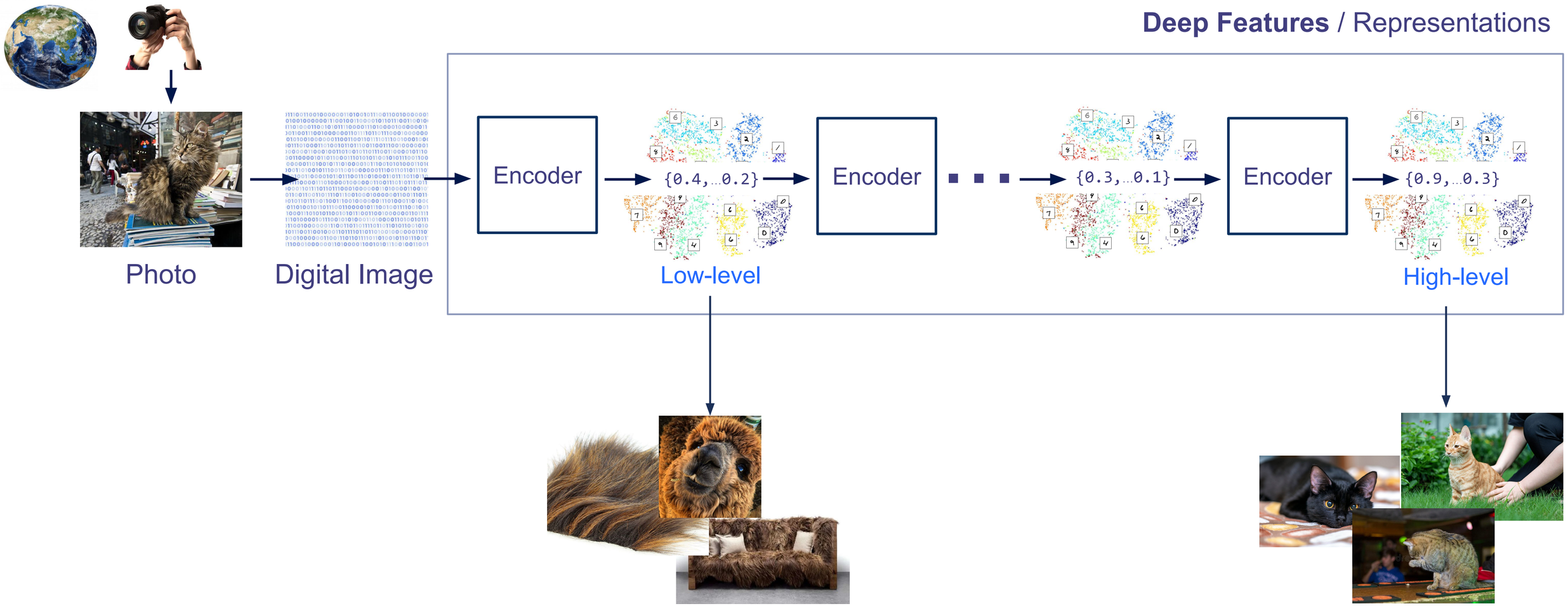
$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

Operatively

1. Compute representations from all the images in the database $D = \{I_1, I_2, \dots I_n\}$
2. Compute representation from the query image I_q 
3. Compute the Euclidean distances between q and all the images in D
4. K-nearest-neighbor search: sort these distances in decreasing order (or, in other words, by increasing similarity) and take the first k results

Different representations, different similarities

By taking representations at different layers of the deep network, we give a different meaning to our similarity measure



