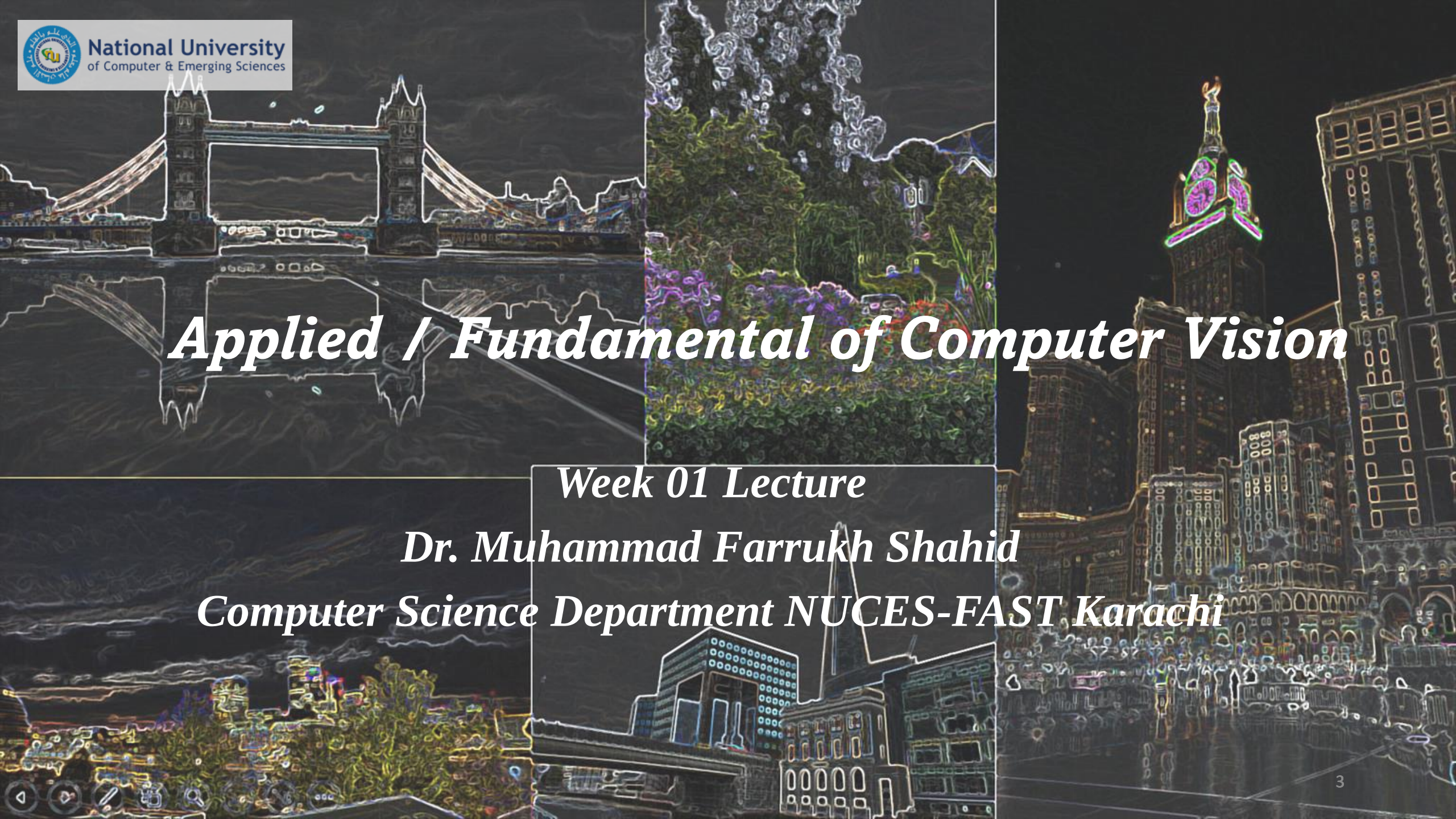
https://careers.google.com/jobs/results/8921886874603591O-student-researcher-2023/



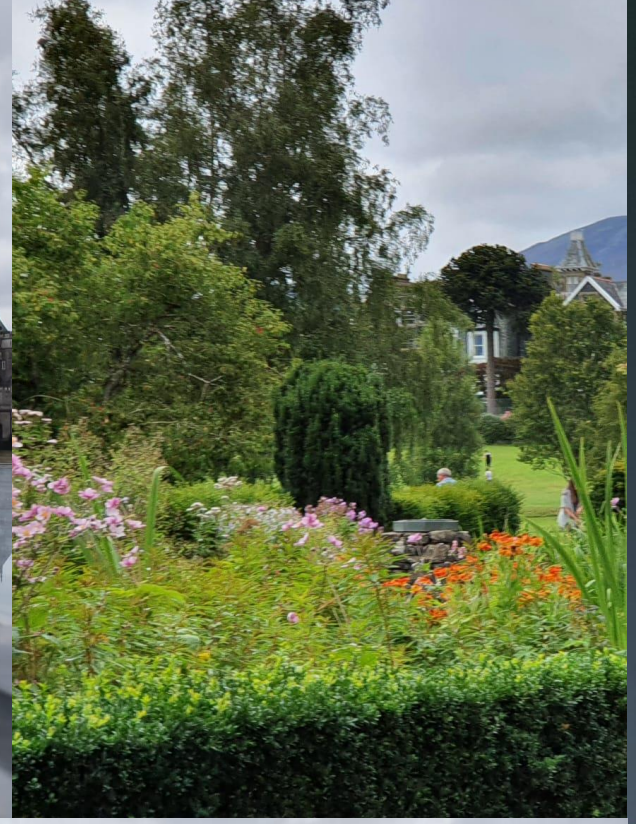
National University of Computer & Emerging Sciences

Applied / Fundamental of Computer Vision

Week 01 Lecture

Dr. Muhammad Farrukh Shahid

Computer Science Department NUCES-FAST Karachi




DAR AL-TAWHID

National University
of Computer & Emerging Sciences

Presentation Profile

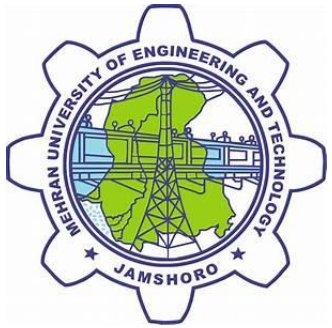
- ***Introduction***
- ***Motivation***
- ***Objectives***
- ***Computer Vision***
- ***History***
- ***Key challenges***
- ***Some examples***
- ***Advantages and Disadvantages***
- ***Conclusion***

About myself

QUEEN MARY
UNIVERSITY OF LONDON

UNIVERSITÀ DEGLI STUDI
DI GENOVA

Joint Doctorate in the AI-enabled Cognitive Radio Internet of Things Networks



Bachelor (Telecommunication)
Masters Communication Systems and Networks

Teaching Experiences at National and International as a Lecturer and Researcher

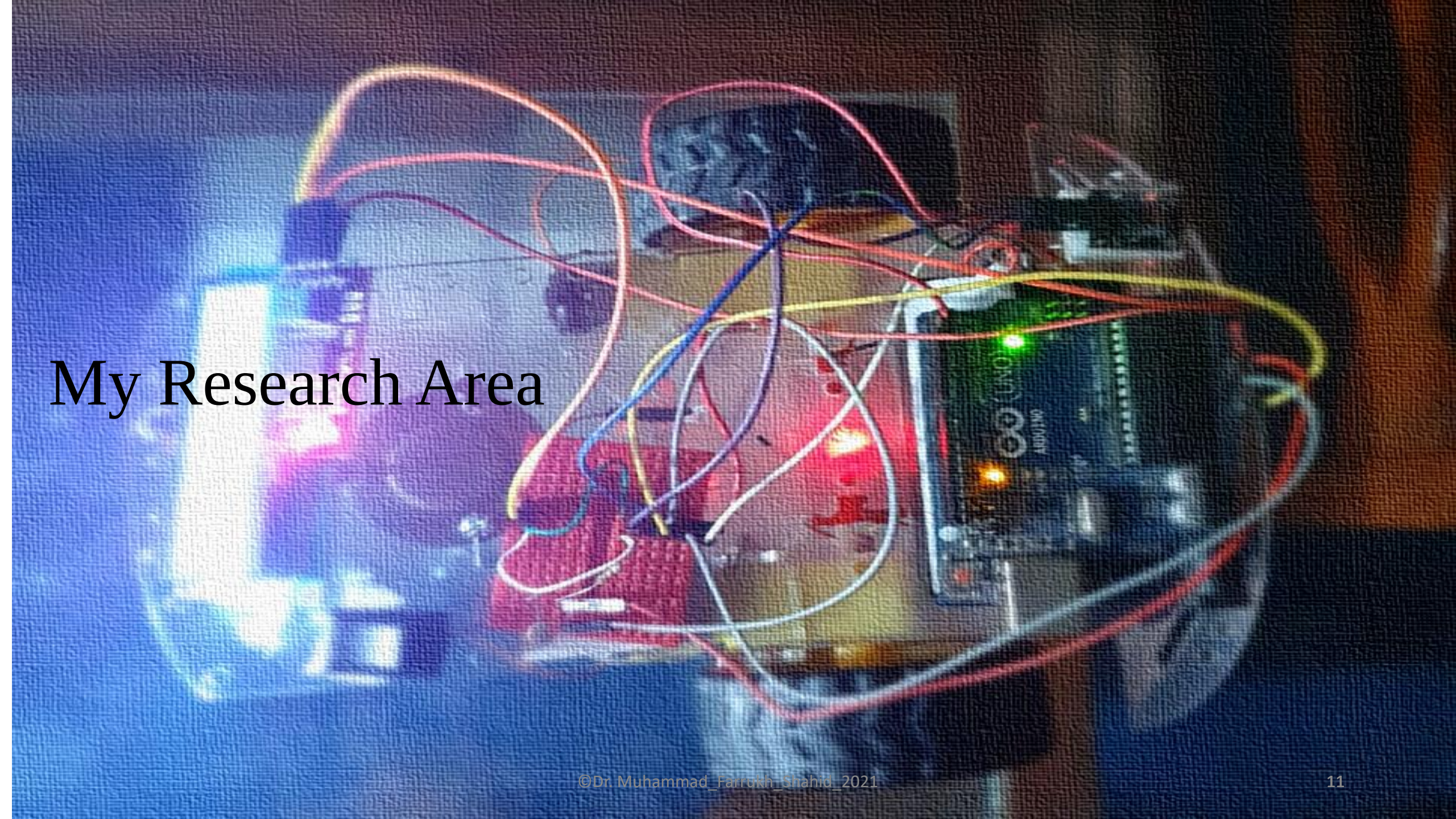
National University
of Computer & Emerging Sciences

Assistant Professor in the Computer Science Department

Faculty of Artificial Intelligence and Data Science

FAST National University of Computing and Emerging Science (NUCES) Karachi Pakistan.

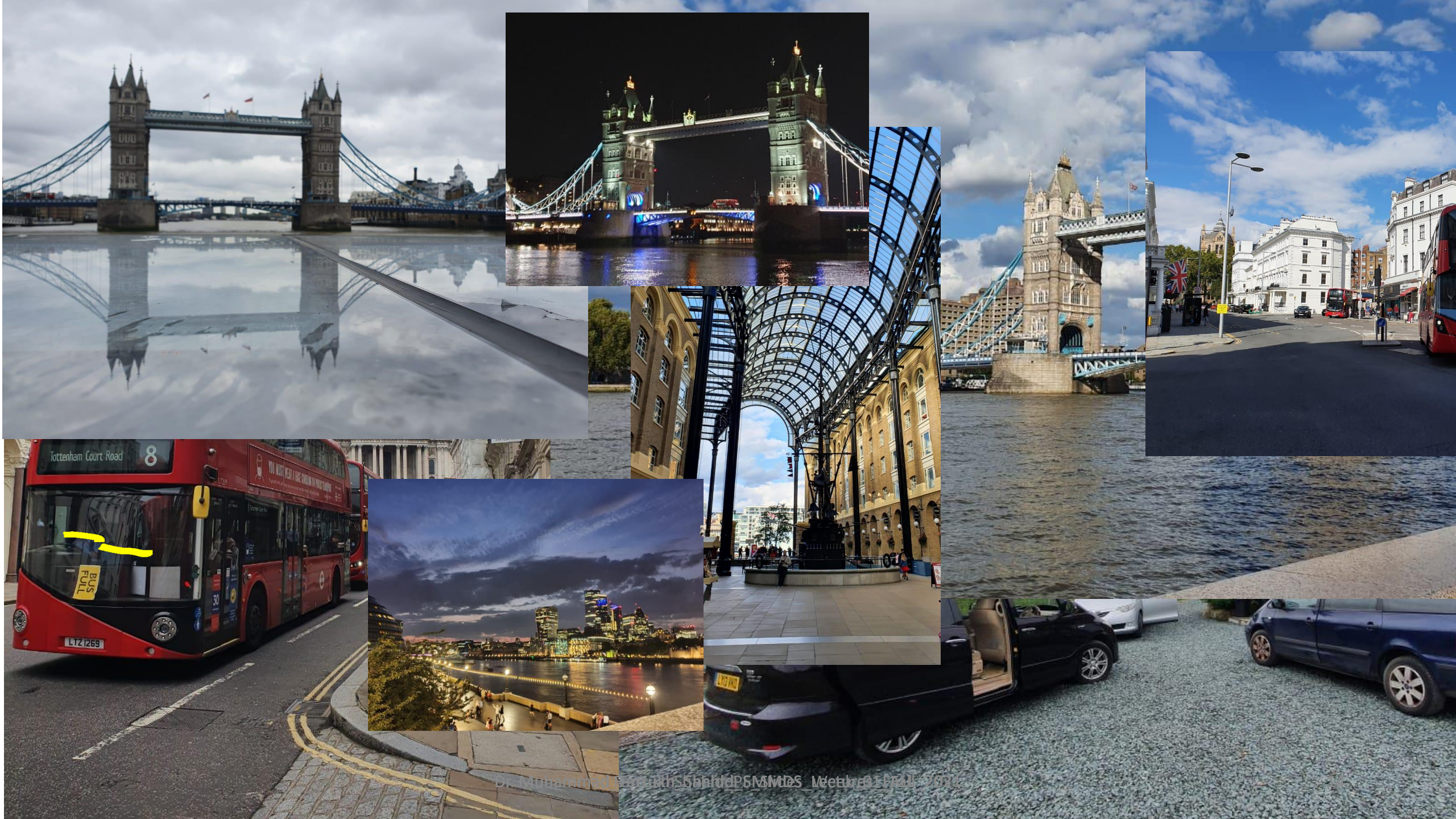
My Research Area

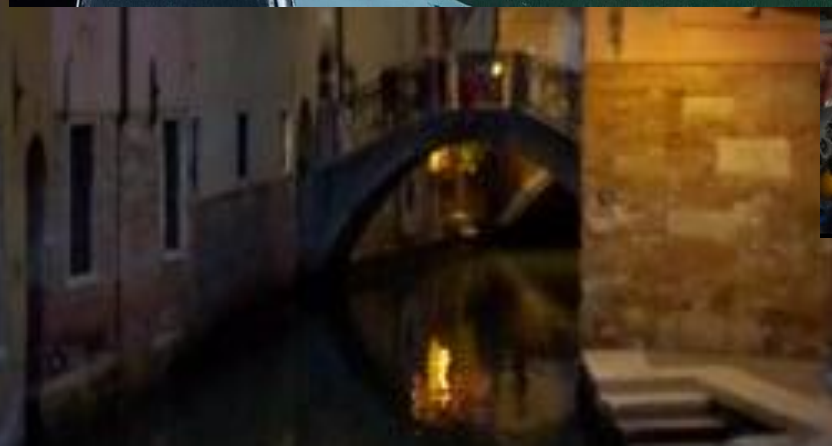


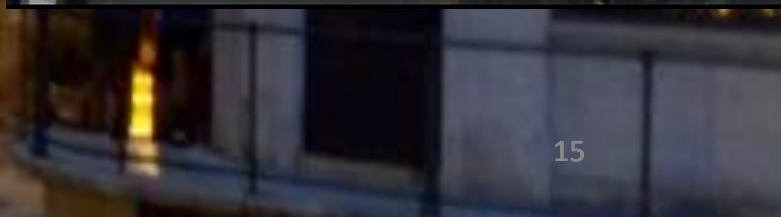
My Research Area (Not Limited But Majors)

- AI techniques and Probabilistic Models for anomalies detection in Cognitive Radio Internet of Things Networks (CR-IoT Network)
- Deep Learning Models for Autonomous Vehicles and Unmanned Aerial Vehicles (UAVs)
- Signal Processing techniques for 5 / 6 G Mobile Networks

My Passion along with Teaching and Research


Tottenham Court Road
8
BUS FULL
LTZ 1269







Recent Projects (Under my Supervision)

1. Generative models on Smart Agriculture
2. Smart Agriculture Using Machine Learning
3. Automated mangrove forest mapping and change detection using Landsat-8 data and machine learning
4. Smart Medical Devices
5. Recent Advancements for deep learning and NLP

6. Identifying and detecting human posters for diverse applications (Prayers and Gym)
7. Automous Crop management framework for Agriculture
8. Smart Medical Devices
9. AI enabled smart agriculture system
10. Data driven approach to build smart medical devices

Recent Projects (International Collaboration)

1. **Block Chain based system for bus approval process**. *(Working as an External Investigator with King Abdul Aziz University KSA Principal Investigators)* ***Status:*** *Accepted by the Institutional Funding Program for Research and Development (Indexed Scientific Publication) research grant number IFPIP: 620-611-1443*

2. AI enabling approach using Deep learning for detecting abnormalities in CR-IoT for medical application. (*Working as an External Investigator with King Abdul Aziz University KSA Principal Investigators*) Status: Project proposal submitted and tentatively approved, waiting to get official approval.

Recent Achievements

Won the biggest ICT Award of Pakistan P@sha Award On 2nd November 2022 Lahore

ZEENAT.AI

ZEENAT.AI at International Level

International Asia Pacific ICT Alliance(APICTA) Awards 2022

held on 9th December 2022 in Islamabad Pakistan

National University
of Computer & Emerging Sciences

Virtual Apparel TryRoom

A Next-Gen Apparel Shopping Experience

Abstract

Virtual Apparel Try Room is a complete size solution for online apparel stores to facilitate their customers with size recommendations and help them virtually try-on clothes through AR on any of their browser-based devices leading to decreased returns and increased revenue for the apparel stores. It included three basic modules:

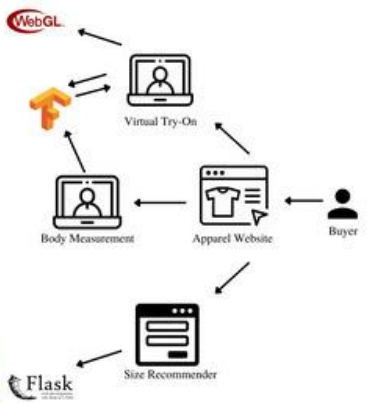
1. Size Recommender
2. Body Measurements
3. Virtual Try-On

Size recommender inputs basic information of the user such as height, weight and age and recommends size with the help of a machine learning algorithm.

Body Measurements module measures users' lengths of key body parts through computer vision on top of TensorFlowJS's pose estimation models

Virtual Try-On module tries to give the best and most realistic view of how apparel will look using computer vision and AR technology

Workflow



Objective

We want to revolutionize the fashion e-commerce industry by allowing online apparel stores' customers to have a futuristic shopping experience as well as increase the revenue of these online stores by cutting the cost of frequent returns which are very costly.

Since size uncertainty and fashion dilemma are two major reasons for buyers to return apparel, what measures can be taken to counter these two problems? If a user gets to have the perfect fit in the first go, they are very unlikely to opt for a return or an exchange. To enable online apparel stores to deliver a perfect fit in the first go, we landed on this solution which helps the buyers to get their body measurements and subsequently their size recommendations as well as virtual try on to have an idea of how the apparel will look on you once you wear it.

Features

- SaaS Model. Can be integrated into websites easily
- Intuitive flow for better user experience
- Very minimum buyers' involvement
- **Body Measurement** has a minimal error threshold
- ML model for **Size Recommendation** has 86% accuracy.
- **Virtual Try-On** is very responsive and maps apparel in real-time
- The buyer doesn't require to leave the website. VATR is integrated in the apparel stores' website

Team

Supervisor
Dr. Muhammad Farrukh Shahid
(m.farrukhshahid@nu.edu.pk)
Co-Supervisor
Mohsin Ali (mohsin.ali@nu.edu.pk)
Group Members
Efshal Ahmed (k180352@nu.edu.pk)
Syed Zeerak Ibrahim (k181197@nu.edu.pk)
Jaafar Bin Farooq (k181294@nu.edu.pk)

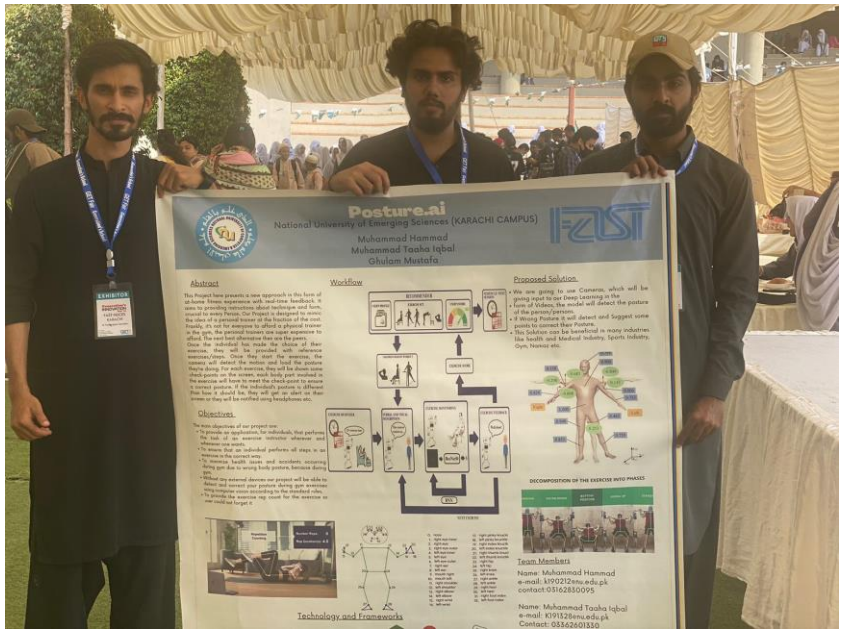
Project Achievements

- Incubated at National Incubation Center(NIC) Karachi May 2022
- 1st Prize Winner in Project Exhibition held by Securiti.ai at Developers' Day 22
- 2nd Prize Winner HackFest held at IBA Karachi
- 2nd Prize Winner in Hackathon at Developers' Day 22

Incubated at National Incubation Center NIC Karachi May 2022

1st Prize Winner in Project Competition held by Scurinty.ai at FAST May 2022

2nd Prize Winner HackFest held at IBA Karachi 2022



Won the highest Innovative Project Idea Award in Generation Innovative Fair 2022 Karachi

Poster.AI

AI GENERATED VIDEO CHAPTERS

FINAL YEAR PROJECT (2018-2022)

ABSTRACT#

E-learning environments are heavily dependent on videos as the main media to deliver lectures to learners. Despite the merits of video-based lectures, new challenges can paralyze the learning process. Challenges that deal with video content accessibility, such as searching, retrieving, explaining, matching, organizing, and even summarizing these contents, significantly limit the potential of video-based learning. In this project, we propose a 2 step approach which includes a) segmenting a video into coherent topics and b) generating titles for every segment. For the segmentation step, we make use of sentence transformers for generating sentence embeddings which are then passed to an LSTM model that outputs the segmentation boundaries. For title generation, we fine-tune the T5 Transformer on our dataset to generate one line summaries for every segment. We compare our models performance with a baseline TextTiling algorithm. Our model beats the baseline by a considerable margin and is able to achieve state-of-the-art 92% accuracy and 0.18 pk on the VTSSUM dataset, a benchmark dataset for the task of video segmentation and summarization. We further analyze our model size and find that we can build our model with many fewer parameters while keeping good performance, thus facilitating real-world applications.

USER INTERFACE

OBJECTIVES

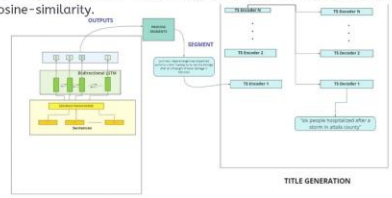
The goal of this project is to successfully implement a supervised text segmentation based on neural LSTM architecture. Our motivation for this project is twofold:

- Text segmentation is an important task within NLP, but is also an extremely useful task for a variety of purposes. In NLP, text segmentation is important for other tasks like summarization, context understanding, and question-answering.
- ·Furthermore, segmentation in general is an extremely useful task. Our main nspiration for this project was realizing that audio content is exploding across platforms (such as podcasts, YouTube, etc) and has become a very popular medium for both learning and entertainment. While tools for transcription have been developed, audio content remains difficult to efficiently search, navigate through, and index. For example, lecture videos can be hours long, and being able to segment an episode into relevant chunks could greatly increase a user's efficiency in learning from online lectures.

ARCHITECTURE

For the text segmentation task, we are using a neural model of Bidirectional LSTM networks. LSTMs keep longer term memories to be able to encode long range dependencies. They do so using gate which control what information can flow to a memory cell state. A bidirectional LSTM helps capture information from both directions into embeddings.

Our first sub network, which we will refer to as the Sentence Representation Network, computes 768-dimensional embeddings of the input sentences using a pretrained Sentence-Transformer network (SBERT). SBERT is a modification of the BERT network using Siamese and triplet networks that is able to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity.



Our second neural network is a bidirectional LSTM with two layers. We input a sentence embedding into this LSTM, and obtain a distribution as output. Lastly, we run the output of the second LSTM through a classification layer consisting of fully connected layers to obtain a sequence of n vectors ,where n is the number of sentences. During validation/testing, we then apply a softmax function to obtain segmentation probabilities for each of those vectors. For training, we are using cross entropy loss with an Adam optimizer.

For the title generation task, we fine-tuned a Text-to-Text Transfer Transformer(T5) on the VTSSUM dataset. the T5 Transformer consists of a typical encoder-decoder architecture similar to BERT. T5 is trained on 3 objectives that include: language modeling (predicting the next word), BERT-style objective (which is masking/replacing words with a random different words and predicting the original text), and deshuffling (which is shuffling the input randomly and try to predict the original text).