Elements of text-to-image search

Using text as a query

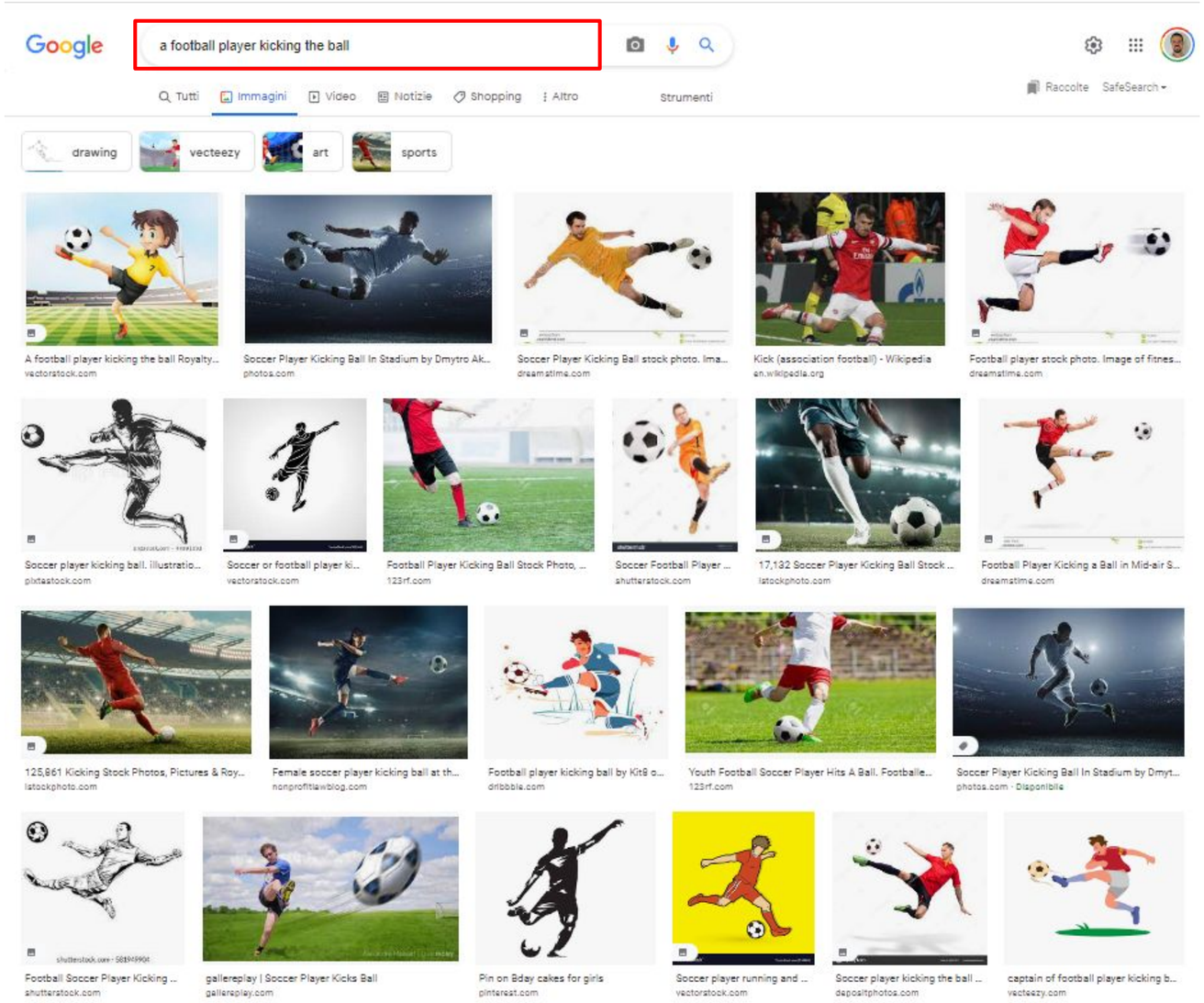
- As of now, we used an image as query for retrieving other images
  - image → image
- What about using another modality as a query?
  - text → image