CONCLUSION

Text segmentation, the task of dividing a document into contiguous sections that are semantically and contextually meaningful, is a field of research that will benefit from advancements in sentence representation. In turn, text segmentation has great potential to aid the NLP tasks of information extraction and summarization. While previous work in this domain has primarily investigated unsupervised methods, there appears to be potential for improvement through supervised methods.

For our model, we focused on a supervised LSTM-based model to predict segmentation. Through trial and error, we learned the importance of a having strong baseline and consistent evaluation metric. We also recognized that models must be trained and evaluated on similar data in order to make valid comparisons between results.

FUTURE WORK

Despite our model doing very well, we worry that it leans toward predicting 0 for all examples (specifically, classifying each segment as not ending a segment). Although this leads to good evaluation scores, this output is not helpful to obtain accurate segmentations of a particular document.

For future work, we would like to address the limitations of our baseline by attempting stronger alternatives for creating sentence representations. One interesting alternative takes a weighted average of word vectors, then modifies them using Principal Component Analysis/Singular Value Decomposition. Another potentially interesting method would be to use deep averaging networks to obtain sentence representations. In addition to improving our baseline, we would like to experiment with adding Encoder layers in the neural network

SUPERVISOR
Dr. Zulfiqar Memon

CO-SUPERVISOR
Dr. Farooque Shahid

MEMBERS
Muhammad Ahmed - K180256
Yusha Arif - K181289
Abdul Musawwir - K180185

PyTorch Flask

Got selected by the 10Pearls for the FYP consultation

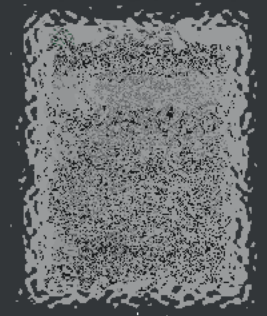
Course Description (Interesting Ingredient)

Marks Distribution

Assessment Item	Number	Weight (%)
Assignments	3	**5**
Quizzes	3	**5**
Computer Vision Challenging Task	1	10
Midterm Exam	2	**10 each**
Project	1	**10**
Final Exam	1	**50**

Teaching Material:

Text Book(s)	**Title**	Computer Vision: Algorithms and Applications Second Edition
	Author	Richard Szeliski
	Publisher	Springer Nature Switzerland AG, 2022
Ref. Book(s)	**Title**	Computer Vision: Principles, Algorithms, Applications, Learning
	Author	E.R. Davies
	Publisher	Academic Press, Elsevier Inc., 2018



Week	Course Contents/Topics	CLO
01	Introduction to CV, brief history, image formation	1

National University

of Computer & Emerging Sciences

02	Image representation, RGB/Grayscale, Otsu/K-means thresholding	1
03	Review of coordinate systems, Image transformations	1
04	Histogram equalization, Matte and Compositing	2
05	Linear nonlinear filtering, Fourier transform filtering	2
06	Fourier transform, filtering in frequency domain, geometric transformations	2
07	Separable filtering, SVD, PCA	2
08	Distance transform, Binary image processing/Morphological Image Processing	3
09	Viola-Jones Algorithm, Integral image, Haar features	3
11	Hough transform, gradient images, HoG	3
13	SIFT, Blob detection (LOG/NLOG), Intro to DL for Computer Vision	4
14	Intro to object detection and YOLO, RoI Pooling, mAP	4
15	Intro to YOLO v7, model reparameterization and scaling, layer aggregation	4

GOOGLE CLASS ROOM

BCS-8A

diiusmf

https://classroom.google.com/c/NTg1ODE5NTcyMTg4?cjc=diiusmf

GOOGLE CLASS ROOM

BCS-6A

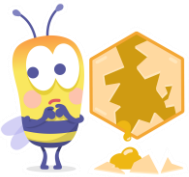
pwmrupu

https://classroom.google.com/c/NTg1ODIyMzg3MjEw?cjc=pwmrupu

Class Room Policies – Pay attention

- Don’t come into the class if you are late
- Don’t sit in my class for time passing or you have less interest.
- Attend the class with full motivation and passion.

Remember you are in the class to learn something new.



Course Policies – Pay attention

- Assignments must be submitted with in ***due dates***.
- Late submission will be subjected to the penalty which is as follow:

After ***2*** days of deadline ***30 %*** of marks deductions
After ***4*** days of deadline ***40 %*** of marks deductions
After ***5*** days of deadline ***100 %*** of marks deductions

- Student contact hours in my office are

Friday 10:00 AM - 1:00 PM

Above all:

- Maintain Discipline in the class. Not even in class overall in your personality.
- May be lecture is not interesting for you. But for someone who wants to learn something so let him / her to learn.
- We have to grow as a **Nation** not individual.

Communication rules and Contact hours

- For your queries related to the course, send an email to the following email address with clearly mention your Class and Student ID in the SUBJECT of an email.

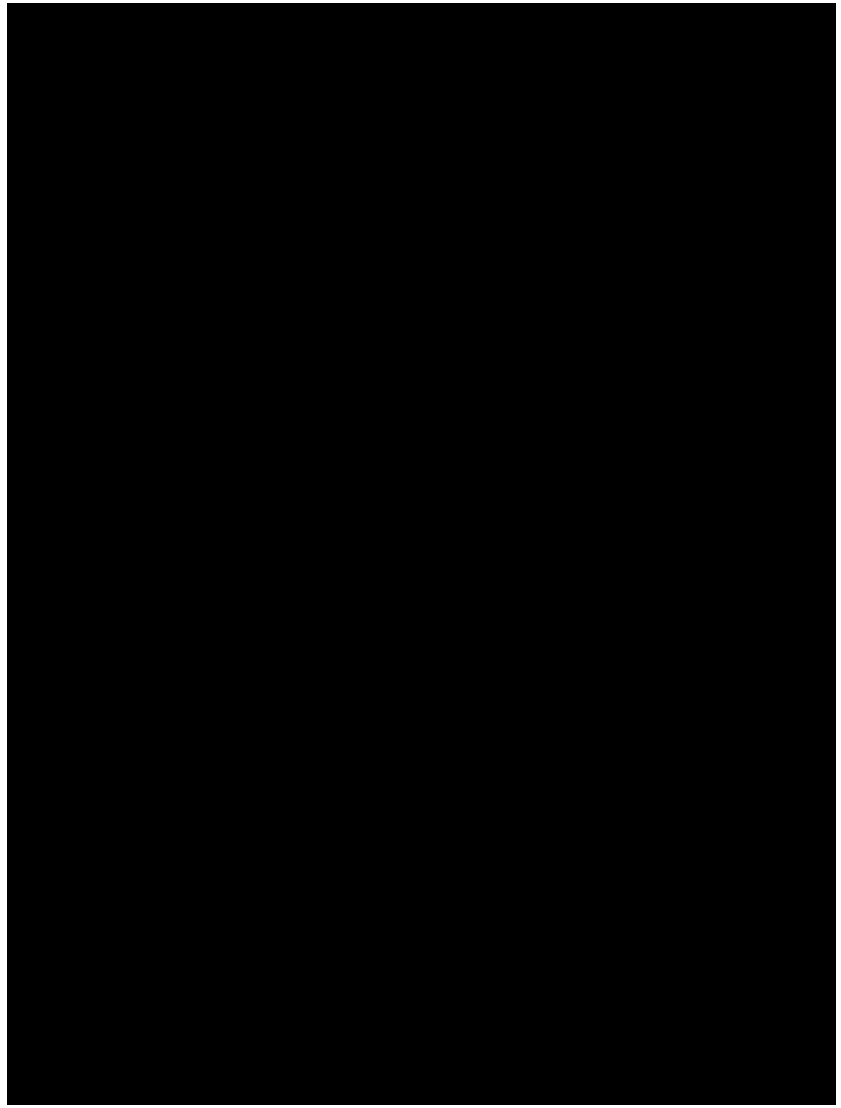
mfarrukh.shahid@nu.edu.pk

Or you can visit the office (in contacting hours) **Academic Block 3 Room no 06**

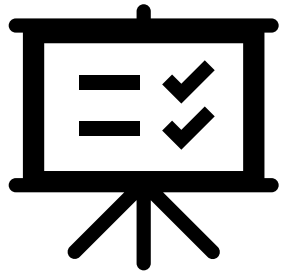
Or discuss in the class (this is highly subjected to the time availability in a class)

My Office Location

Room no 06
Academic Block 3



*The **Course** objectives are follows,*



Course Objectives

- How Computer Vision has been changing the world ?
- From theory to the practical
- Latest trends of CV in different applications
- Give you full insight to chose your Final Year Project.
- Developing International Collaboration
- Roadmap for Higher Education Abroad

Top CV conferences

Rank	Conference Details	Impact Score
1	**Computer Vision and Pattern Recognition** 18-06-2023 - 22-06-2023 - **Vancouver**	63.10
2	**Neural Information Processing Systems** 28-11-2022 - 09-12-2022 - **New Orleans**	42.30
3	**International Conference on Computer Vision** 11-10-2021 - 11-10-2021 - **Montreal**	40.60
4	**European Conference on Computer Vision** 24-10-2022 - 28-10-2022 - **Tel Aviv**	33.20
5	**International Conference on Machine Learning**	

Top CV research groups

1. **Arizona State University** - Center for Cognitive Ubiquitous Computing (CUbiC) |
2. **Arizona State University** - Image, Video, and Usability Lab
3. **Boston University** - Image and Video Computing Group
4. **Brown University** - Computer Vision
5. **California Institute of Technology** - Computational Vision
6. **Carnegie Mellon University** - CI2CV Computer Vision Lab
7. **Colorado State University** - Computer Vision
8. **Columbia University** - Computer Vision Laboratory |
9. **Cornell University** - Computer Vision Group

Find more on https://computervisiononline.com/blog/awesome-computer-vision-groups

Computer Vision Research Groups

- **Austria:**
  - Technical U of Graz Computer Vision Group.
  - Technical Univ. of Vienna Pattern Recognition and Image Processing Group.
  - Johannes Kepler Univ. Computer Vision Research Group.

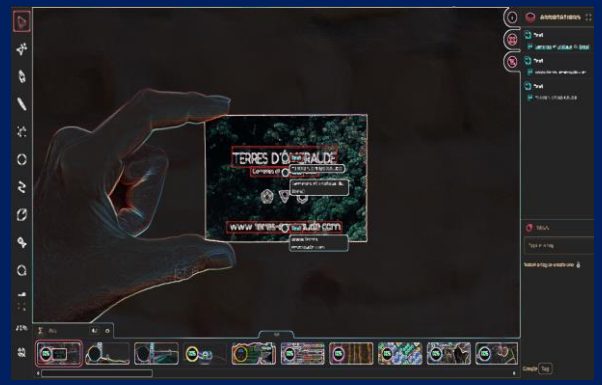
- **Australia:**
  - ANU Biorobotic Vision Group.
  - Univ. Auckland Tamaki Campus, Computer Vision Unit.
  - Univ. Curtin AI and Computer Vision Group.
  - Univ. Melbourne Computer Vision and Machine Intelligence Lab.
  - Univ. Western Australia Robotics and Vision Research Group.

- **Canada:**
  - DalTech Computer Vision and Image Processing Lab.
  - Laval Univ. Computer Vision and Systems Lab.
  - McGill Center for Intelligent Machines.
  - Simon Fraser Univ. Computational Vision Lab.
  - UBC Laboratory for Computational Intelligence.
  - Univ. Saskatchewan Computer Vision, Graphics and Image Processing Group

https://home.isr.uc.pt/~peixoto/PXResGrp.html

Top CV Projects

1. People counting tool
2. Colors detection
3. Object tracking in a video
4. Pedestrian detection
5. Hand gesture recognition
6. Human emotion recognition
7. Road lane detection
8. Business card scanner
9. License plate recognition
10. Handwritten digit recognition
11. Iris Flowers Classification
12. Family photo face detection
13. LEGO Brick Finder
14. PPE Detection
15. Face mask detection
16. Traffic light detection

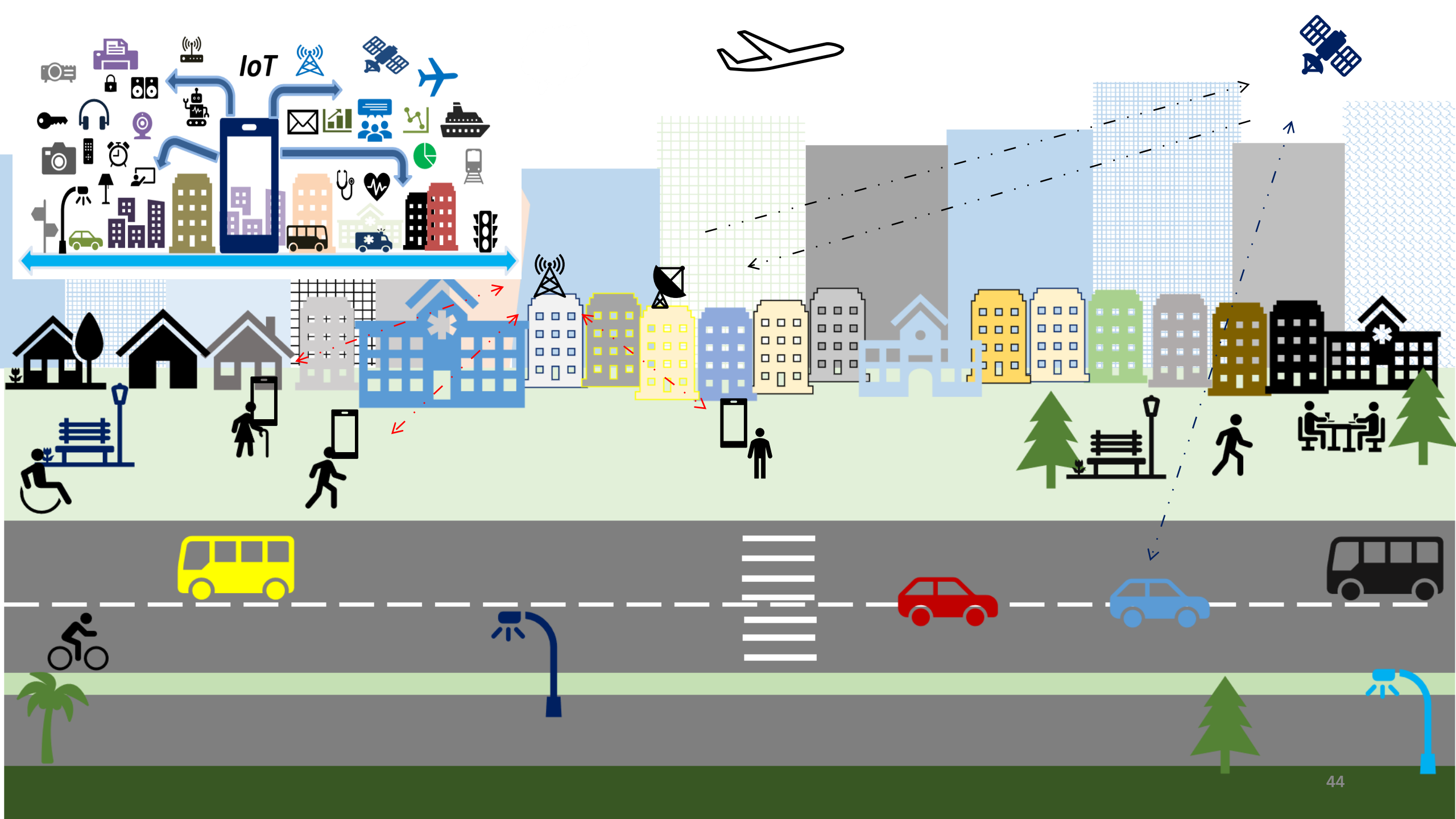


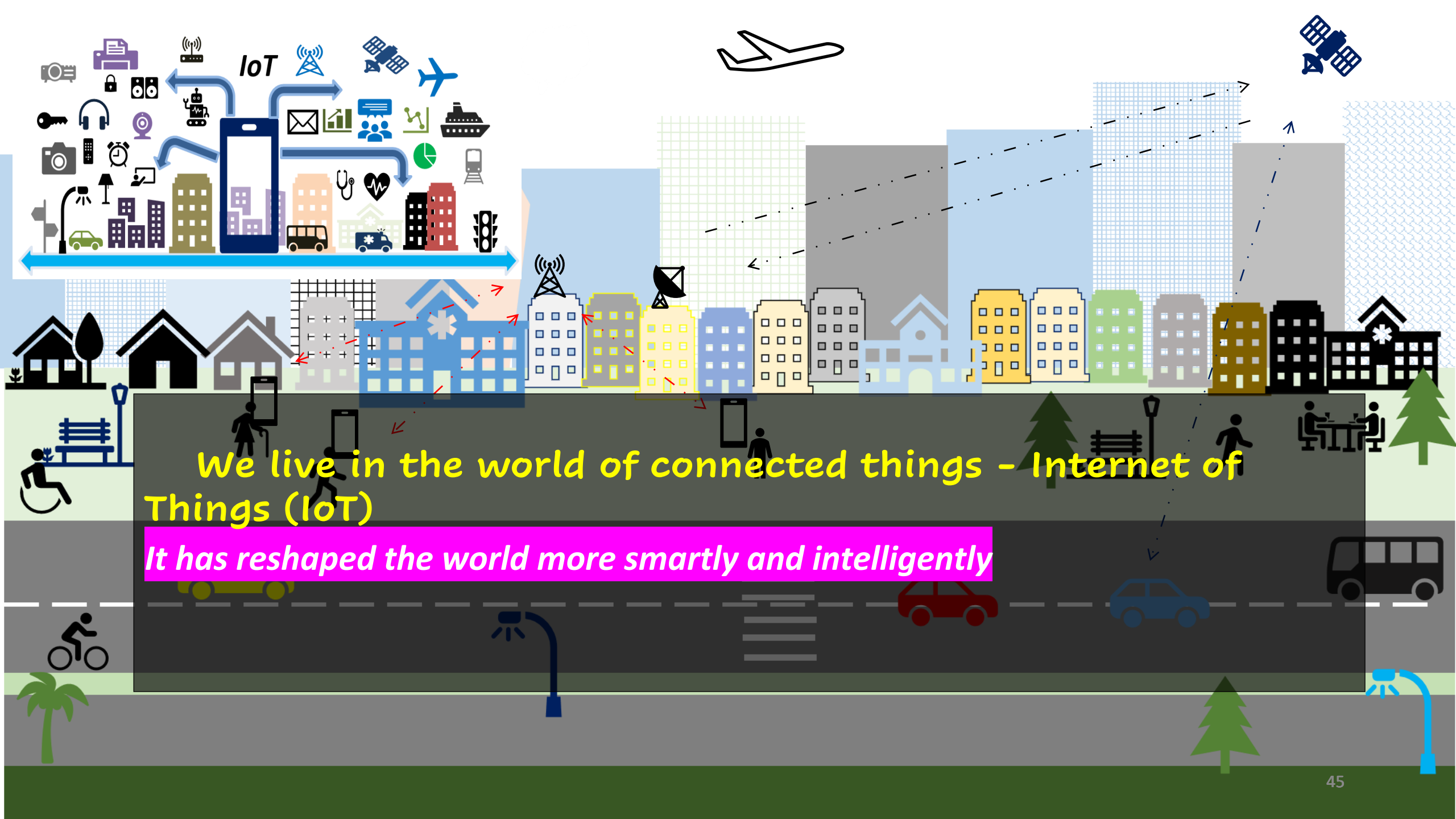
PPE Personal Protective Equipment

https://www.v7labs.com/blog/computer-vision-project-ideas#h13

Let’s explore


Objects everywhere


IoT



We live in the world of connected things - Internet of Things (IoT)

It has reshaped the world more smartly and intelligently

Internet of Things (IoT) is a network of various interconnected objects to provide services to the user [1].