image → image

text → image

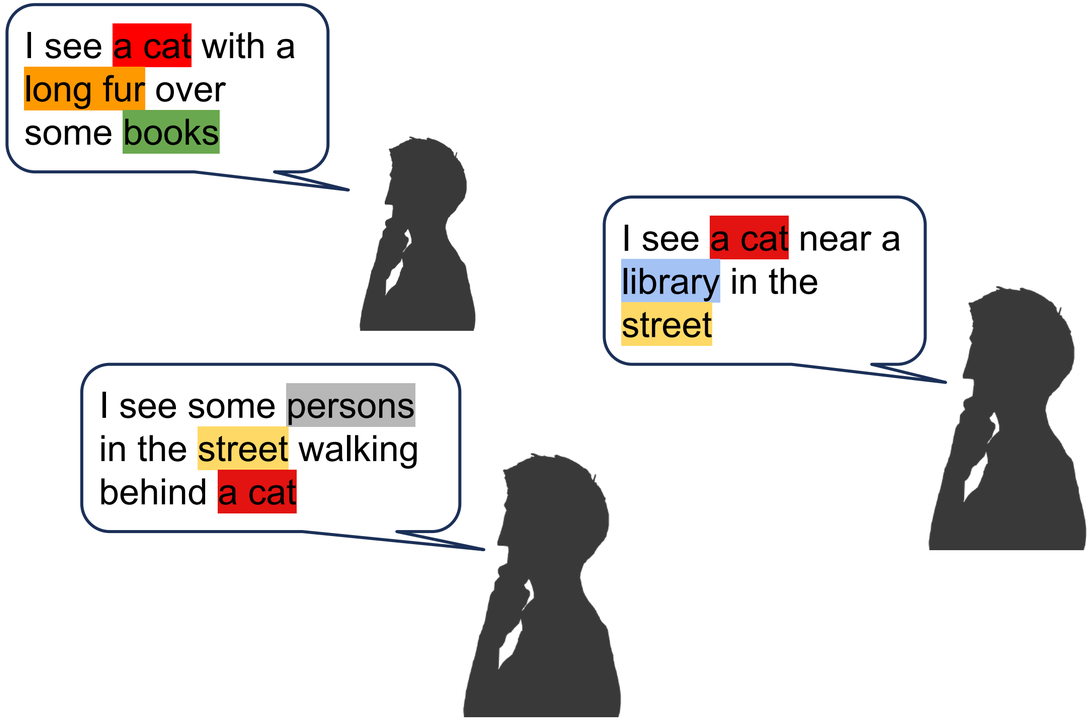
"A football player kicked the ball"

Google image search



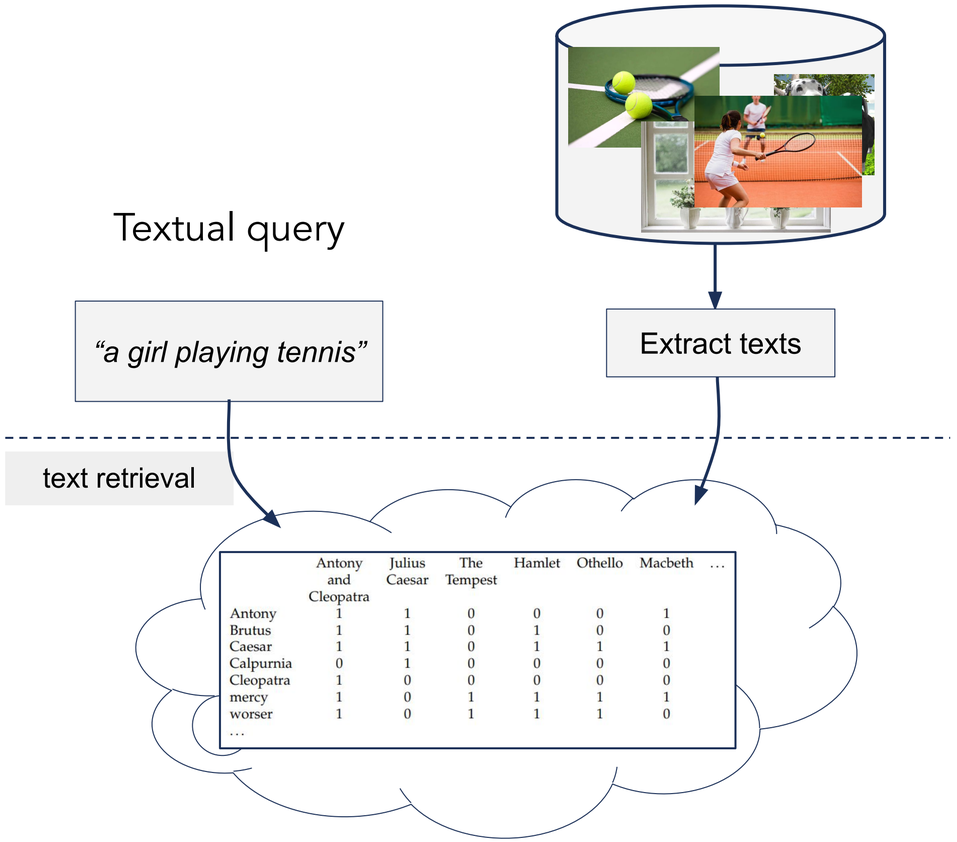
Advantages

- In many cases it is not convenient to search using an image as a query
  - Imagine to always search images in Google using other images
- The natural language is natively less ambiguous than an image