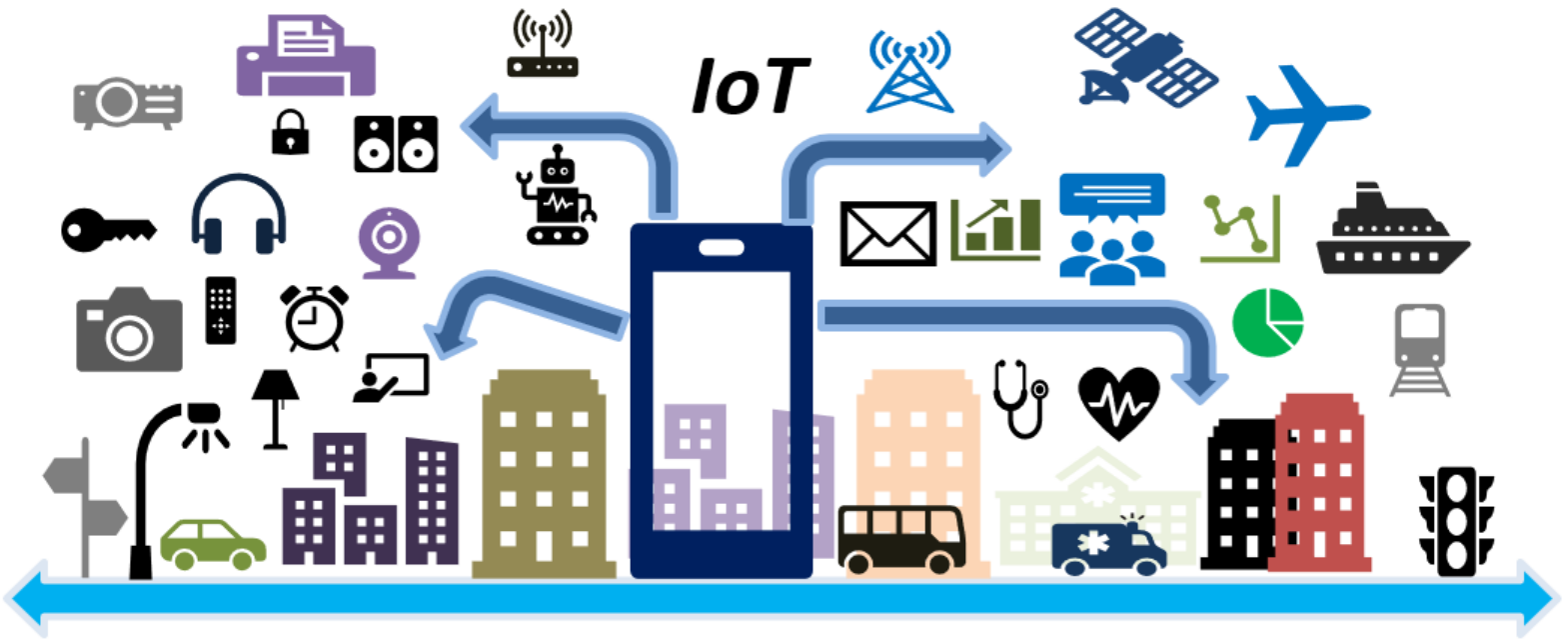
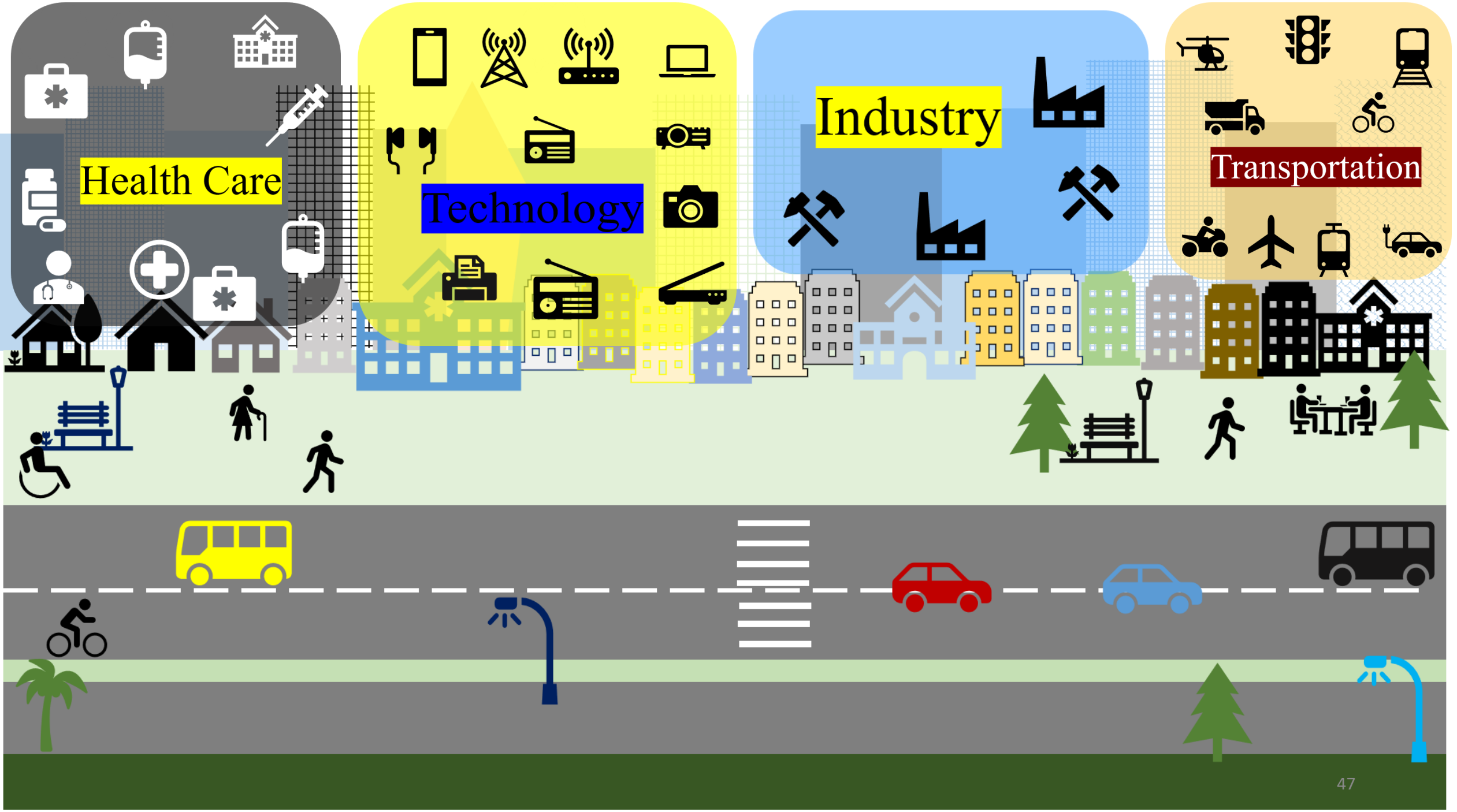
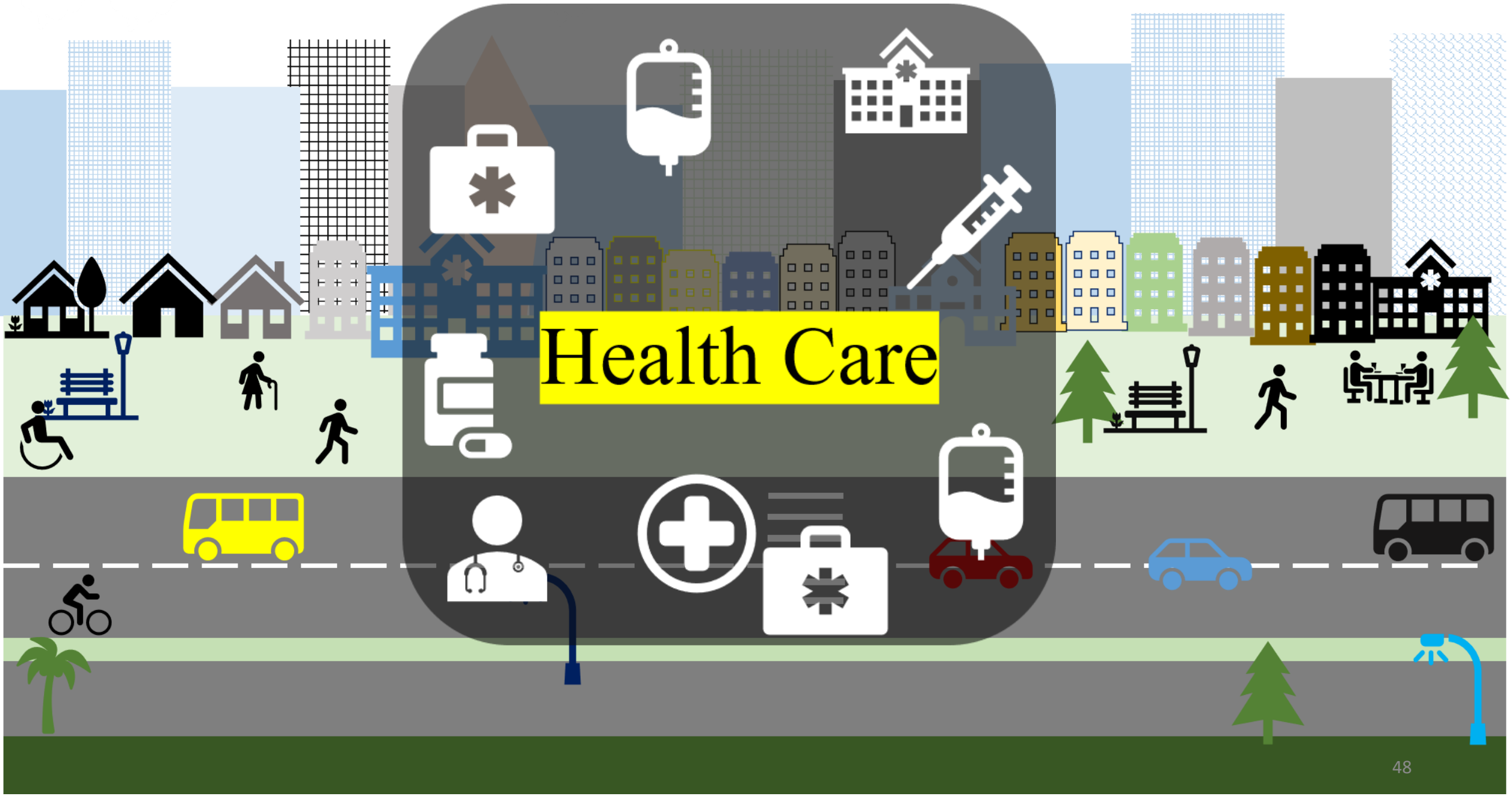
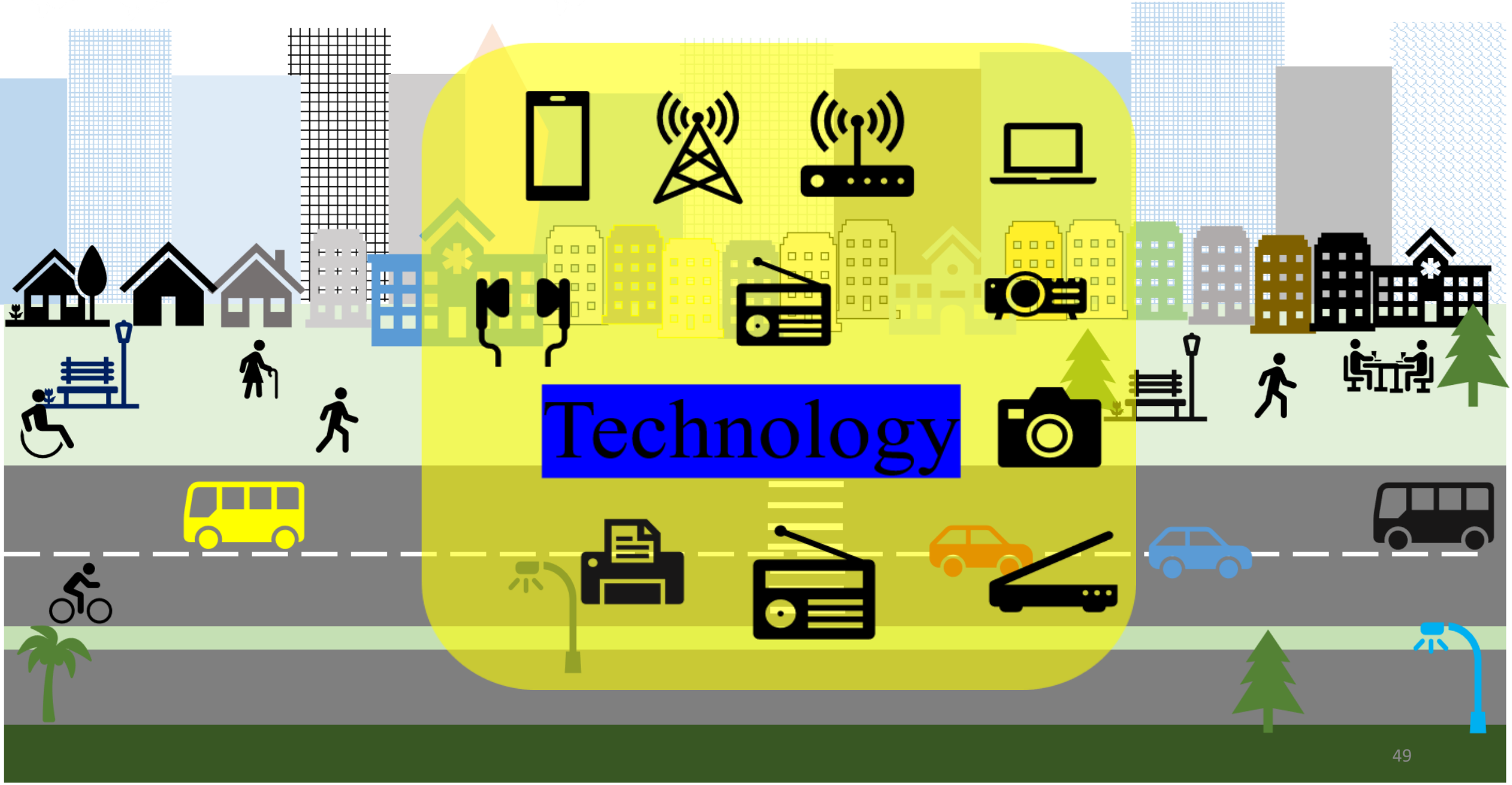


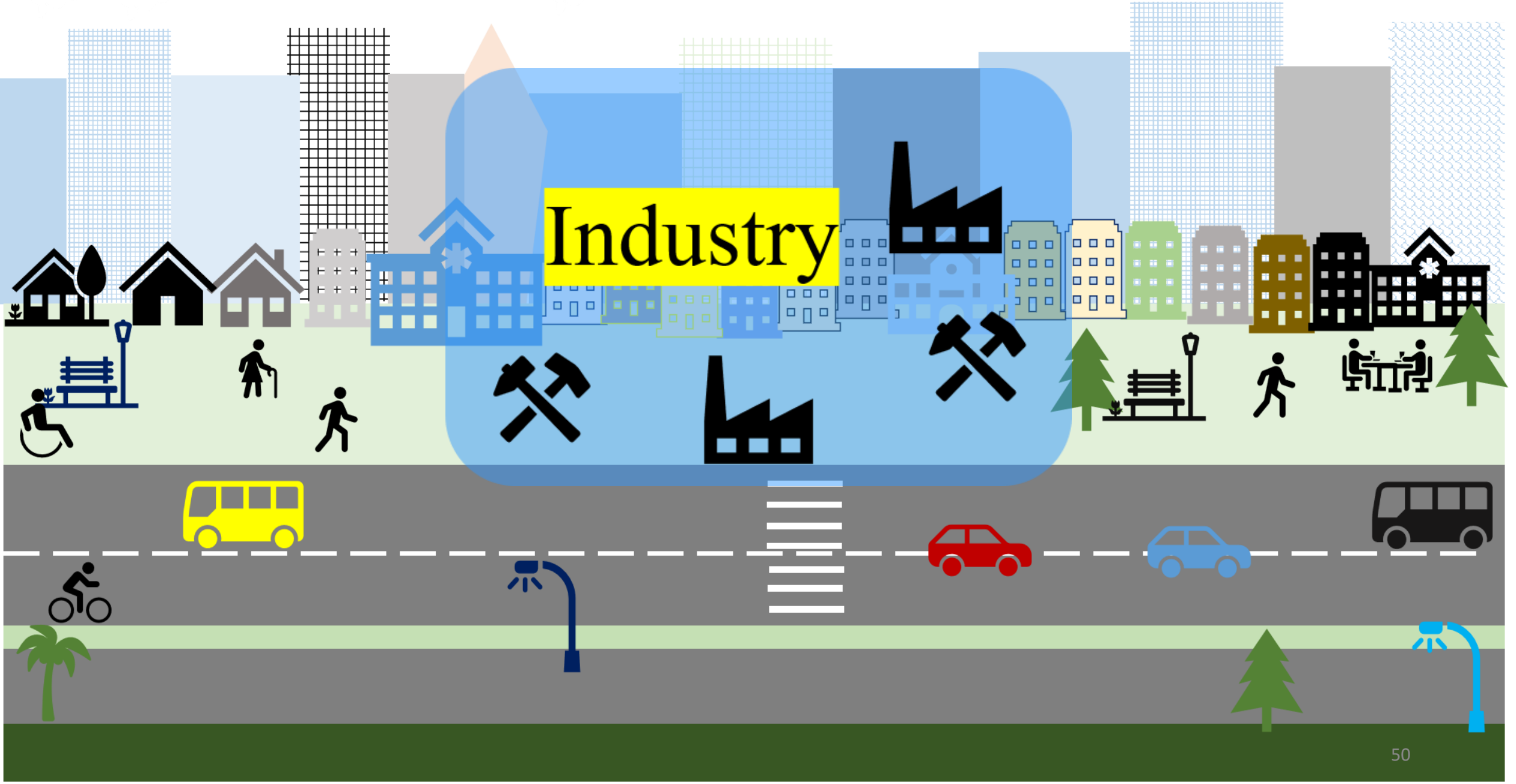
Figure 1 Internet of Things Network

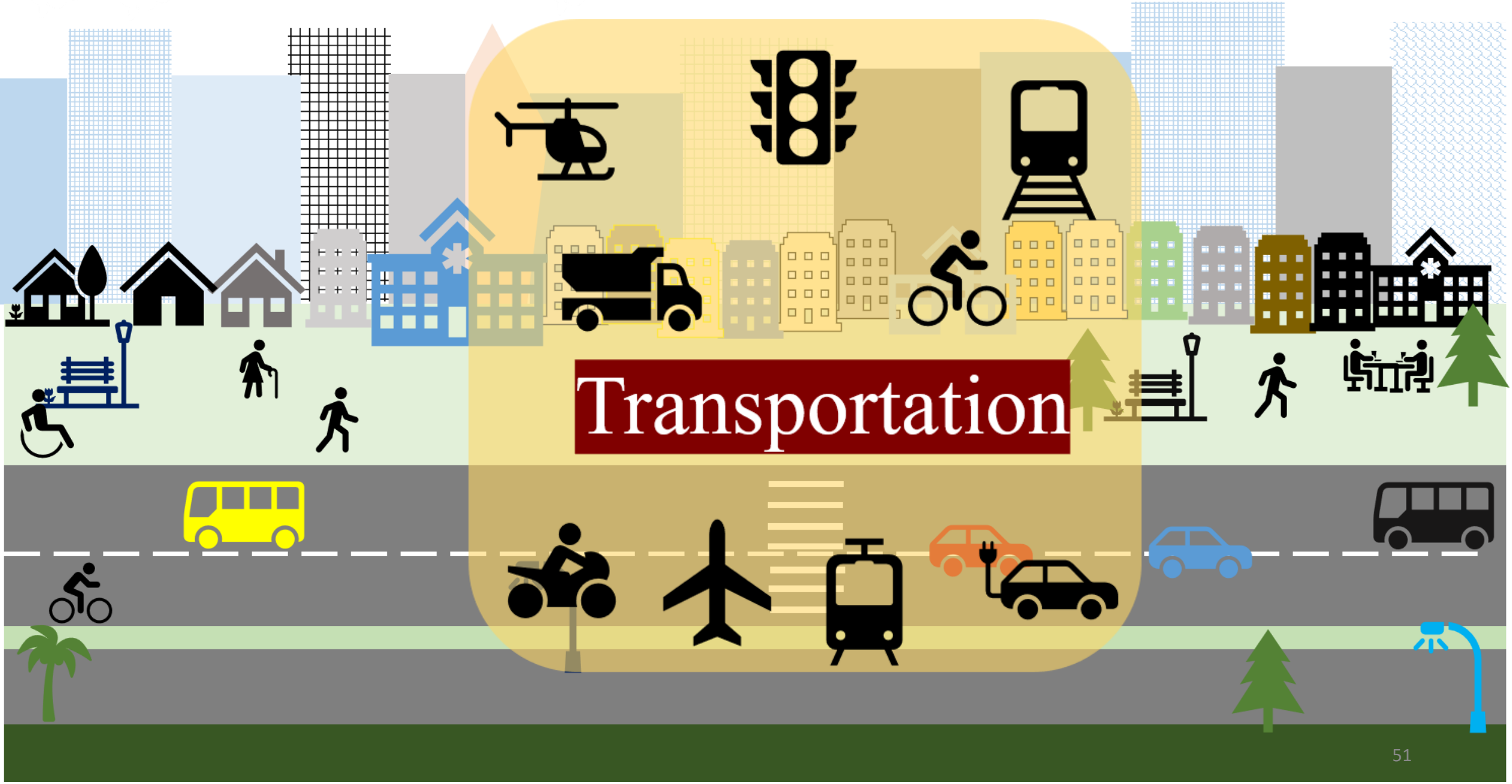
[1] Y. Saleem, N. Crespi, M. H. Rehmani, and R. Copeland, “Internet of things aided smart grid: Technologies, architectures, applications, prototypes, and future research directions,” IEEE Access, vol. 7, pp. 62962–63003, 2019

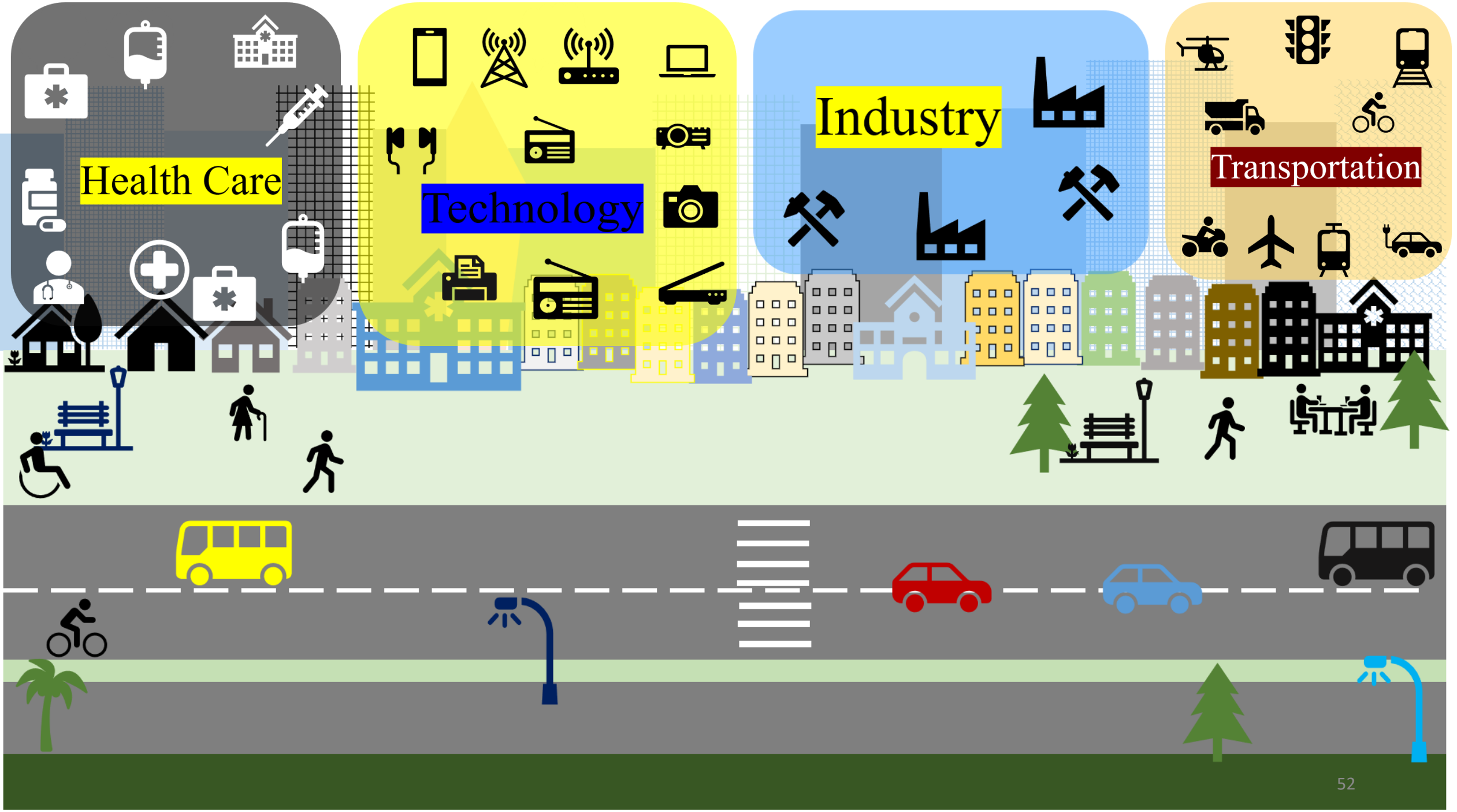

Health Care
Technology
Industry
Transportation

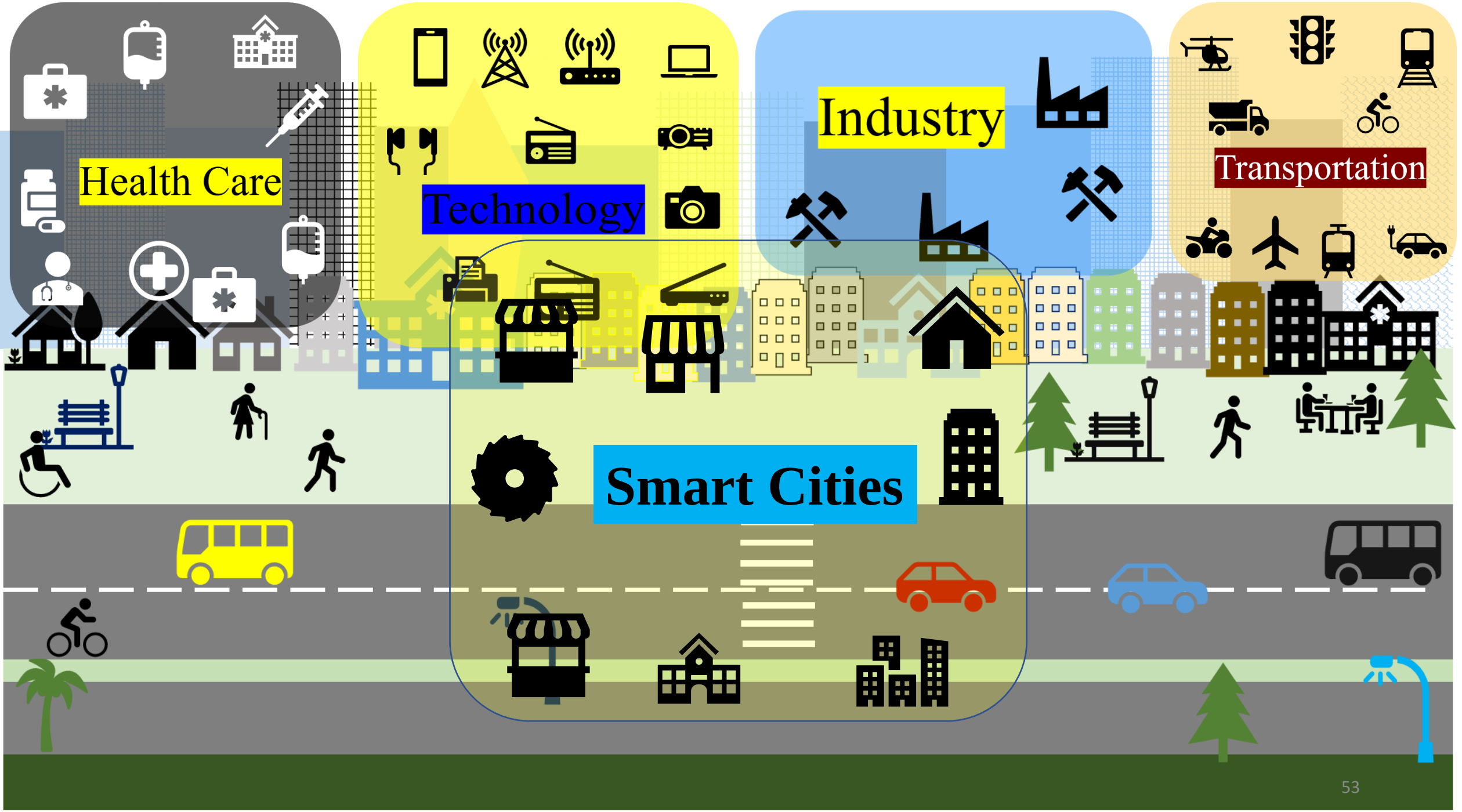

Health Care

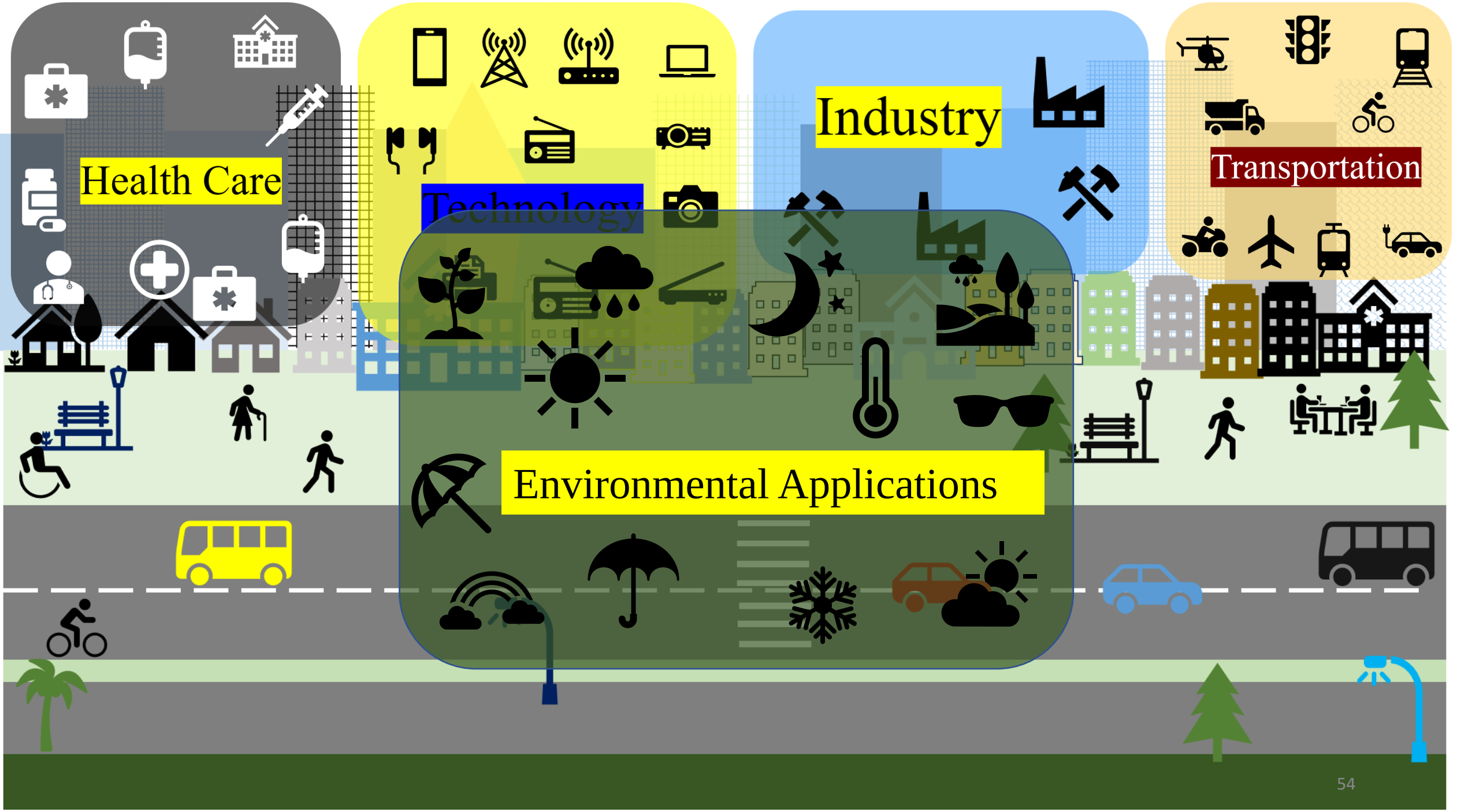

Technology


Industry


Transportation


Health Care
Technology
Industry
Transportation


Health Care
Technology
Industry
Transportation
Smart Cities


Health Care
Technology
Industry
Transportation
Environmental Applications

Connection in IoT Network

Wired Technology

Wireless Technology

Spectrum Scarcity problem in Wireless Network

Self-Awareness *(SA)* Capabilities [1]

Figure 2 CR-IoT Network

Vulnerable to attacks

[1] A. Toma *et al*., "AI-Based Abnormality Detection at the PHY-Layer of Cognitive Radio by Learning Generative Models," in *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 21-34, March 2020, doi: 10.1109/TCCN.2020.2970693.

Cognitive Radio (CR)

- Cognitive Radio (CR) was devised in 1991 by Mitola and Maguire to bestow the concept of intelligent radios capable of learning, reasoning, and acclimating to the environment [1].
- An essential feature of CR is the mastery of self-programming and autonomous learning.
- According to Haykin [2], CR radio meliorates spectrum utilization by using brain-empower devices to achieve efficient exploitation of spectrum and reliable communication objectives.
- In essence, it integrates model-based reasoning with software radio technologies to build adaptive, smart, and self-configured radios that learn the operating environment parameters and adapt changes accordingly.

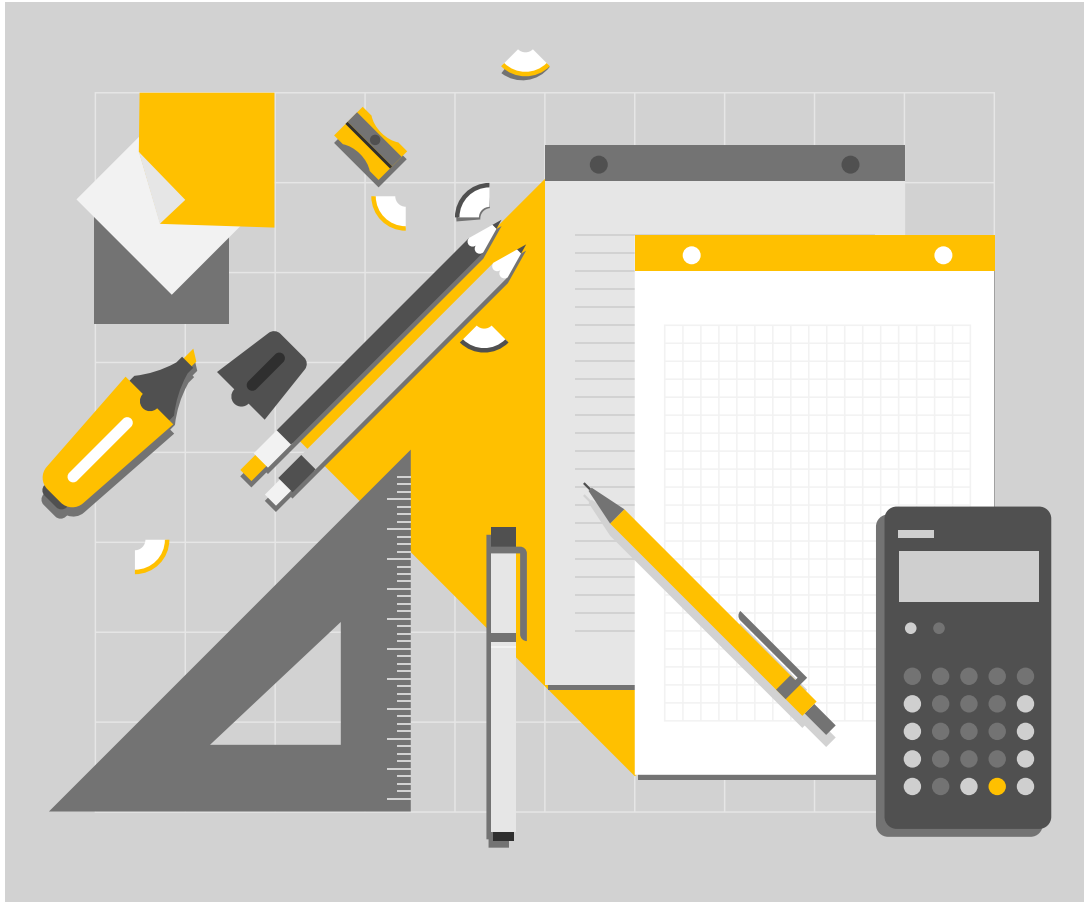
[1] J. Mitola and G. Q. Maguire, Cognitive radio: making software radios more personal,“ IEEE Personal Communications, vol. 6, no. 4, pp. 13{18}, 1999

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," in IEEE Journal on Selected Areas in Communications, vol. 23, no. 2, pp. 201-220, Feb. 2005, doi: 10.1109/JSAC.2004.839380.

Objectives

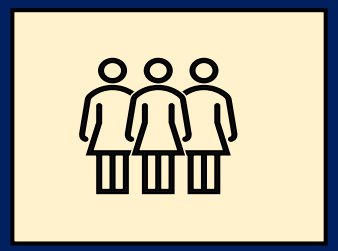
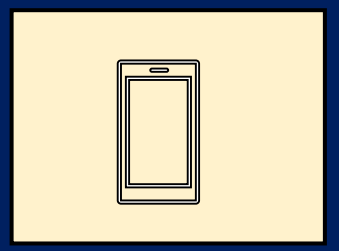
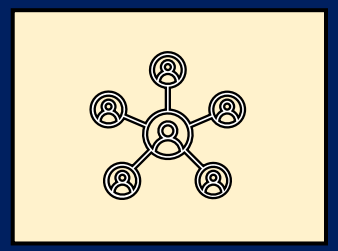
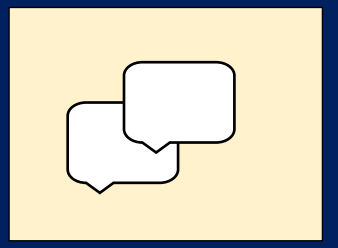
- Building CR-IoT network more protective and attack-free to conduct secure transmission.
- Making CR devices more intelligent, cognitive, and aware by proposing a method to bring SA capabilities into the CR-IoT objects.
- Providing a probabilistic framework to statistically exploit CR signals and model CR dynamic behavior inside a spectrum evolving with time.
- Capturing and detecting abnormalities at PHY-layer in CR-IoT network.
- Develop a data-driven approach that takes advantage of the deep learning method to handle high dimensional radio signals and dynamic Bayesian model to provide spectrum inference at continuous and discrete levels and perform state estimation tasks. In addition to that, the proposed method should be capable of capturing abnormalities in the CR-IoT network spectrum.

180º Perception

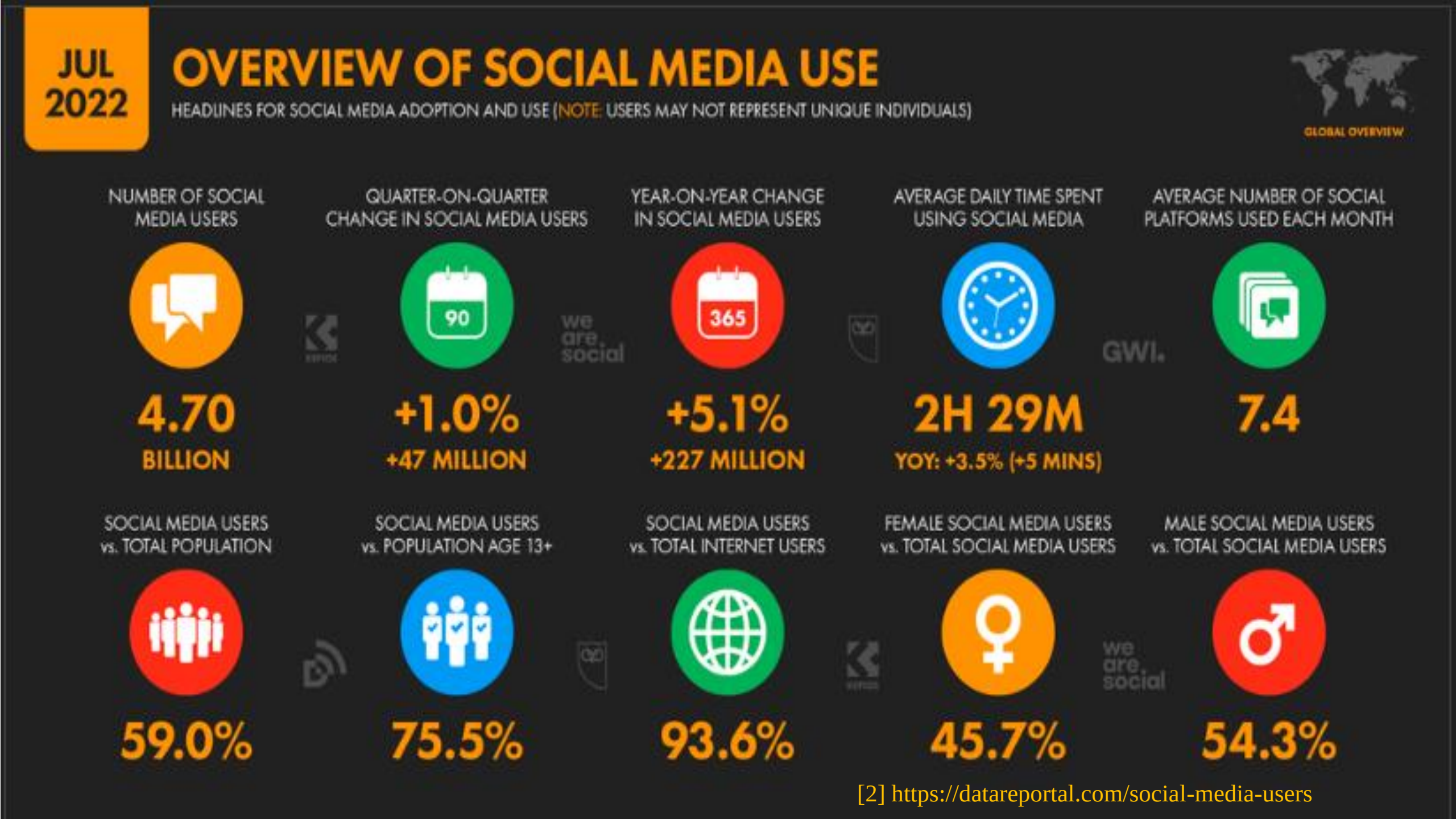


Digital adaptations around the world in 2022

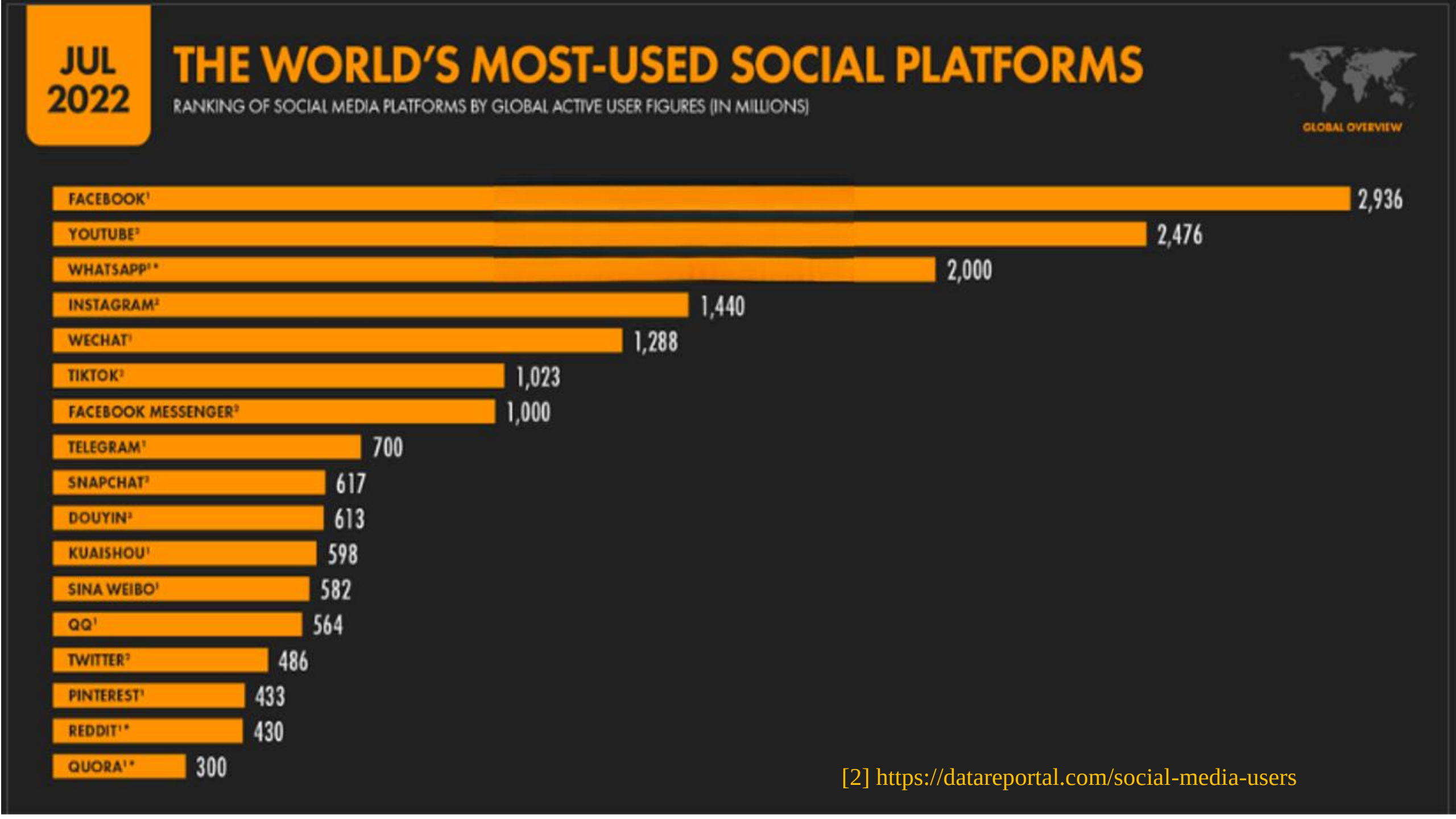
Digital around the world[2]

Total Population	*Unique mobile phone users*	*Internet users*	*Active Social media users*
			
8 Billion 56.6 % *Urbanisation*	5.27 Billion 65.6 % *Population*	4.72 Billion 59 % *Population*	4.25 Billion 53.1 % *Population*

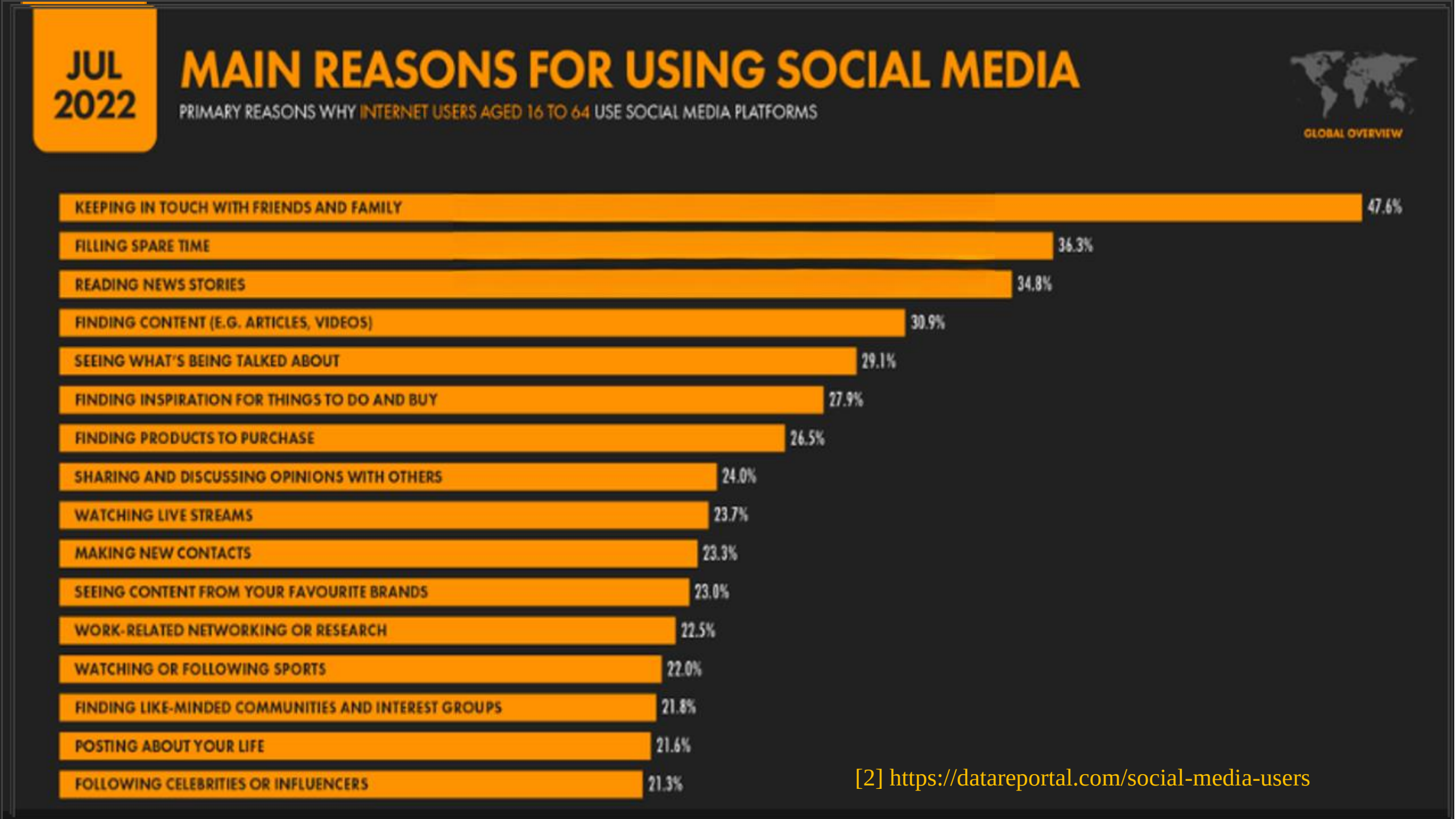
[2] https://datareportal.com/social-media-users



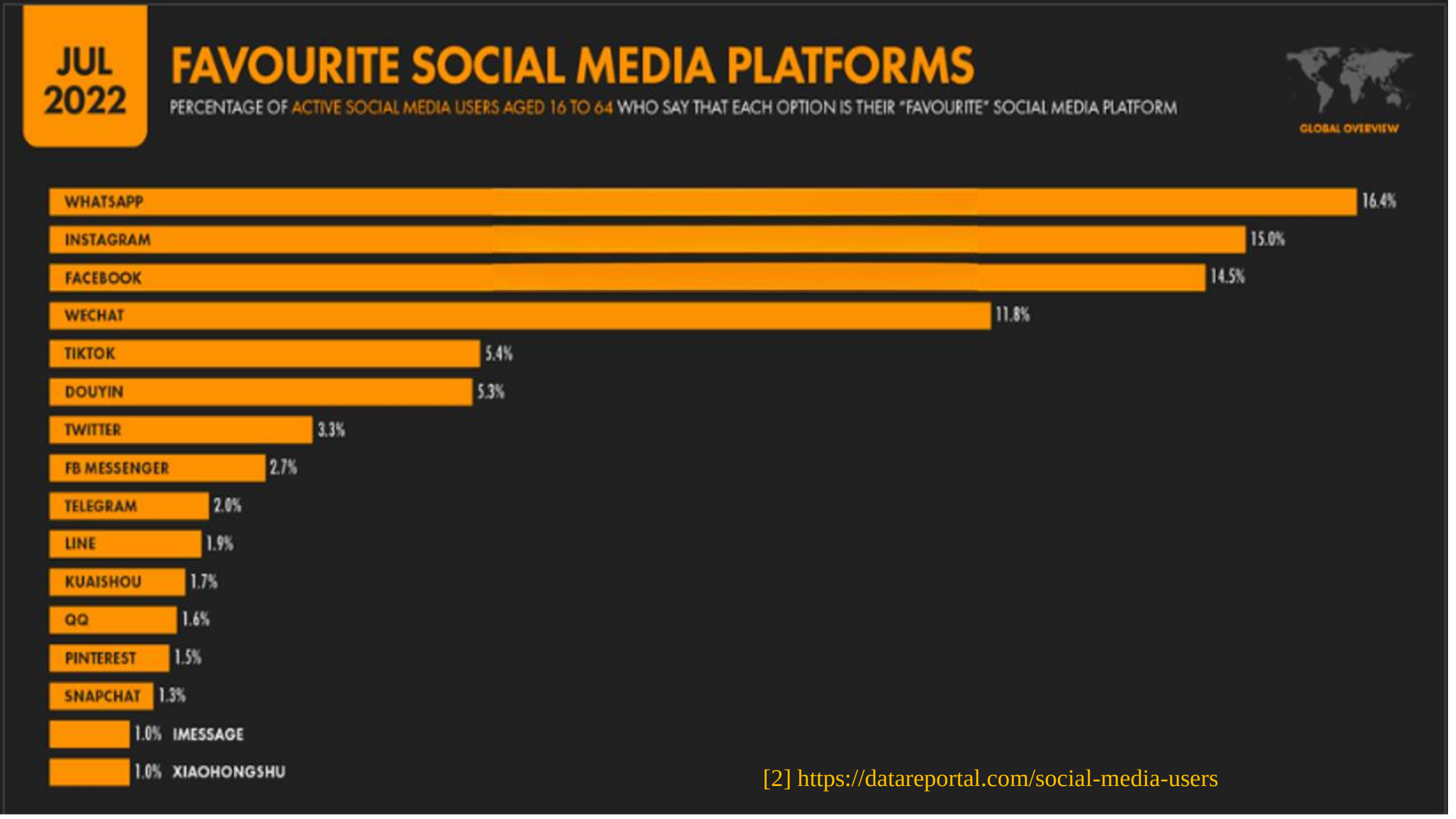
[2] https://datareportal.com/social-media-users



[2] https://datareportal.com/social-media-users



[2] https://datareportal.com/social-media-users



[2] https://datareportal.com/social-media-users

Consequences

Everyday we generate approximately the following amount of data

2.5 quintillion bytes a day
(175 trillion gigabytes of data)

Hence,

How to handle such huge amount of data ?

Data Science (DS) – An emerging paradigm

What is Data Science (DS) ?

Data Science is a data-driven approach which extract information from the given data and make inferences on a data by using scientific tools and techniques.

It is process to discover facts and figures in a given data.

Why do we need DS ?

D

Discover pattern in a data

Make predictions

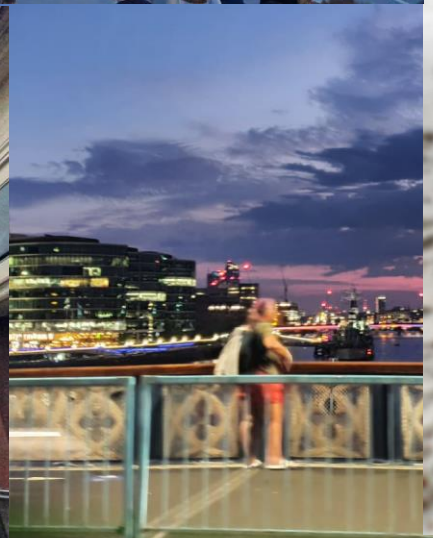
S

Enhance decision making

Data in Data Science

- Audio Signals
- Images
- Videos


VEAR A FACE COVERING ON PUBLIC TRANSPORT
be turned away from the network and could be fined



Computer Vision

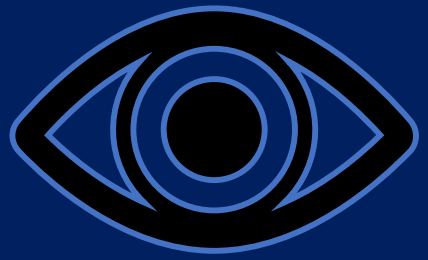
Computer vision tasks seek to enable computer system automatically to see, identify and understand the visual world, simulating the same way that human vision does [1].

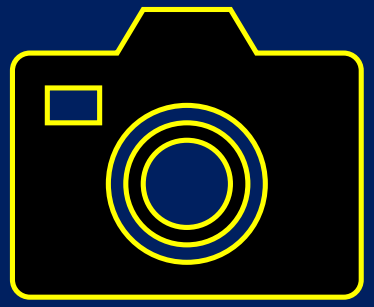
Researchers in computer vision have been developing, in parallel, mathematical techniques for recovering the three-dimensional shape and appearance of objects in imagery.