A trivial solution

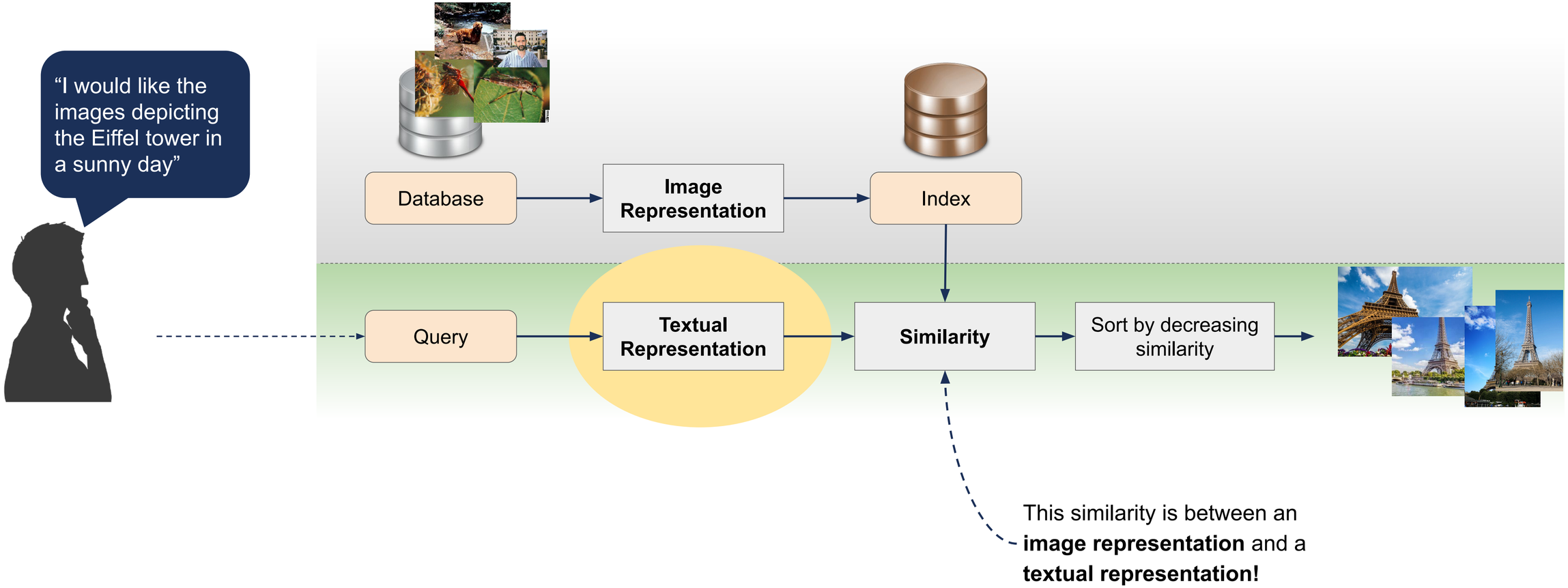
Use the textual metadata associated to the image (e.g., in the *alt* text) to perform a textual search