[1] Computer vision, WikiPedia. [Online]. Available: https://en.wikipedia.org/w iki/Computer_vision

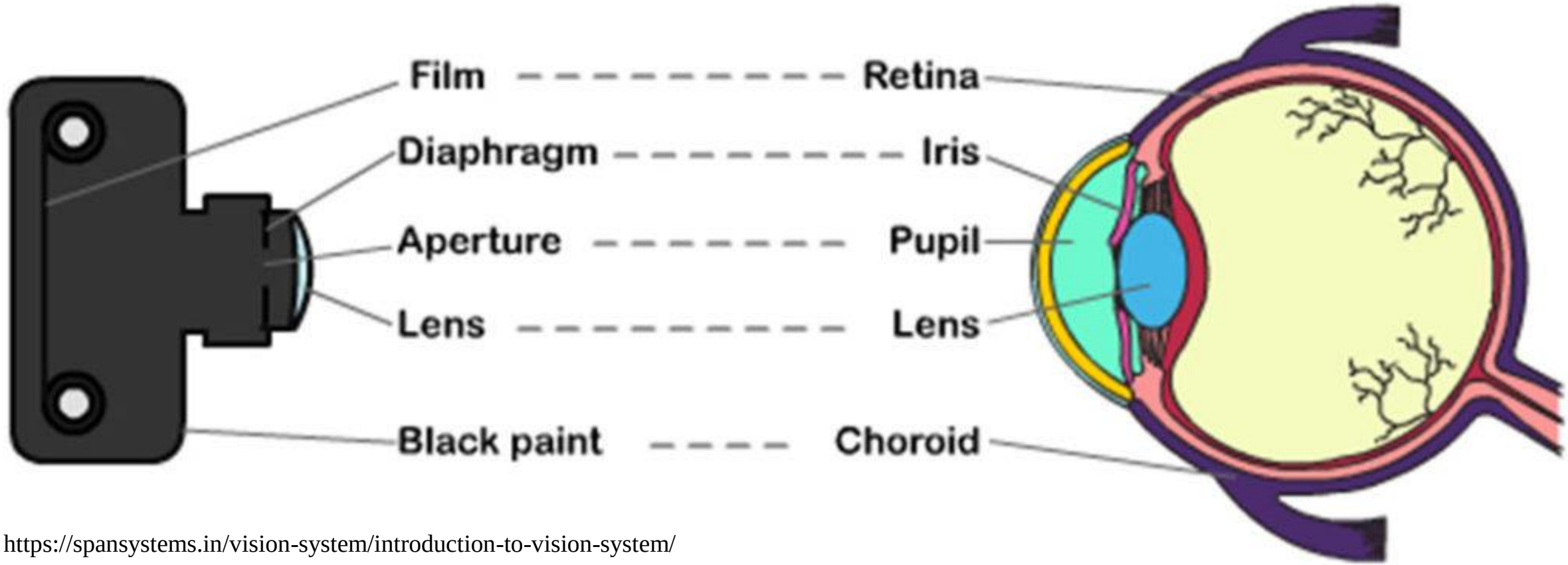
Human vision has no meaning without a beautiful brain

Similarly



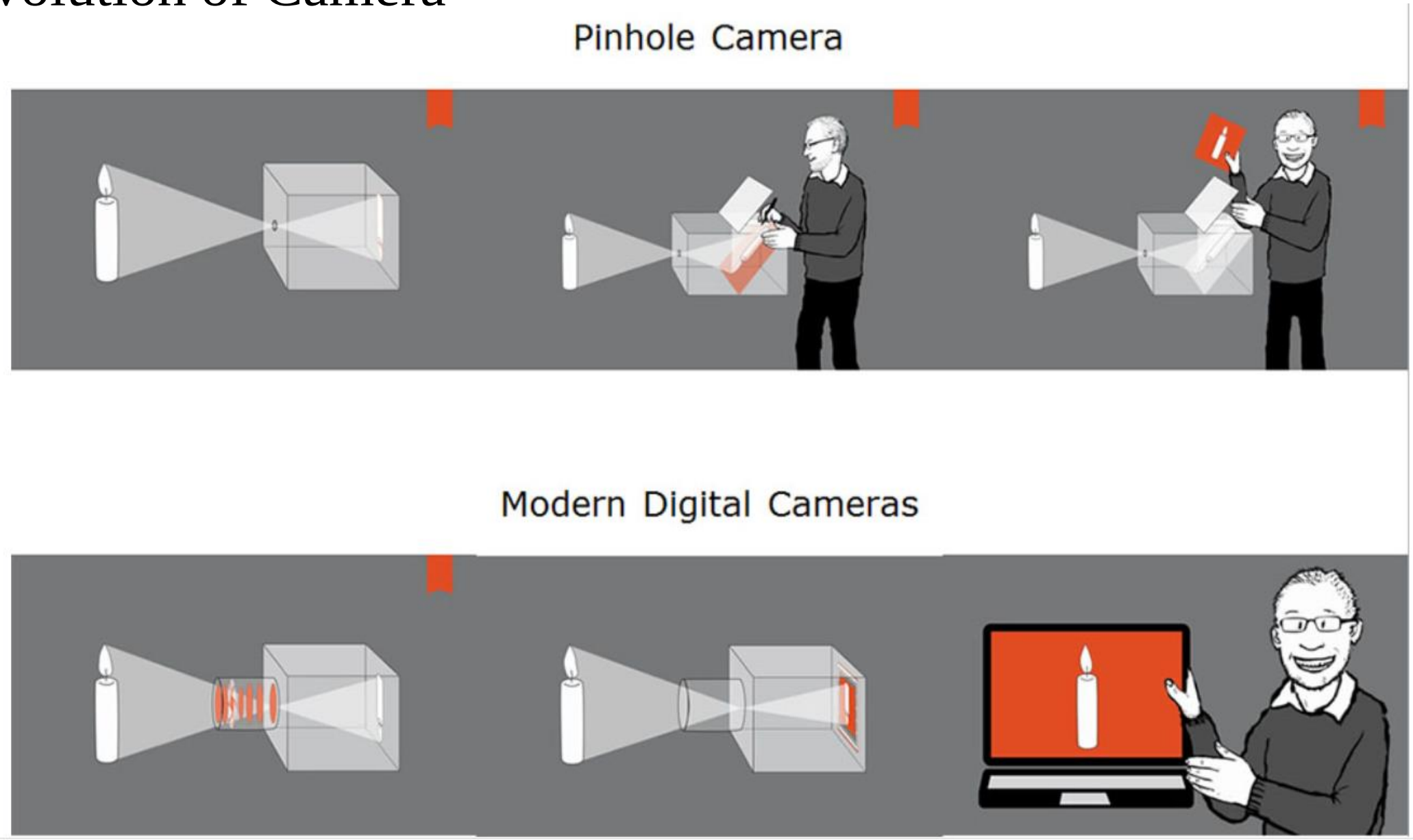


Images from camera has no value without a powerful software

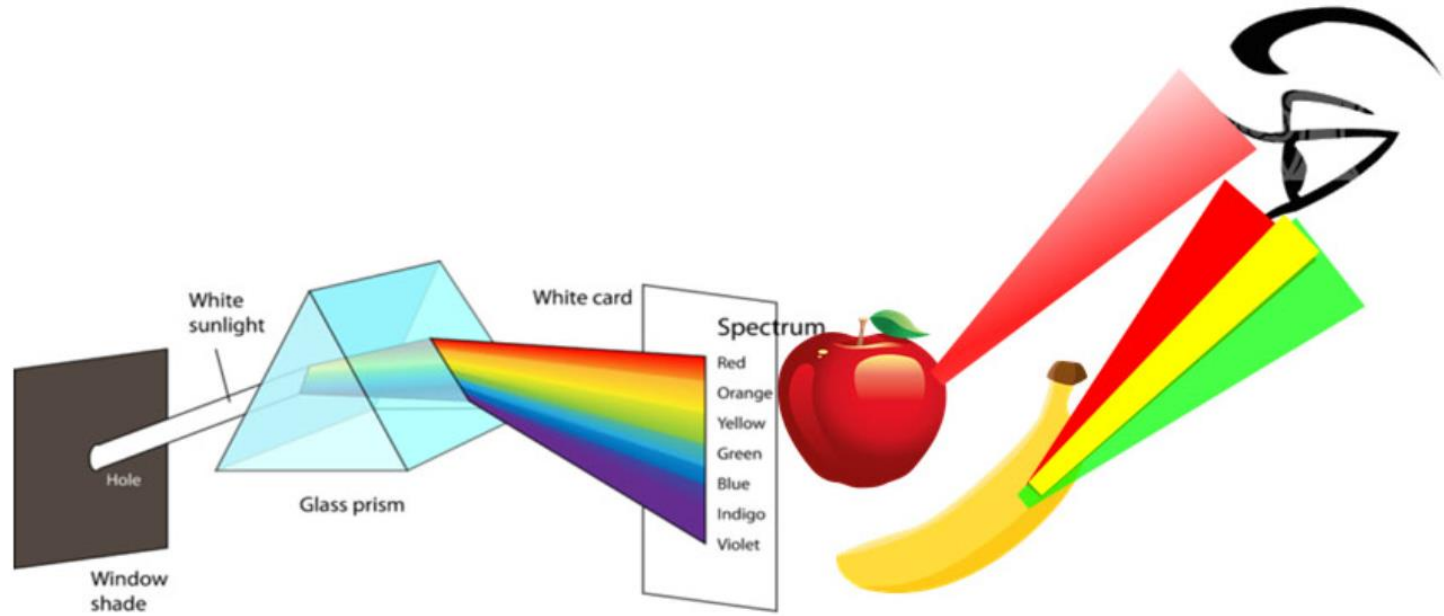


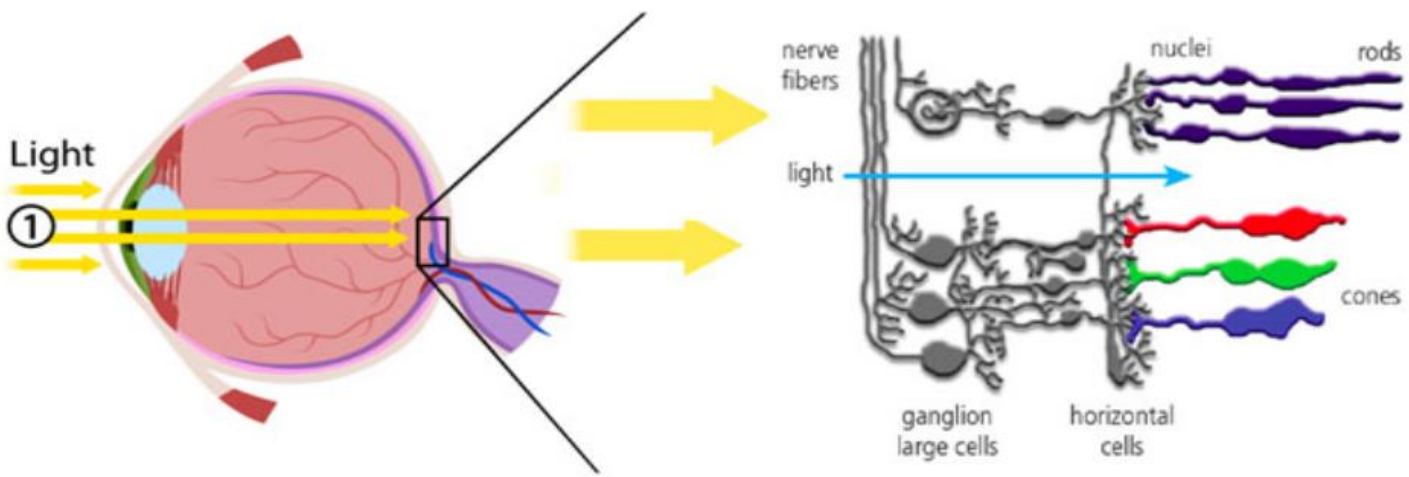
https://spansystems.in/vision-system/introduction-to-vision-system/

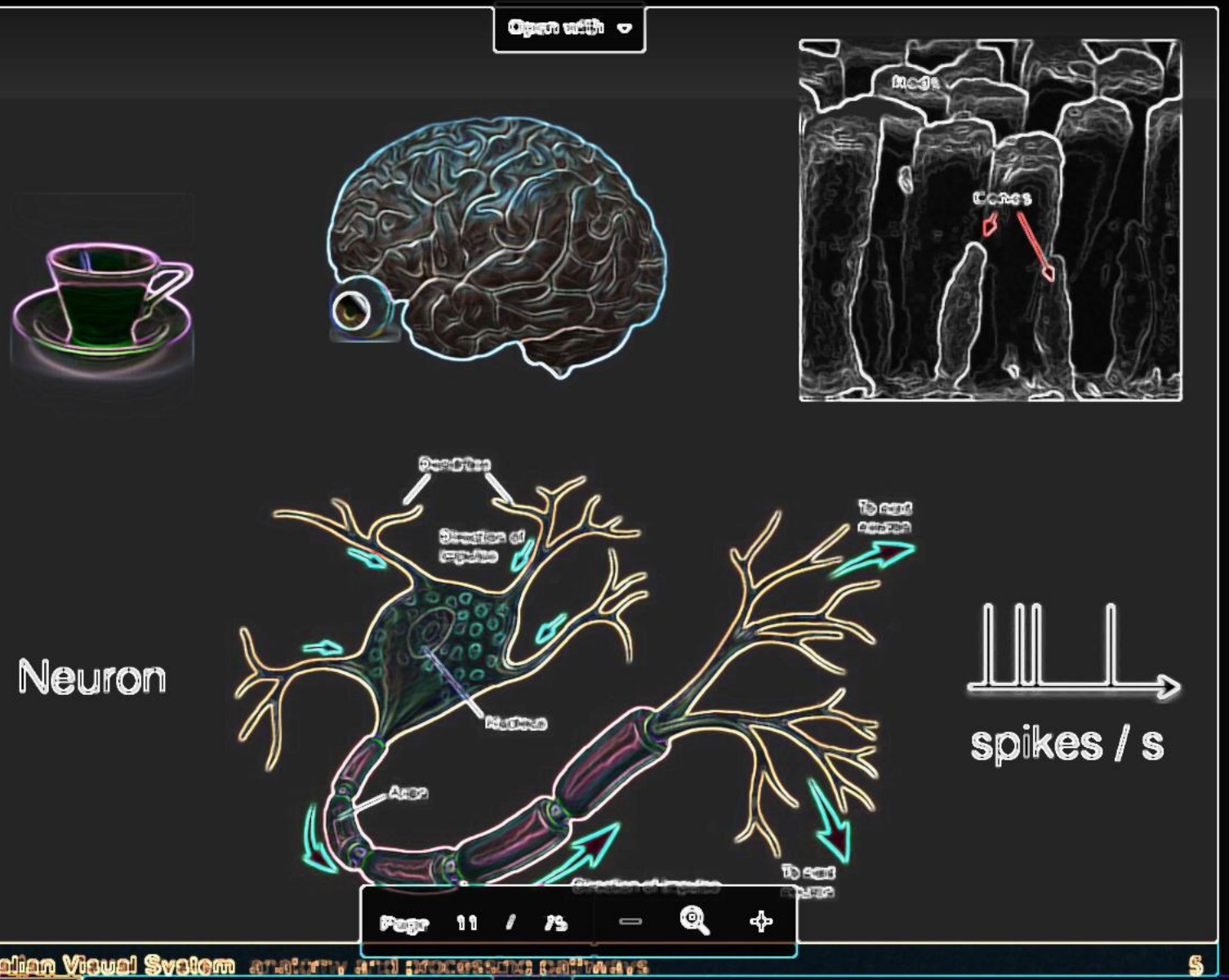
Evolution of Camera

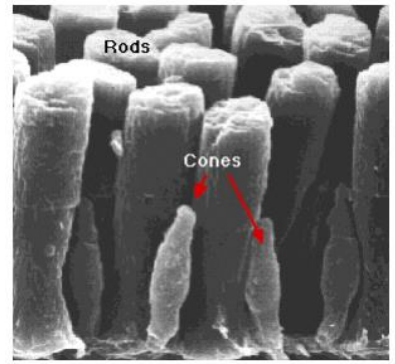


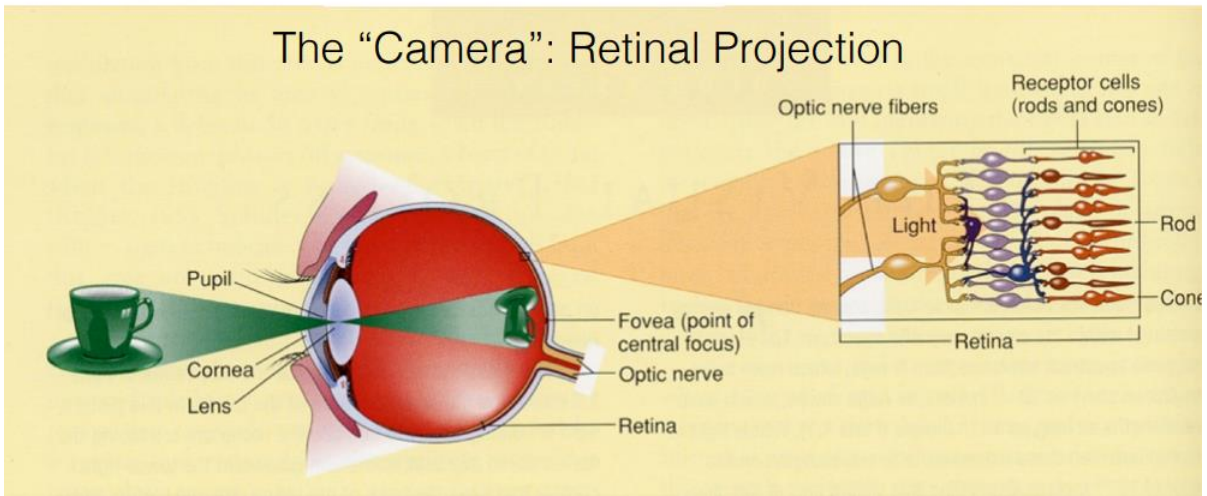
How are color seen to eyes?

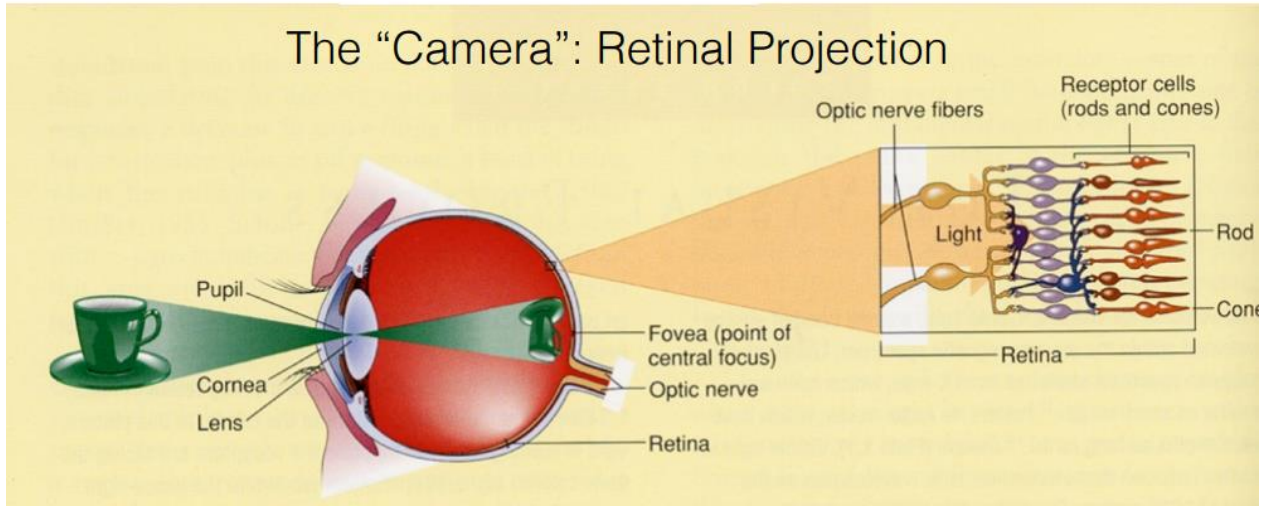





Neuron
spikes / s
Mammalian Visual System



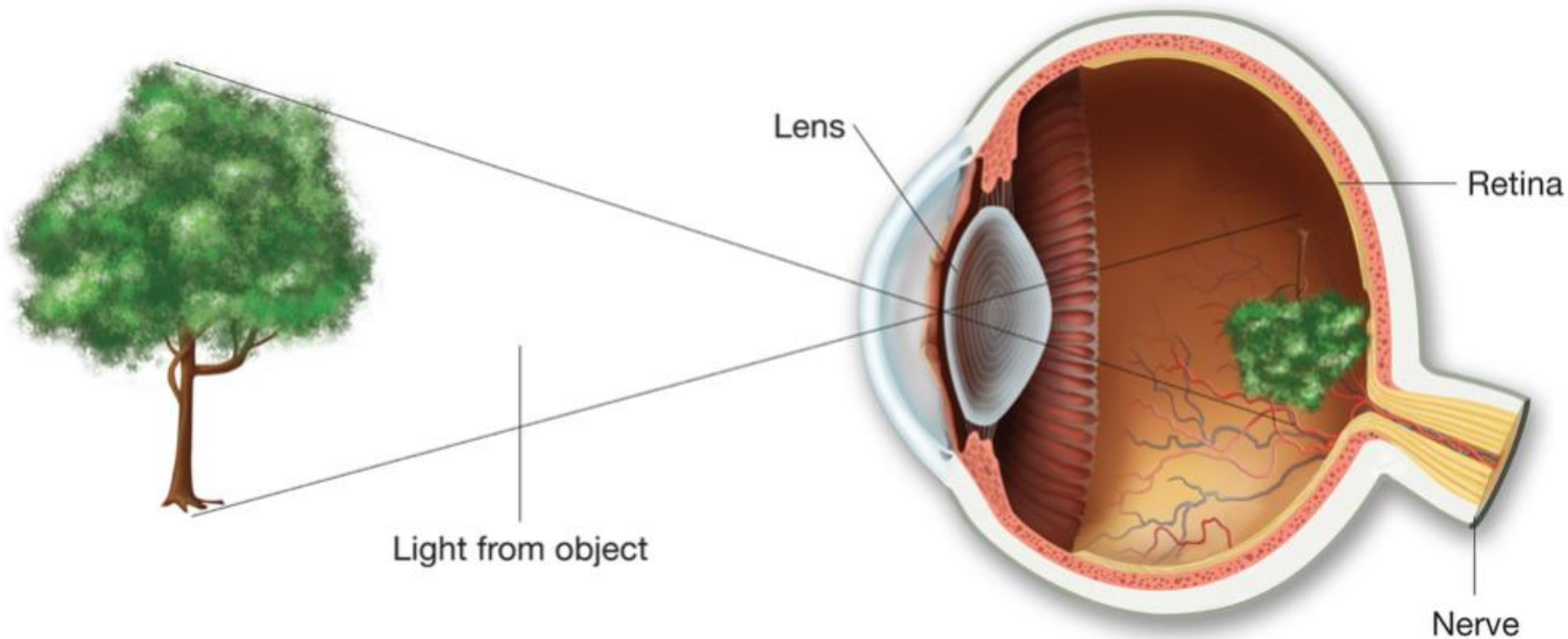




http://vision.stanford.edu/teaching/cs131_fall1415/lectures/lec-10-27_v2.pdf

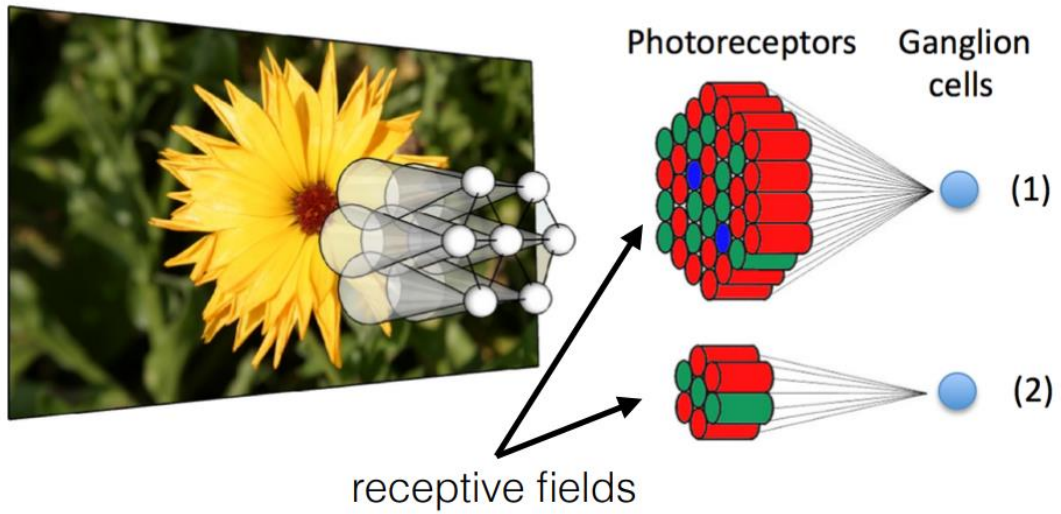
Retinal Projection

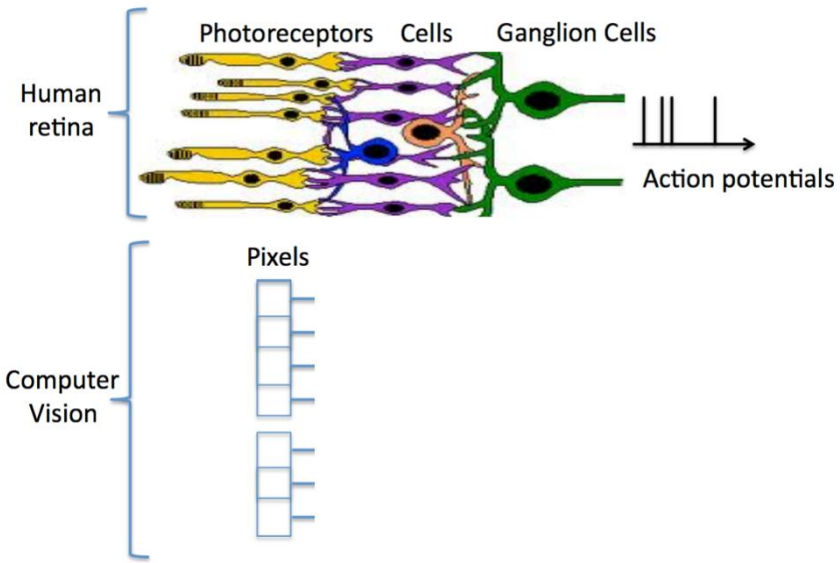
picture is inverted, but spatial relationships are preserved



Receptive Fields

extract similar features at each position in the visual field



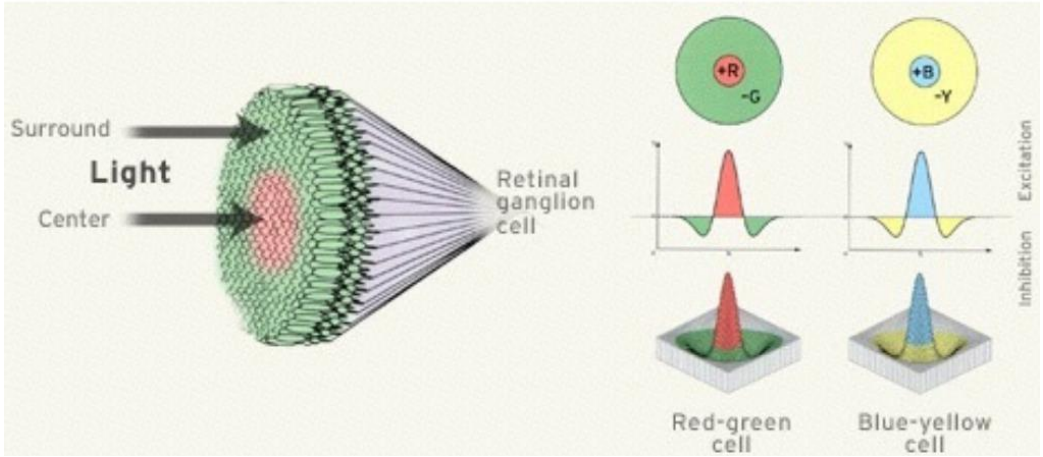


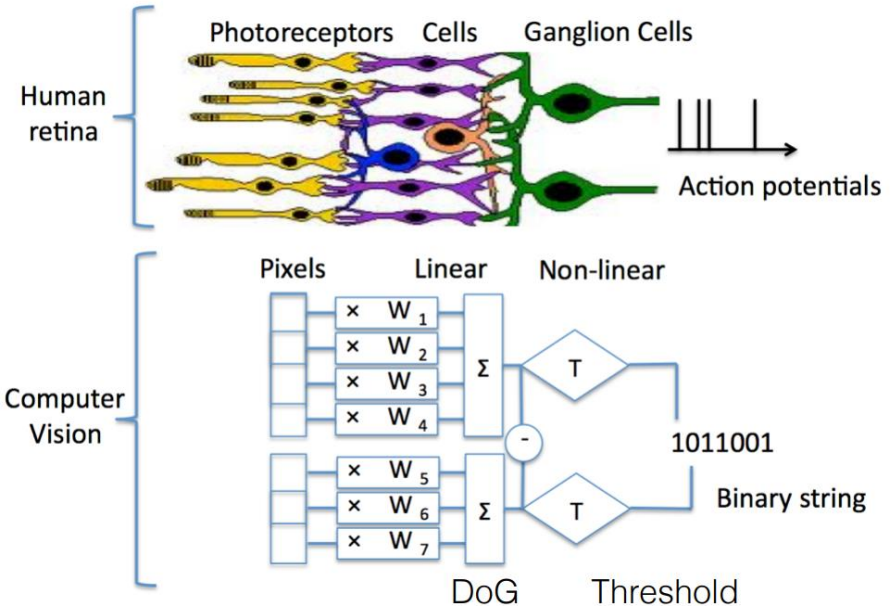
http://vision.stanford.edu/teaching/cs131_fall1415/lectures/lec-10-27_v2.pdf