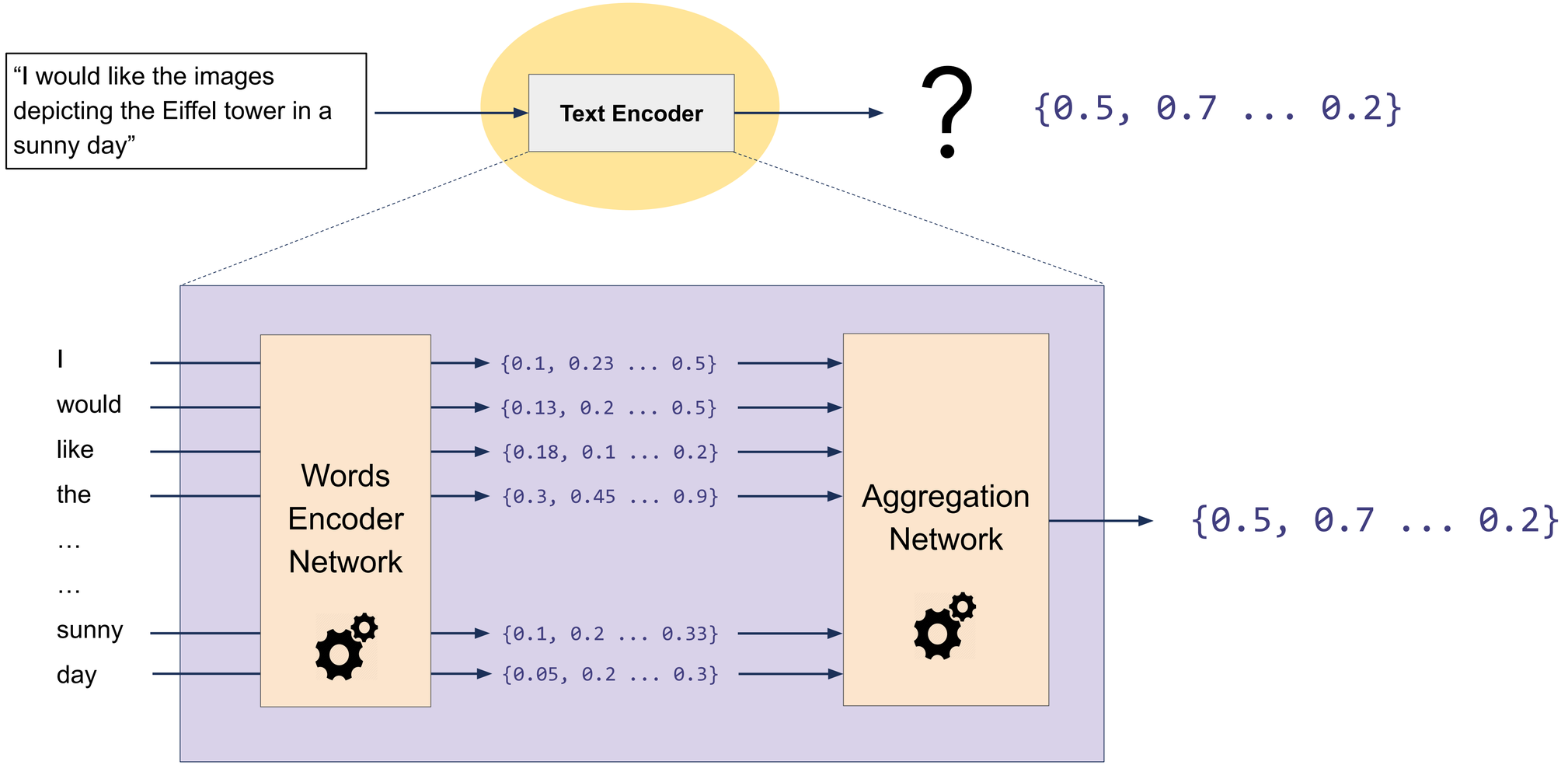
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	…
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
…							

- Google partially works in this way
- What if we don't have textual descriptions for images?