Receptive Fields

extract similar features at each position in the visual field

ON ganglion cells

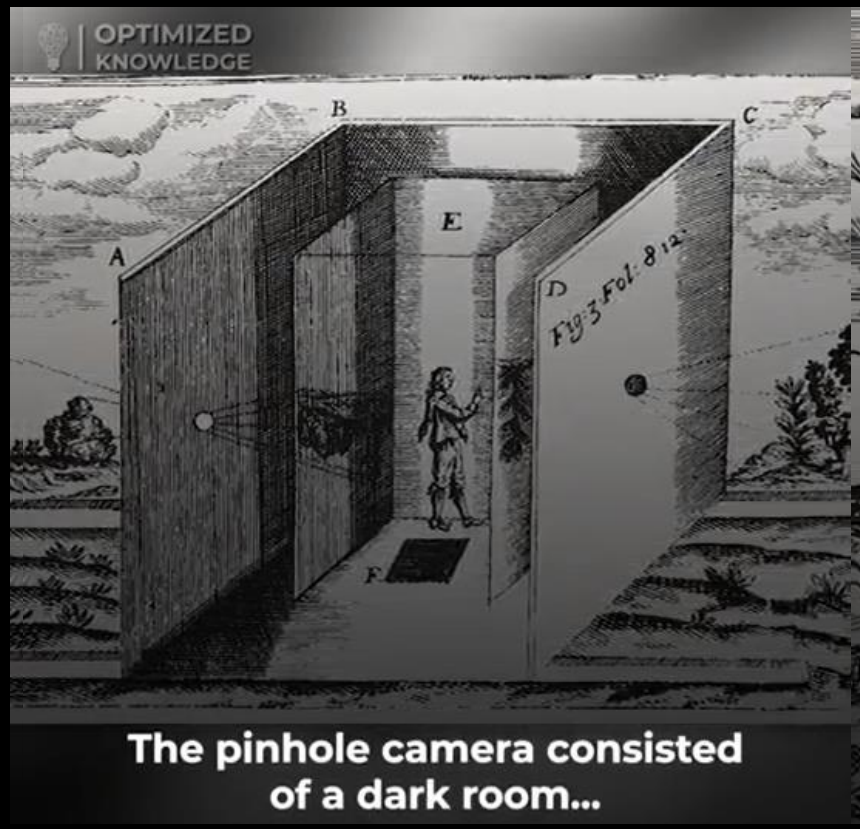




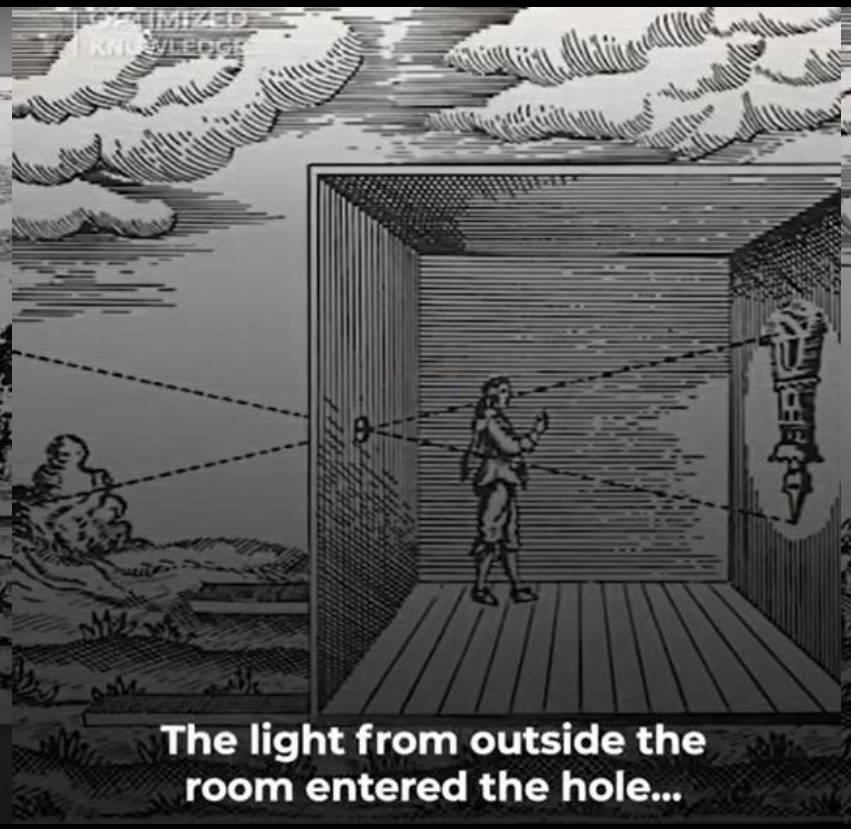
Ibn Al Haytham invented a working model of camera.

Which he called the Albait Almuzlim meaning “the dark room.”

But how did this first camera work?



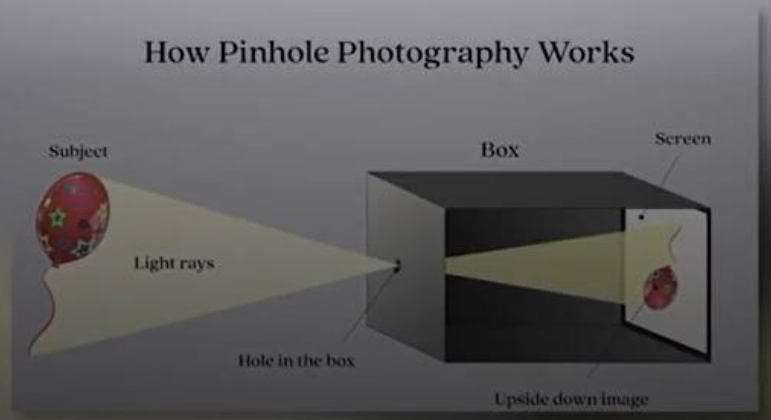
The pinhole camera consisted of a dark room...



The light from outside the room entered the hole...

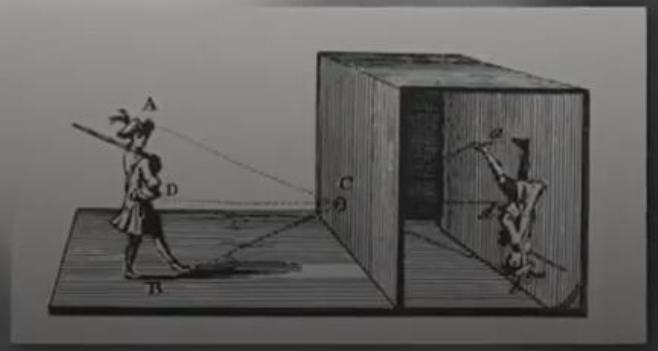
And projected a luminous beam onto the opposing wall.

OPTIMIZED KNOWLEDGE



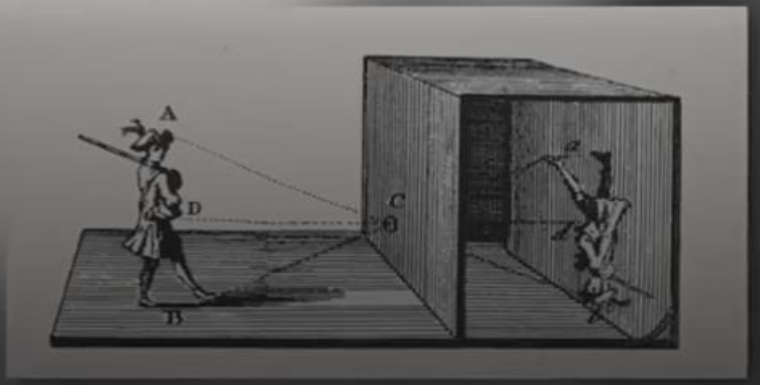
The entered light projected the inverted image of the actual object inside the camera.

OPTIMIZED KNOWLEDGE



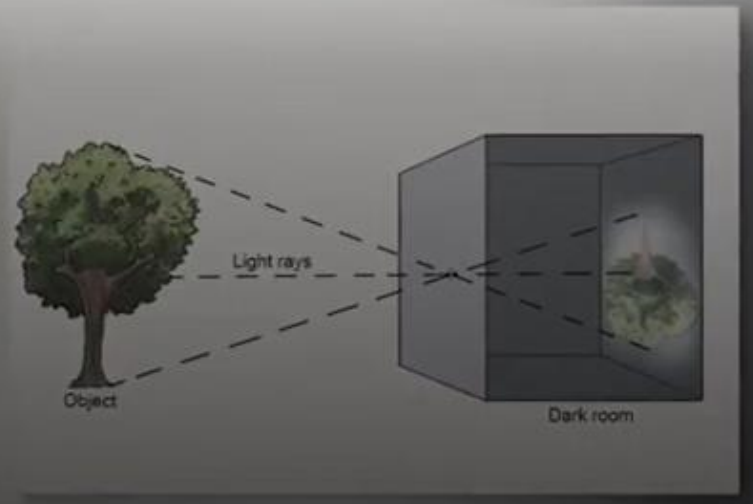
The smaller the hole, the sharper the image appeared.

OPTIMIZED KNOWLEDGE



However, when the hole was too small, the projected image lacked brightness.

OPTIMIZED KNOWLEDGE



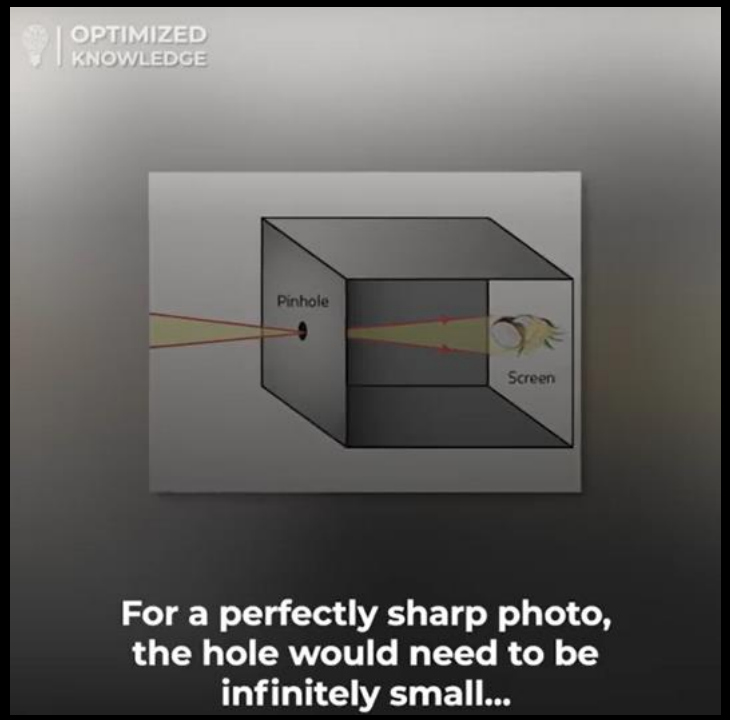
Hence, there was an optimal hole size that gave enough definition and brightness to the image.

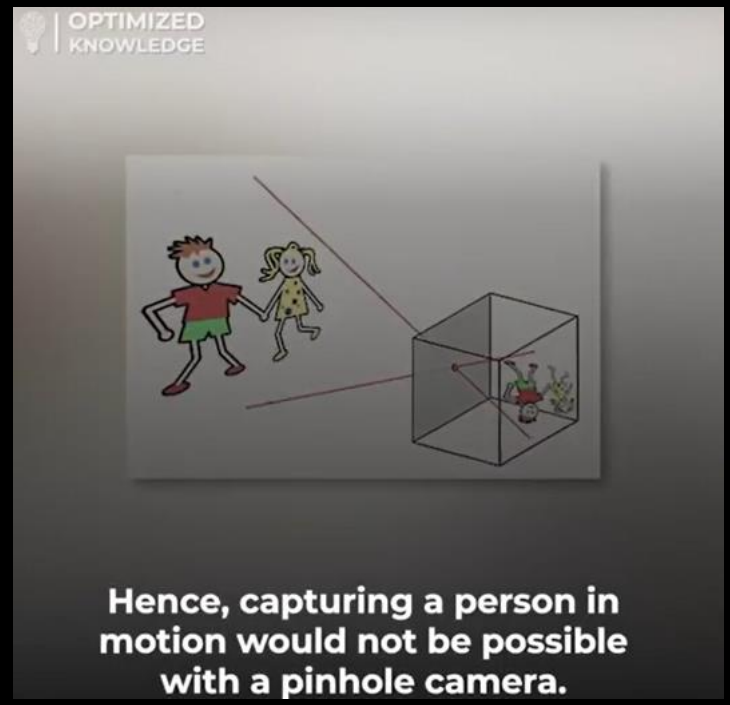
OPTIMIZED KNOWLEDGE

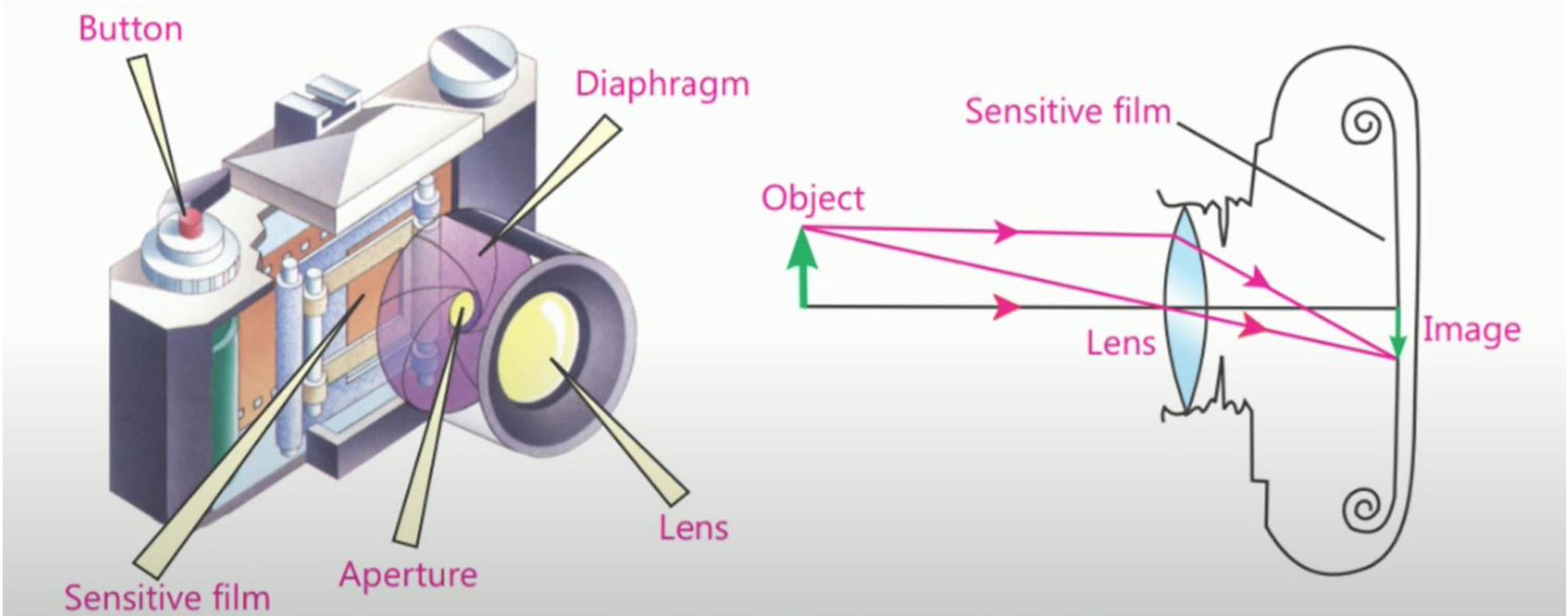
Ibn Al-Haytham was successful in projecting the