Text-to-image similarity search



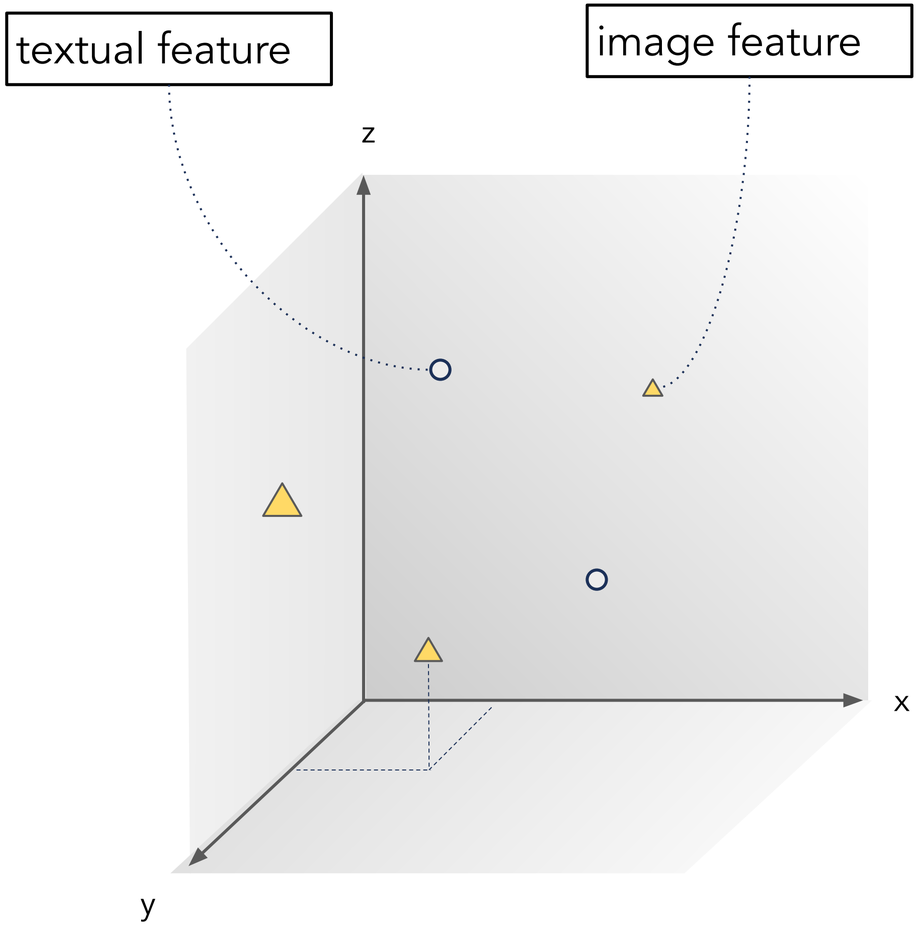
A deep network for texts!