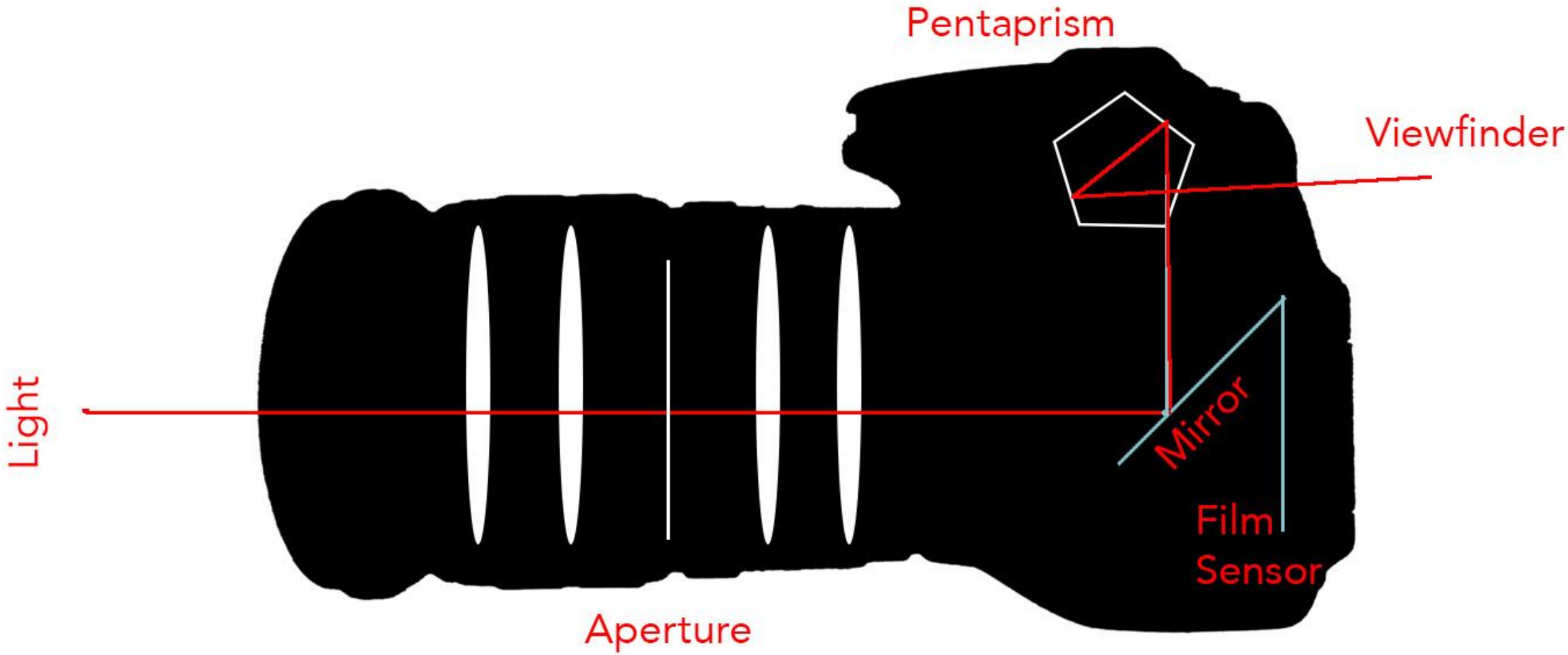
MOVEMENT OF SUN

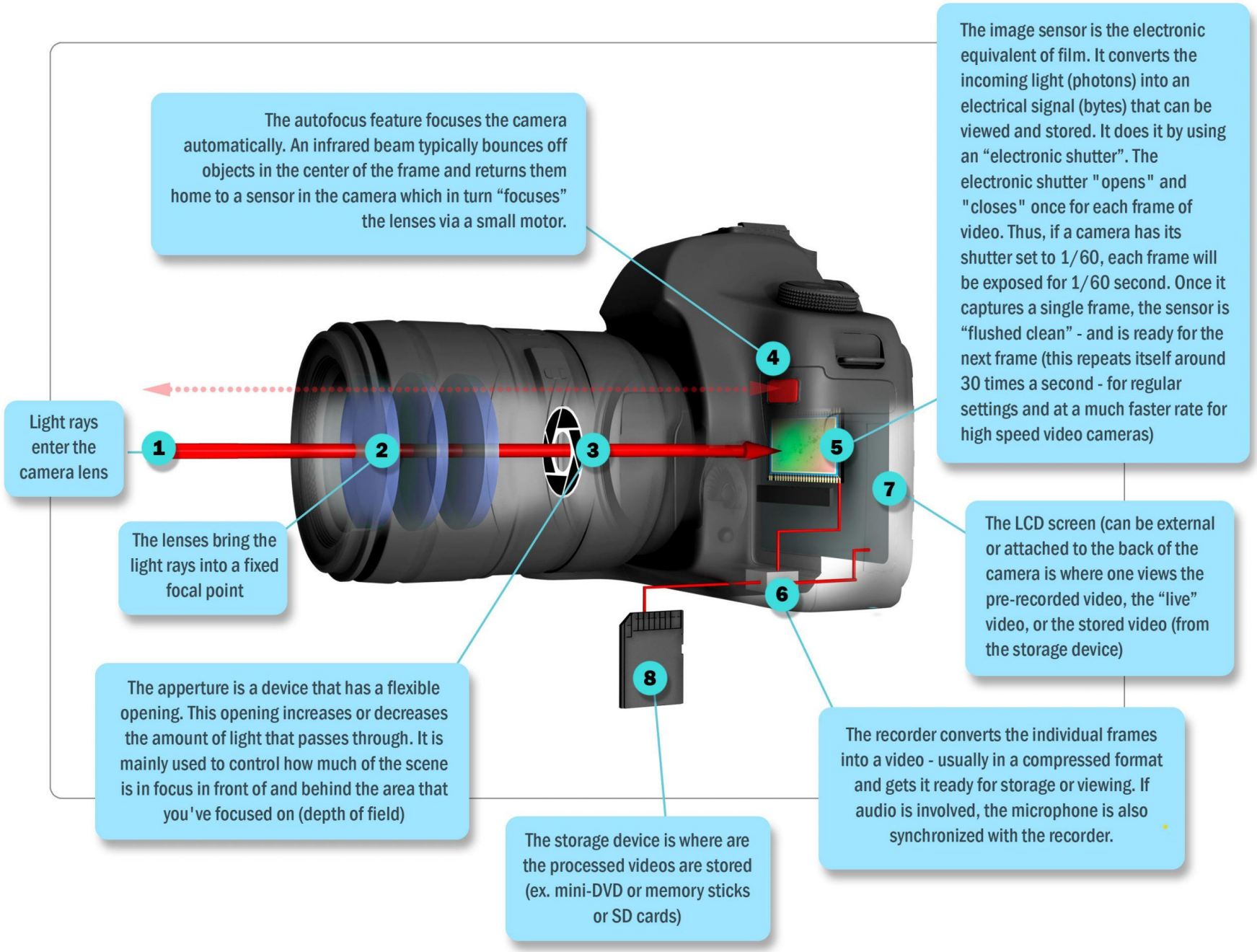
Limitations of Pin-Hole Camera



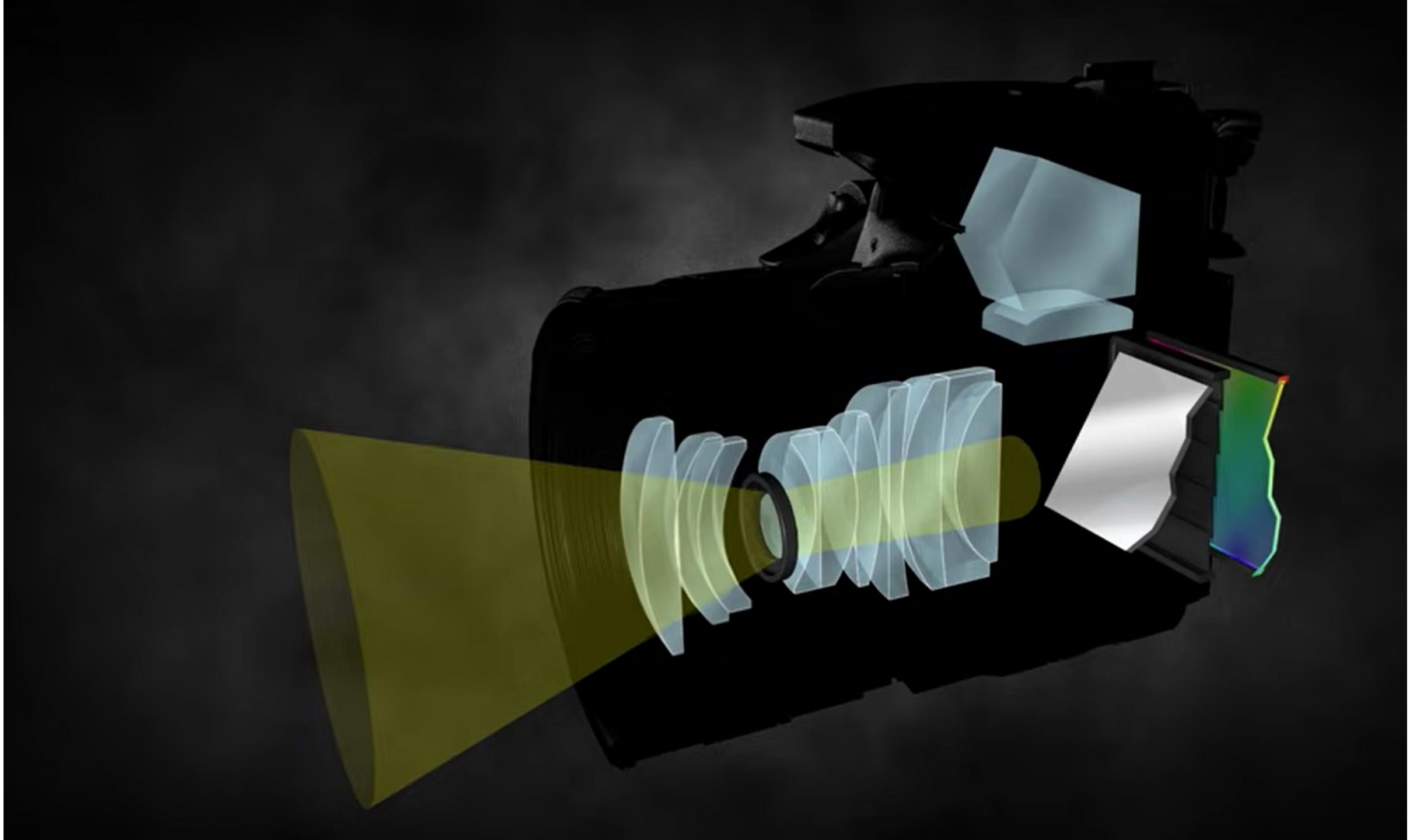


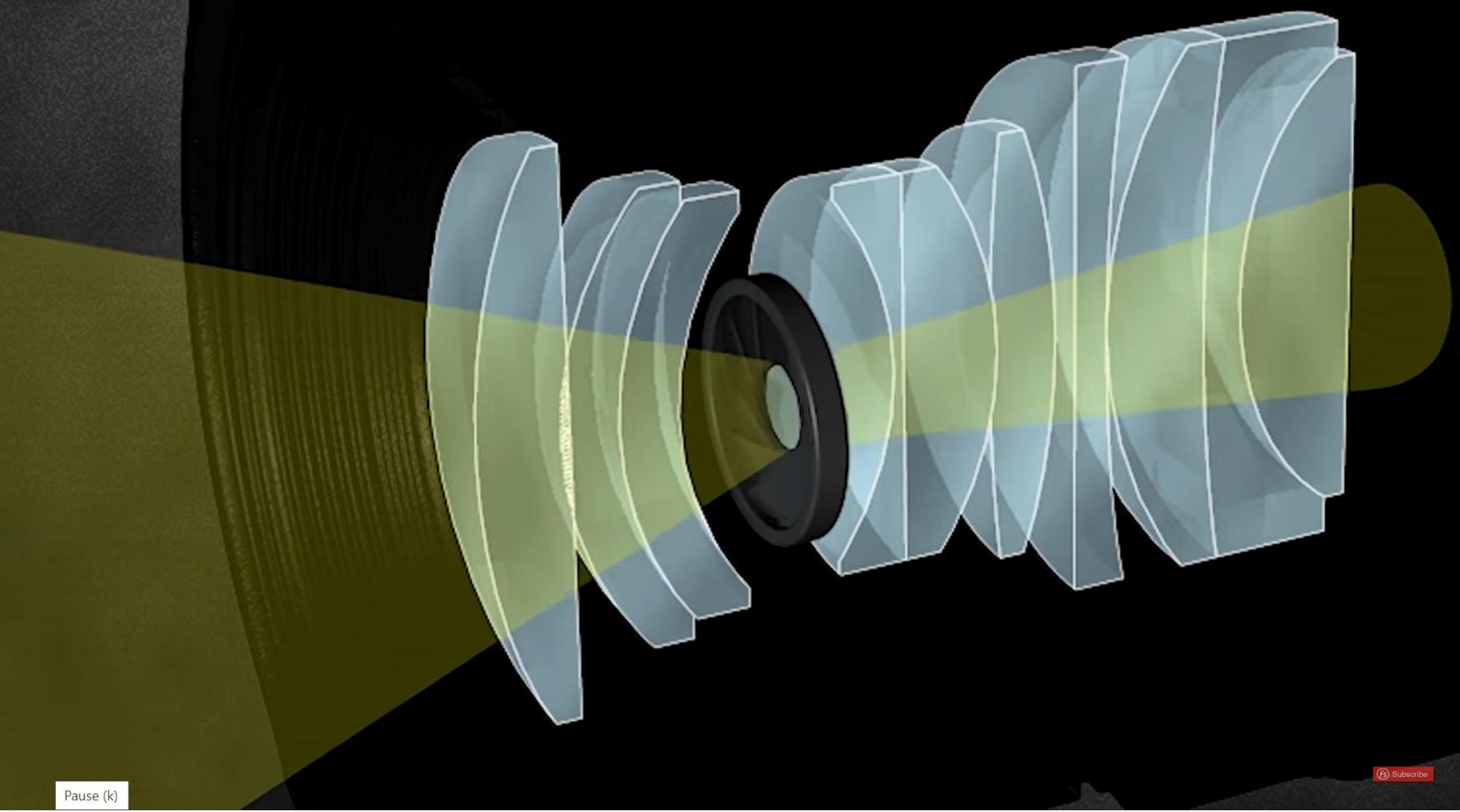

Button
Diaphragm
Lens
Aperture
Sensitive film
Sensitive film
Object
Lens
Image

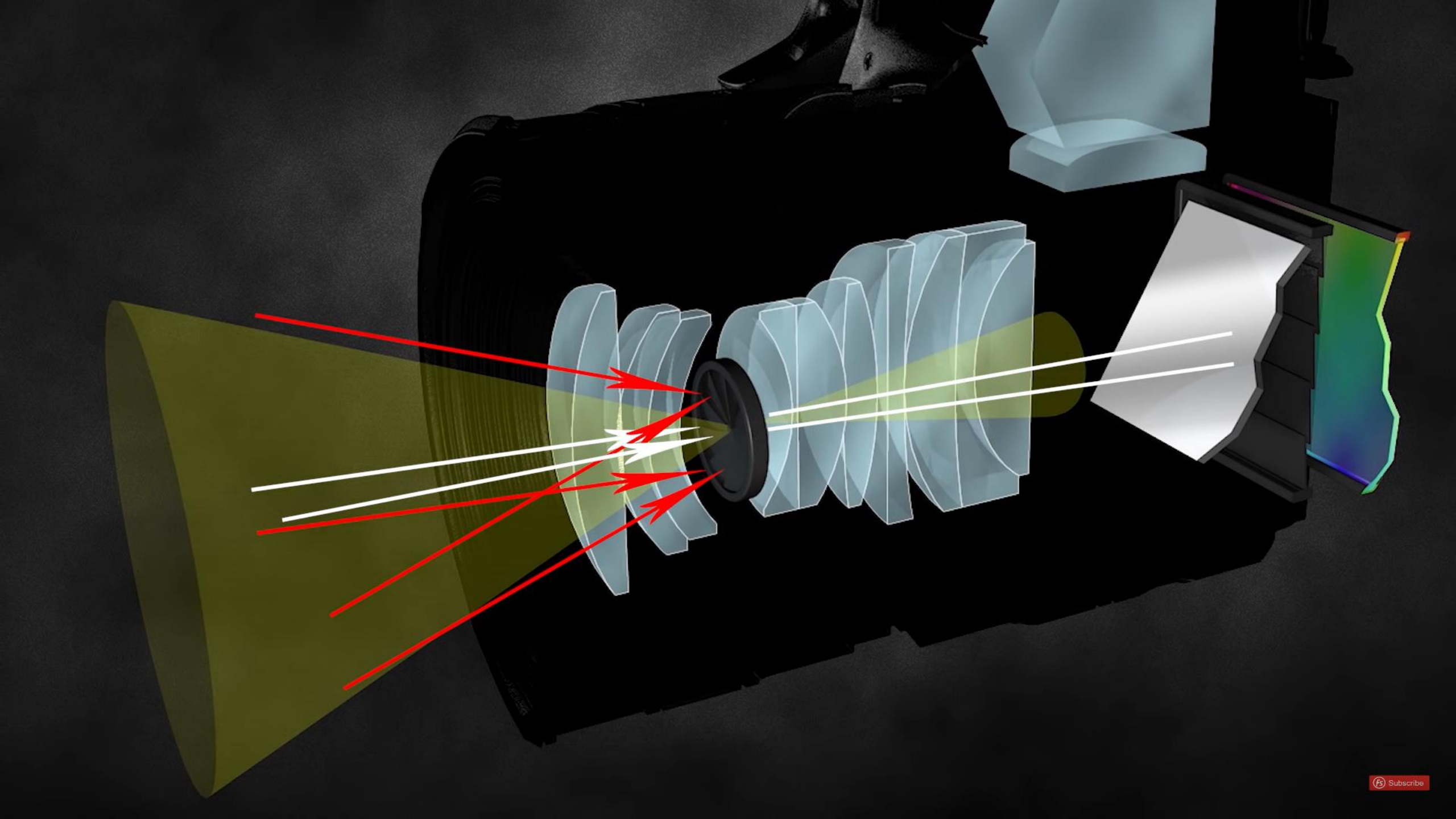

Pentaprism
Viewfinder
Light
Mirror
Film
Sensor
Aperture



https://trovideoproduction.com/2021/02/20/how-a-video-camera-works/




Pause (k)
Subscribe


Subscribe

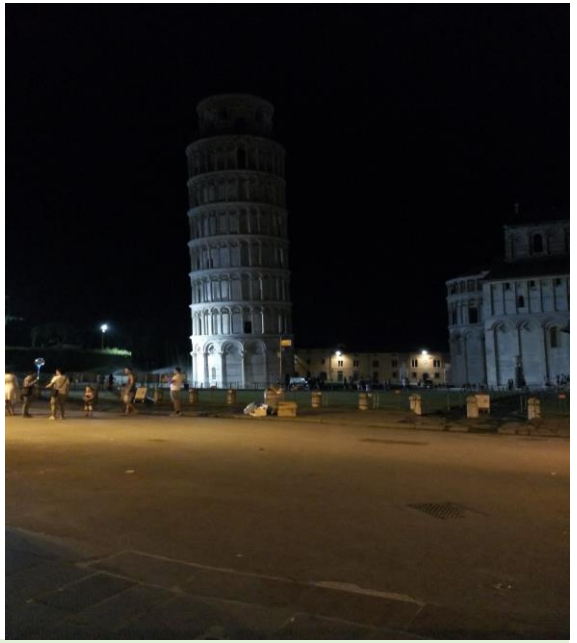
https://www.youtube.com/watch?v=W34sLbAsFhM

Digital Image

A **digital image** is an image composed of picture elements, also known as *pixels*, each with *finite*, *discrete quantities* of numeric representation for its intensity or gray level that is an output from its two-dimensional functions fed as input by its spatial coordinates denoted with *x*, *y* on the x-axis and y-axis, respectively.[1]

[1] *Gonzalez, Rafael (2018). Digital image processing. New York, NY: Pearson. ISBN 978-0-13-335672-4. OCLC 966609831*

Or



An image is defined as a two-dimensional function, **F(x,y)**, where x and y are spatial coordinates, and the amplitude of F at any pair of coordinates (x,y) is called the intensity of that image at that point. When x,y, and amplitude values of F are finite, we call it a digital image.

Digital Image Contd..

In other words, an image can be defined by a two-dimensional array specifically arranged in rows and columns

Digital Image is composed of a finite number of elements, each of which elements have a particular value at a particular location.

These elements are referred to as picture elements, image elements, and pixels.

A **Pixel** is most widely used to denote the elements of a Digital Image.

1.BINARY IMAGE– The binary image as its name suggests, contain only two-pixel elements i.e 0 & 1,where 0 refers to black and 1 refers to white. This image is also known as Monochrome.

2.BLACK AND WHITE IMAGE– The image which consist of only black and white color is called BLACK AND WHITE IMAGE.

3.8 bit COLOR FORMAT– It is the most famous image format. It has 256 different shades of colors in it and commonly known as Grayscale Image. In this format, 0 stands for Black, and 255 stands for white, and 127 stands for gray.

4.16 bit COLOR FORMAT– It is a color image format. It has **65,536** different colors in it. It is also known as High Color Format. In this format the distribution of color is not as same as Grayscale image.

Computer Vision Contd..

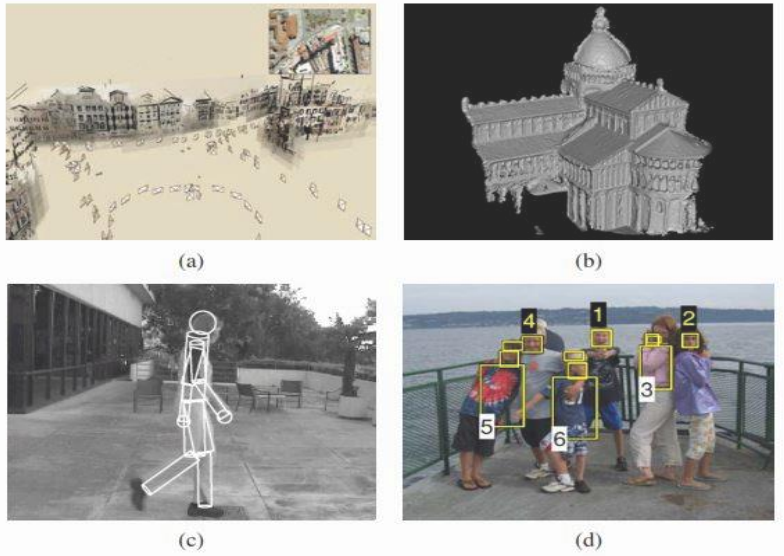


Figure 1.2 Some examples of computer vision algorithms and applications. (a) *Structure from motion* algorithms can reconstruct a sparse 3D point model of a large complex scene from hundreds of partially overlapping photographs (Snavely, Seitz, and Szeliski 2006) © 2006 ACM. (b) *Stereo matching* algorithms can build a detailed 3D model of a building façade from hundreds of differently exposed photographs taken from the Internet (Goesele, Snavely, Curless *et al.* 2007) © 2007 IEEE. (c) *Person tracking* algorithms can track a person walking in front of a cluttered background (Sidenbladh, Black, and Fleet 2000) © 2000 Springer. (d) *Face detection* algorithms, coupled with color-based clothing and hair detection algorithms, can locate and recognize the individuals in this image (Sivic, Zitnick, and Szeliski 2006) © 2006 Springer.

- We now have reliable techniques for accurately computing a **partial 3D model** of an environment from thousands of partially overlapping photographs (Figure 1.2a).

- Given a large enough set of views of a particular object or facade, we can create accurate dense 3D surface models using **stereo matching** (Figure 1.2b).

- We can track a **person moving** against a complex background (Figure 1.2c).

- We can even, with moderate success, **attempt to find and name all of the people in a photograph** using a combination of face, clothing, and hair detection and recognition (Figure 1.2d).

However, despite all of these advances, the dream of having a computer interpret an image at the same level as a two-year old (for example, counting all of the animals in a picture) remains elusive.

Why is vision so difficult? In part, it is because vision is an inverse problem, in which we seek to recover some unknowns given insufficient information to fully specify the solution.

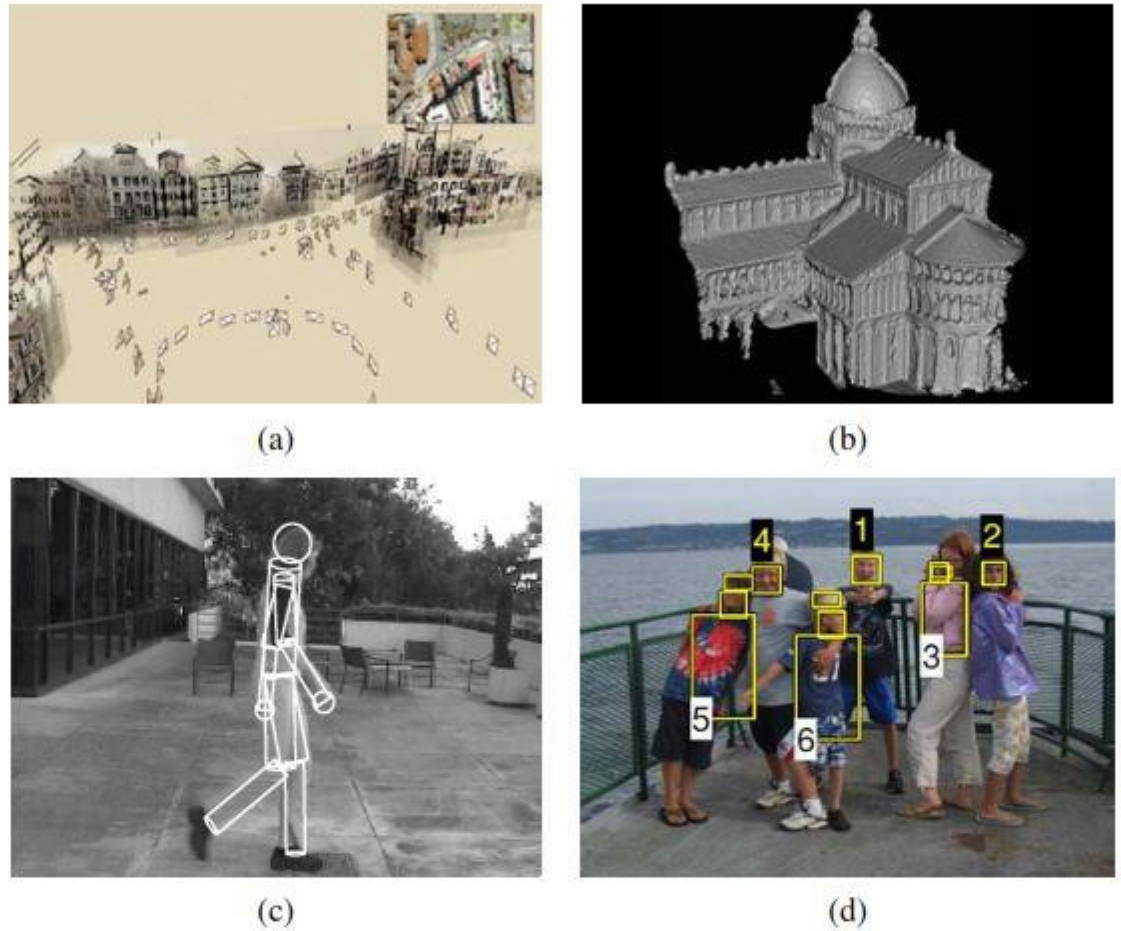


Figure 1.2 Some examples of computer vision algorithms and applications. (a) *Structure from motion* algorithms can reconstruct a sparse 3D point model of a large complex scene from hundreds of partially overlapping photographs (Snavely, Seitz, and Szeliski 2006) © 2006 ACM. (b) *Stereo matching* algorithms can build a detailed 3D model of a building façade from hundreds of differently exposed photographs taken from the Internet (Goesele, Snavely, Curless *et al.* 2007) © 2007 IEEE. (c) *Person tracking* algorithms can track a person walking in front of a cluttered background (Sidenbladh, Black, and Fleet 2000) © 2000 Springer. (d) *Face detection* algorithms, coupled with color-based clothing and hair detection algorithms, can locate and recognize the individuals in this image (Sivic, Zitnick, and Szeliski 2006) © 2006 Springer.

Computer Vision Contd..

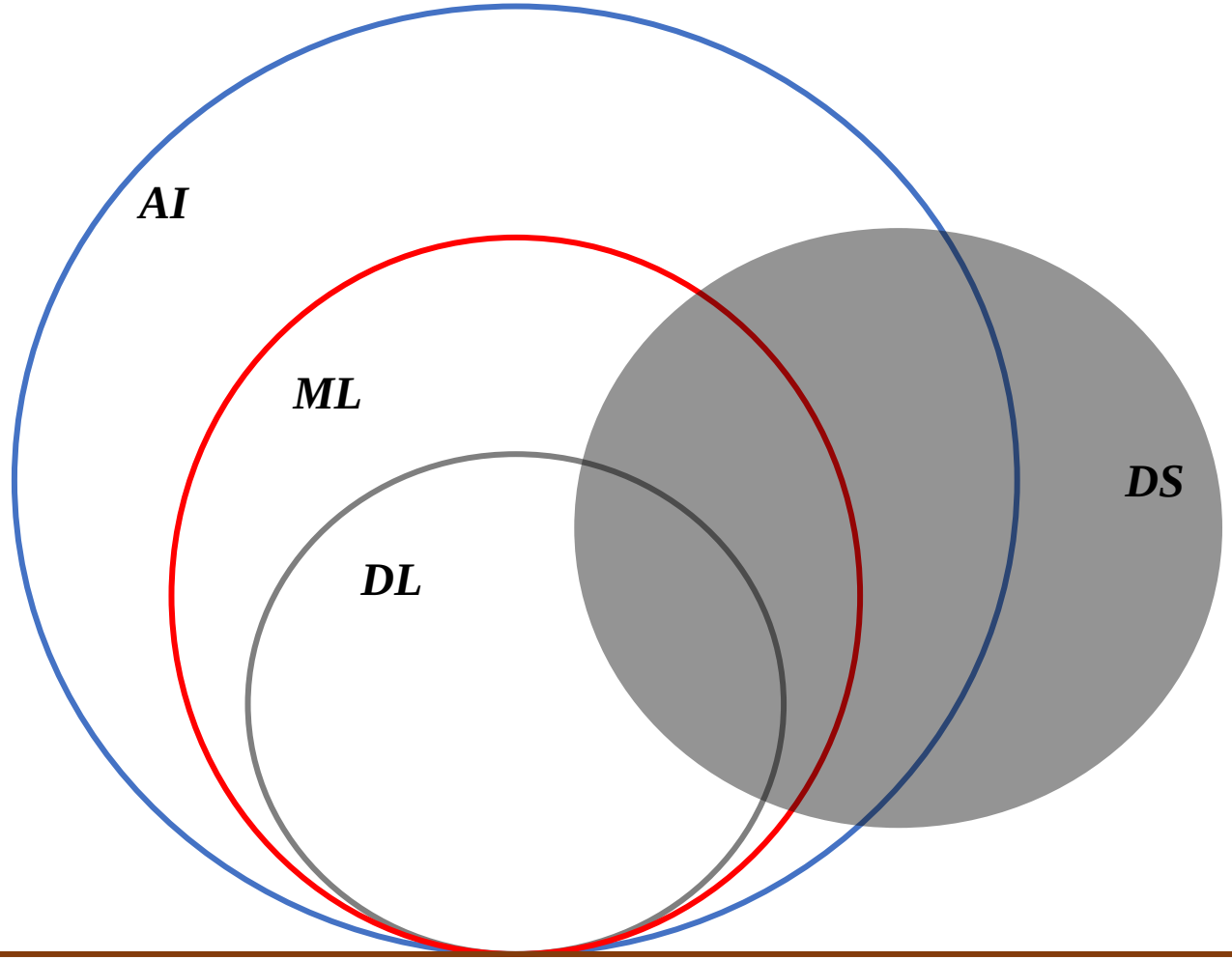
- Early efforts have made a great contribution to the philosophy of **human vision** and the **basic computational theory of computer vision** by exploiting well-designed features and feature descriptors combined with classical machine learning methods [3,4].

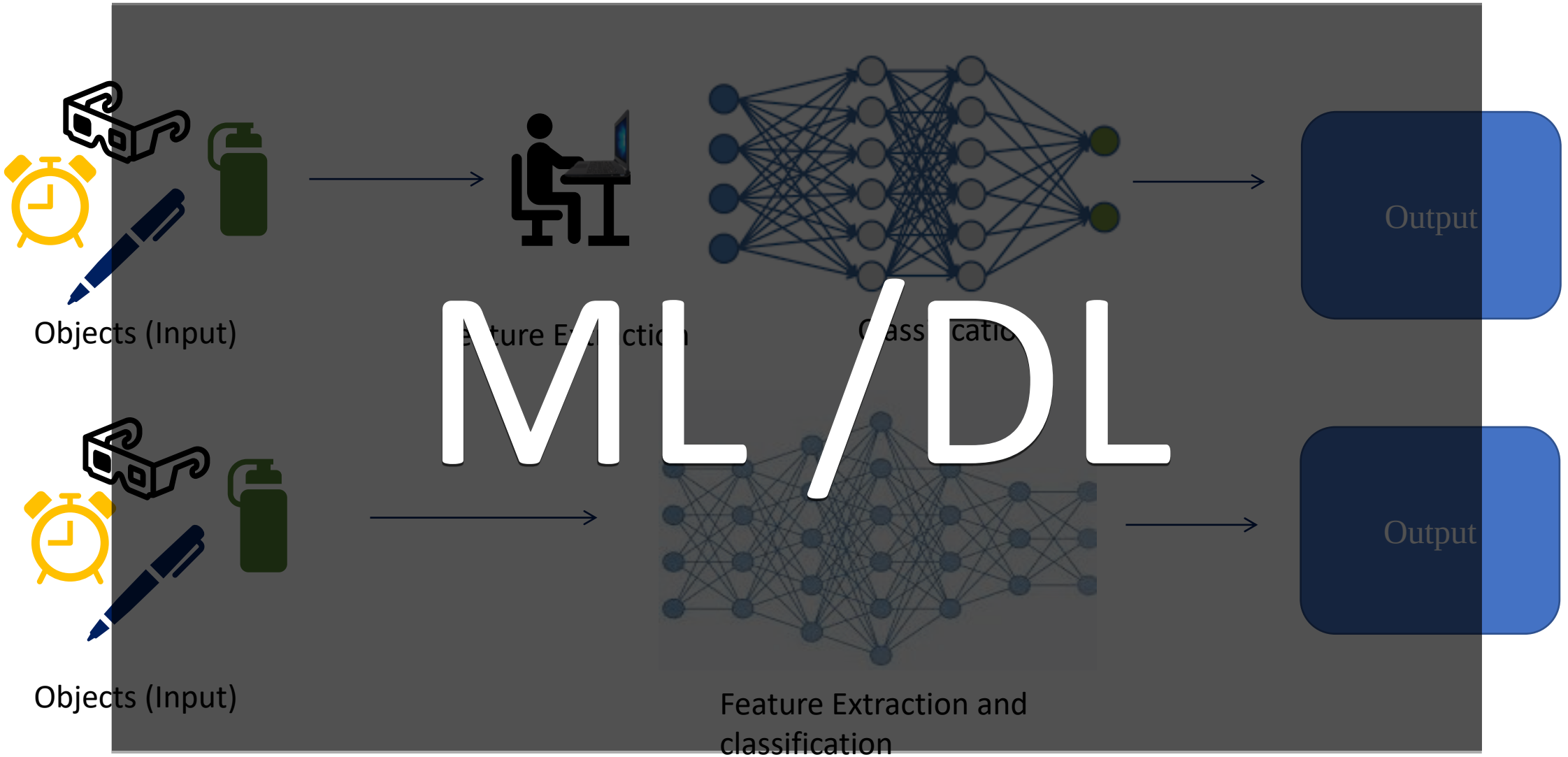
- Fortunately, researchers have believed that computer systems can go beyond regular object recognition and learn to reveal details and insights of the visual world by training them to see trillions of images and videos generated from Internet.