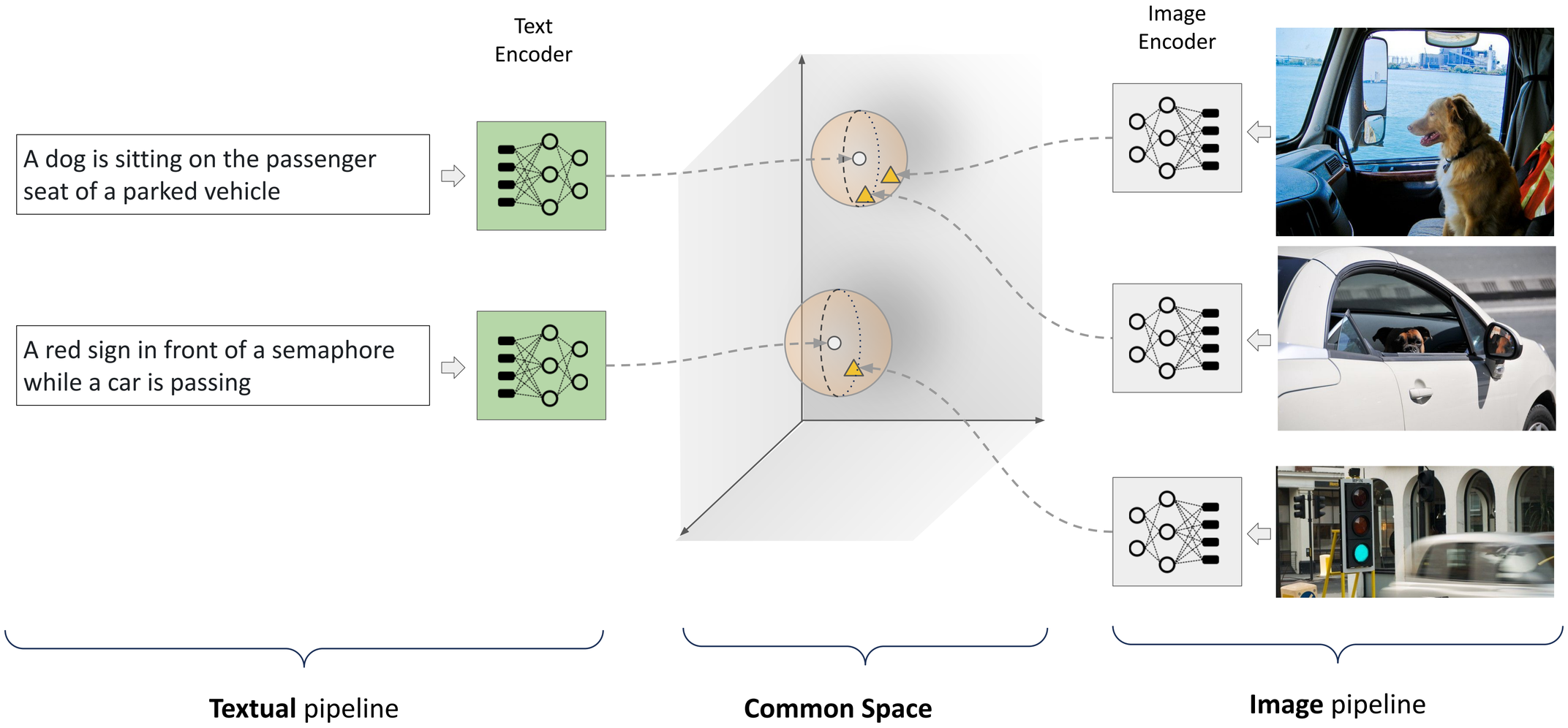
The similarity computation

If the image representations and the textual representations have the same dimensions, they can be compared in the same space!

- We reuse the similarity framework developed before for image-image searches
- We can compute the Euclidean distance between textual and image representations

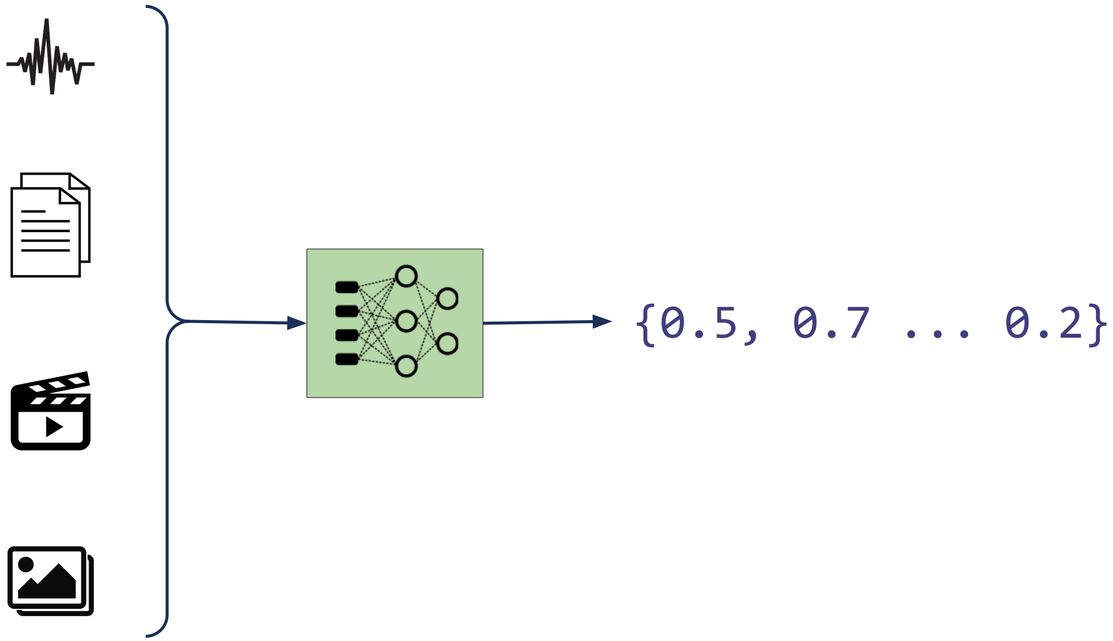


The common visual-textual space



Generalizing to multiple modalities

Every multimedia object can be converted in a numerical representation using deep neural networks



Using k-NN search in a common space, we can easily perform:

text → image

text → text

text → video

image → text

audio → text

text → audio, video

The end!

Questions?