- To nourish the computer brain, the largest image classification dataset “ImageNet” [6] that contains **15 million images** across **22,000** classes of objects was created, upon which the well-known “**deep learning**” technology has demonstrated its overwhelming superiority over **traditional computer vison algorithms** that treat objects as a collection of shape and color features

Computer Vision Contd..

AI / ML / DL / DS




Objects (Input)
Output
ML/DL
Objects (Input)
Output
Feature Extraction and
classification

Computer Vision Contd..

- Deep learning is a particular class of machine learning algorithm, which typically simplifies the process of **feature extraction** and **description through a multi-layer convolutional neural network (CNN)**.

- CNN aims to transform the high-dimension input image into low-dimension yet highly-abstracted semantic output.

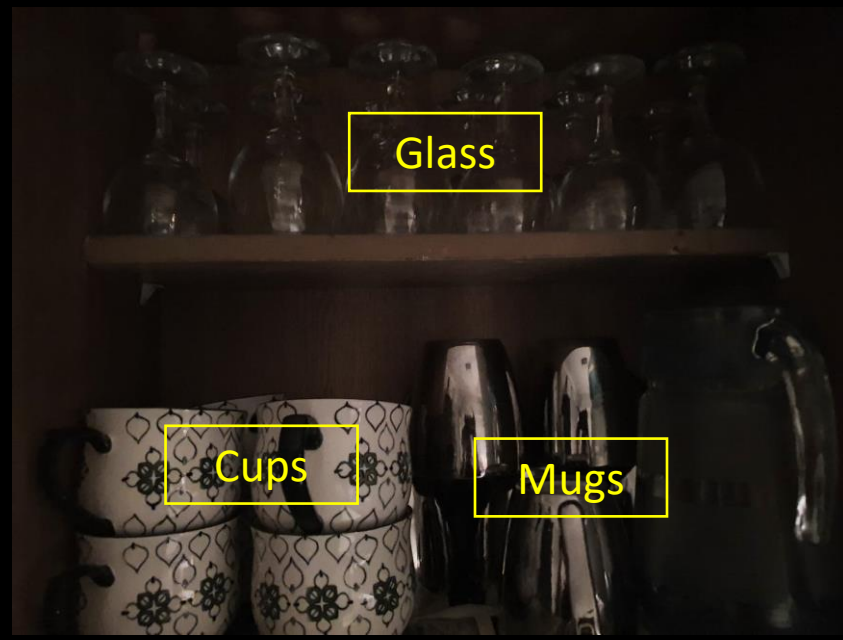
- The deep neural network (DNN) achieve the state-of-the-art performance and bring an unprecedented development of computer vision in both algorithms and hardware implementations.
- In recent years, CNN has become the de-facto standard computation framework in computer vision.

- Numbers of deeper and more complicated networks are developed to make CNNs deliver near human accuracy in many computer vision applications, such as classification, detection and segmentation.

- **The high accuracy, however, comes at the price of large computational cost. As a result, dedicated hardware platforms, from the general-purpose GPUs to application-specific processors, are investigated to optimize for DNN-based workloads.**

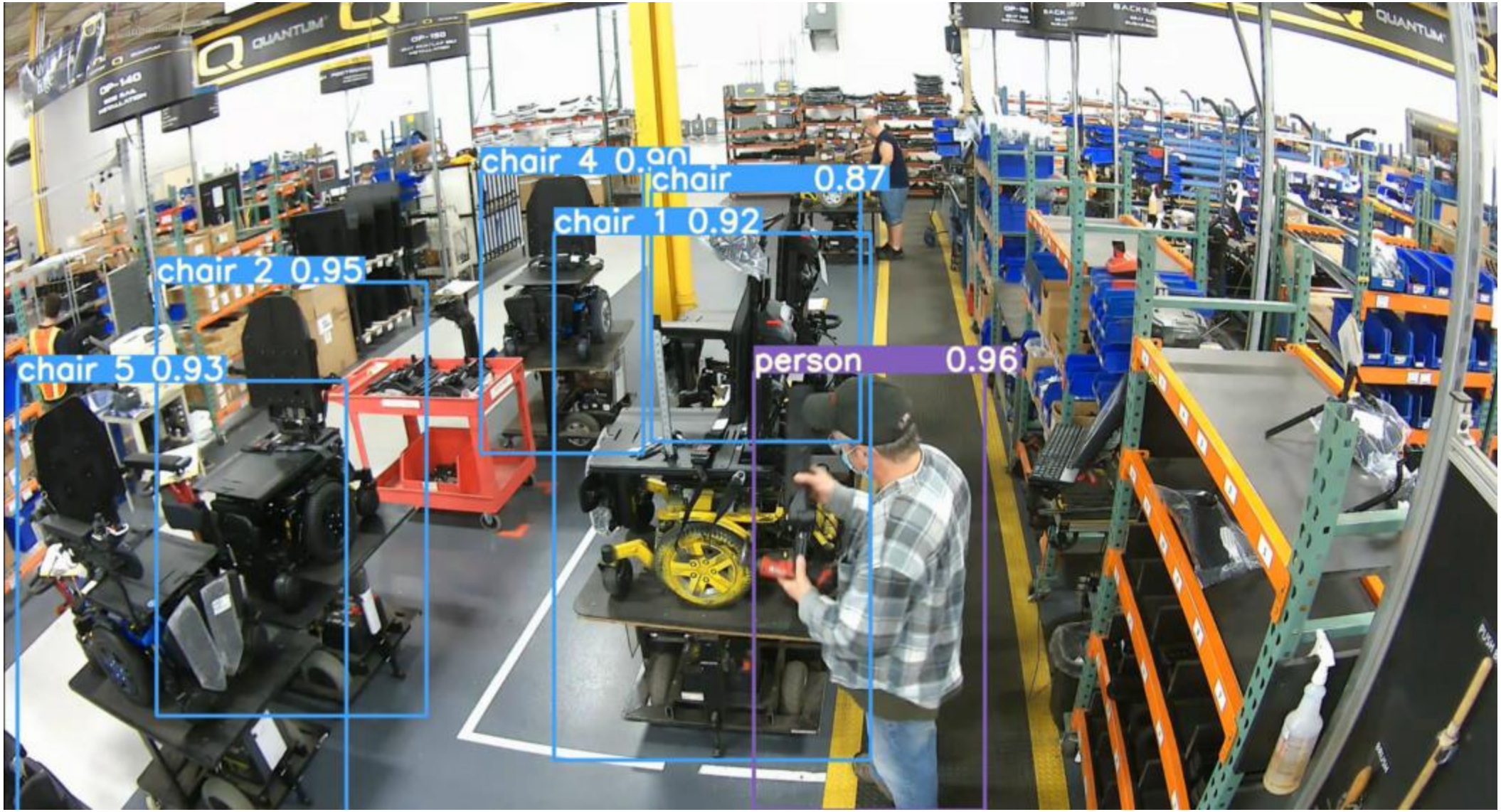
Computer Vision aspects

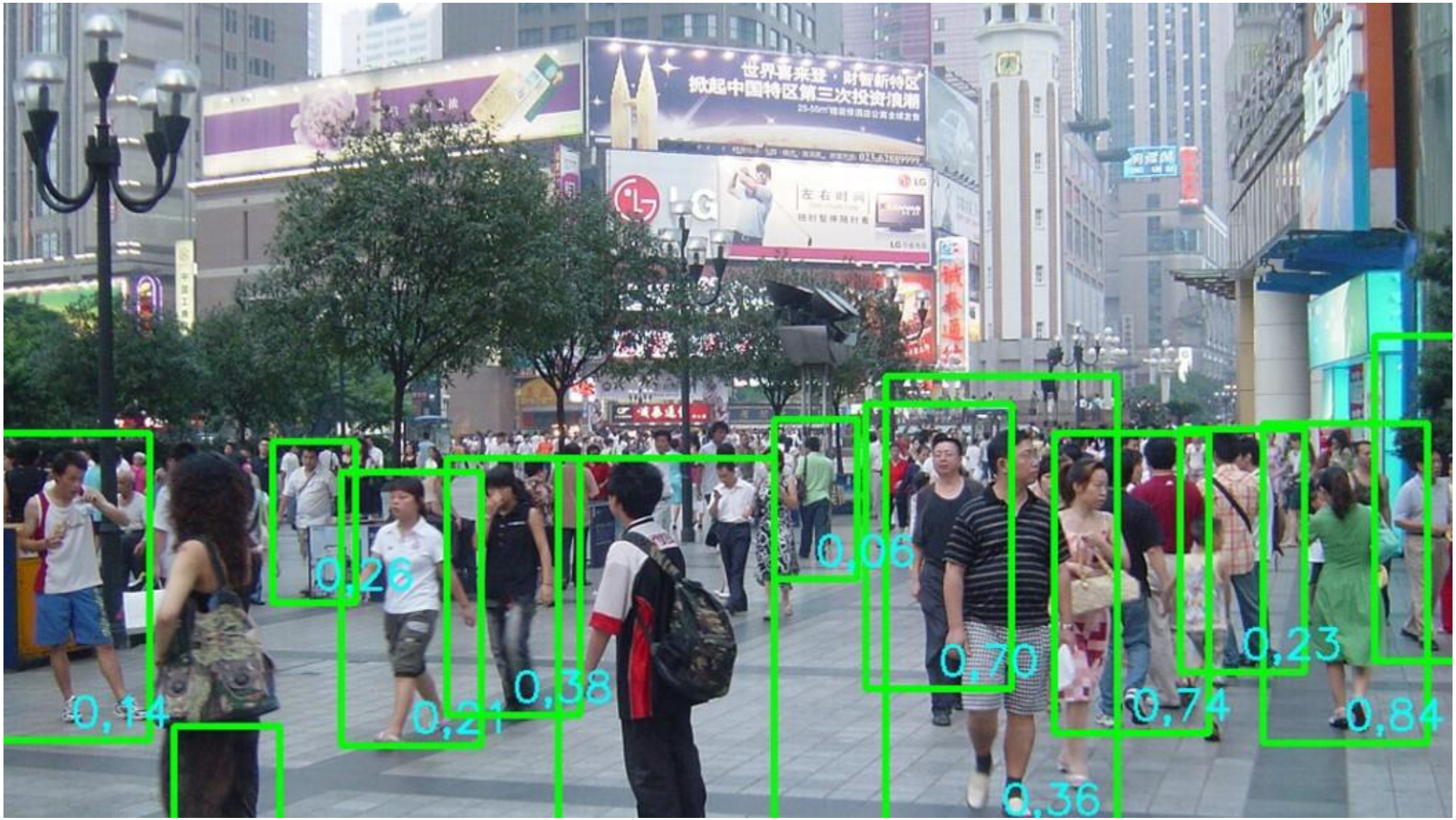
- **Image classification**
- **Object detection**
- **Image segmentation**
- **Visual Tracking**
- **Semantic Segmentation**
- **Image restoration**

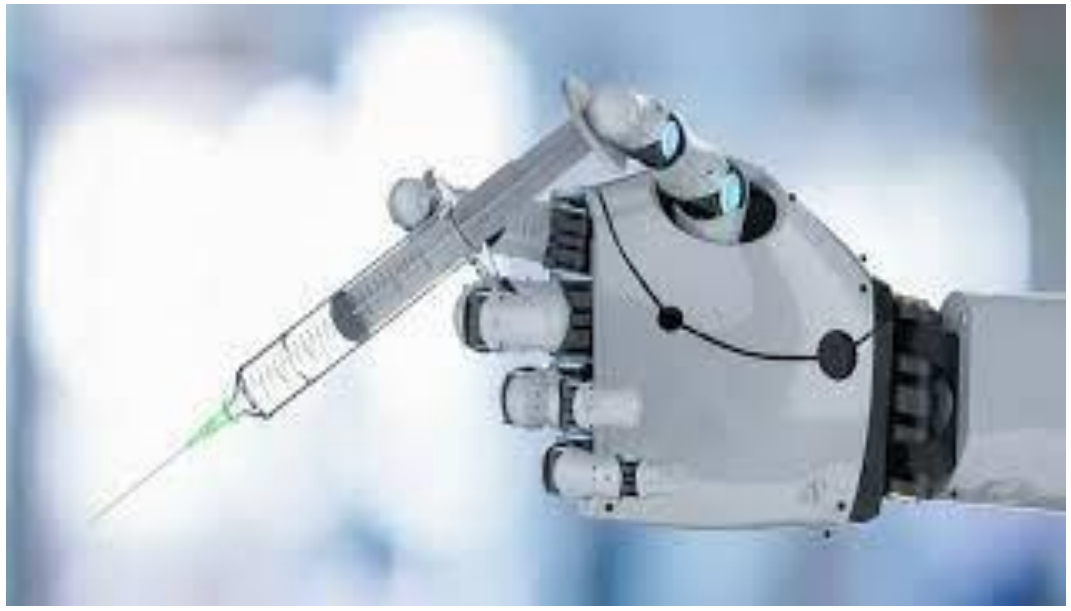



Smart Drone
for Surveying and fertilizer
GPS guided Automatically
Smart Analysis
Control , Monitoring
Data processing, collection
from drones and satellites
Smart Machinery
GPS guided and controlled
Self-driving Systems
Smart Farm Dash
WAYPOINT FLIGHT
Zone 1
Waypoints
Total Distance
95%
WIND
D 150m
H 35m

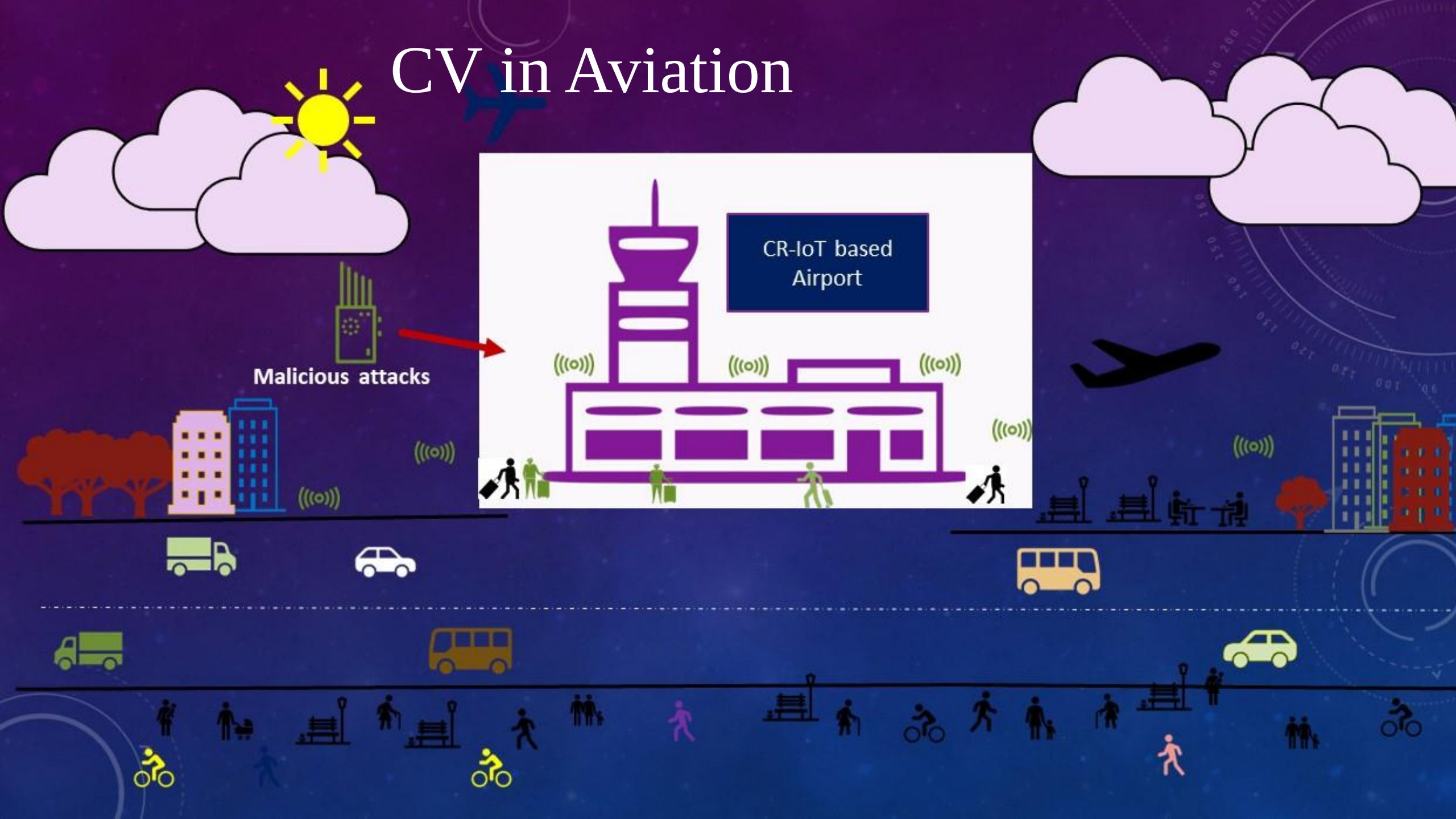

Patient Information
Exam Number
Accession Number
Patient ID
Patient Name
Sex
Birthdate
Age
Weight
Height
History
Protocol Number
Anatomical Selector
Default Protocol
BrightSpeed S SYS#ctb
Ex: 6904
Se: 2
IC I100.56
Im: 81
DFOV 47.3cm
STND/+/E
Acc Num:DM01378616
M 59 QP-545301
DOB: 21 Sep 1954
17 Jan 2014
512
kV 120
mA 318


chair 4 0.90
chair 0.87
chair 1 0.92
chair 2 0.95
chair 5 0.93
person 0.96

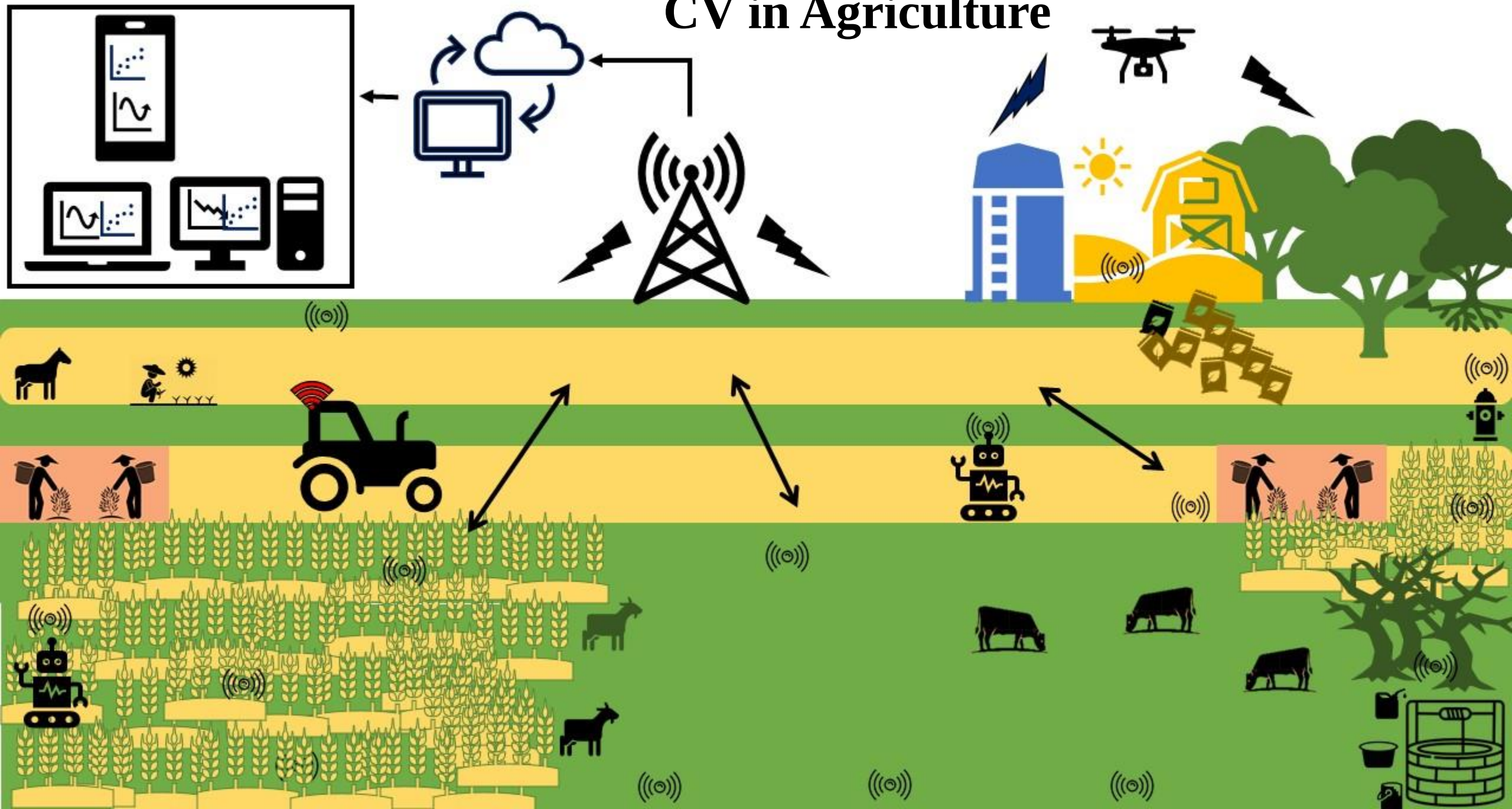

0,14
0,26
0,21
0,38
0,06
0,70
0,36
0,74
0,23
0,84



CV in Aviation



CV in Agriculture



Precision Livestock Counting

Leverage high-accuracy, field-proven methodologies with Plainsight's patent-pending technology and edge computing efficiency to eliminate error-prone manual livestock inventory. If desired, our experts will tailor your counting models to:

- Monitor transport on- & off-loading, feedlot entries/exits
- Automate animal detection & livestock counts across multiple locations
- Integrate insights into operational/financial systems
- Digitize existing receiving process & curate auditable data records
- Re-allocate staff to support other functions
- Easily operationalize repeatable solutions at scale

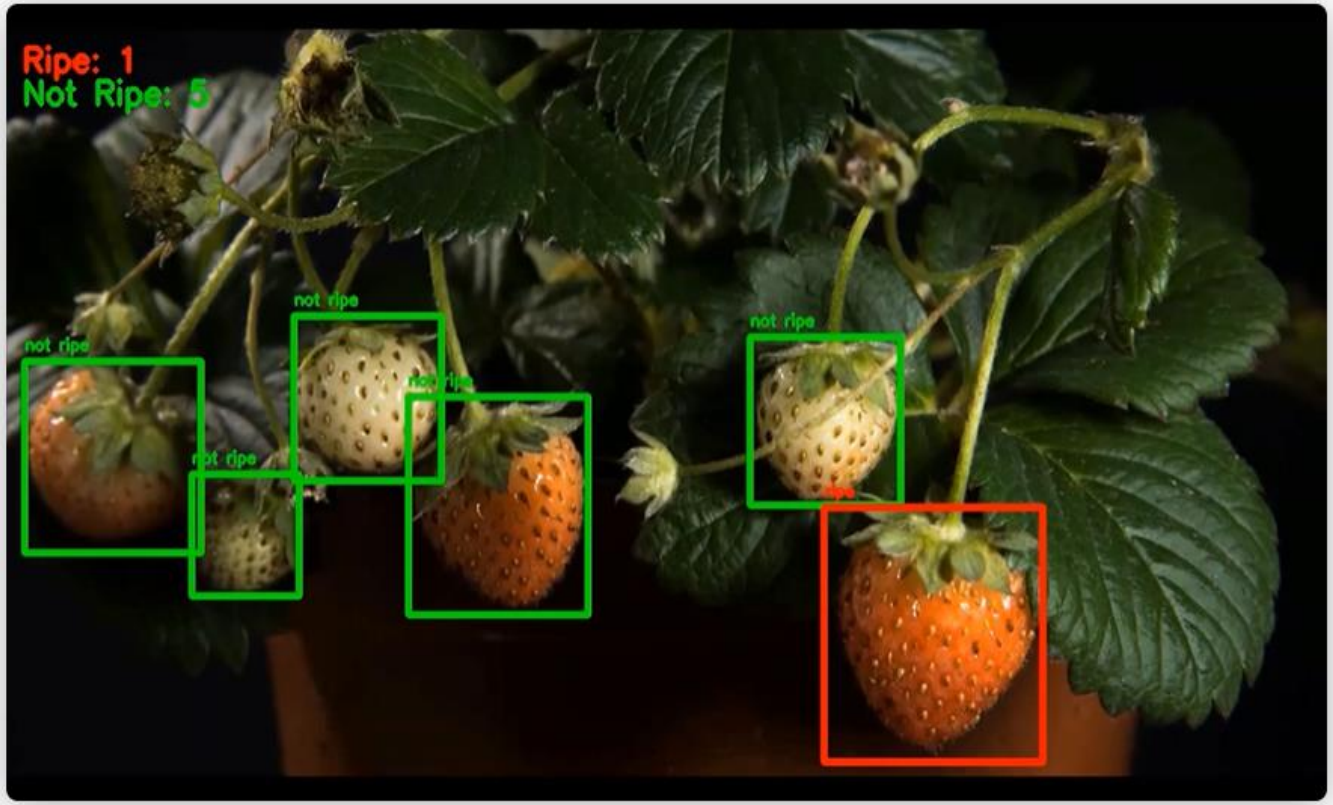
[10] https://plainsight.ai/industries/agriculture/?utm_term=smart%20agriculture&utm_campaign=Ag+Test&utm_source=adwords&utm_medium=ppc&hsa_acc=1094431452&hsa_cam=17450374732&hsa_grp=145057913984&hsa_ad=607275160697&hsa_src=g&hsa_tgt=kwd-335721596098&hsa_kw=smart%20agriculture&hsa_mt=p&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwuaiXBhCCARIsAKZLt3msR8utFAIejhlG4aG-zgfN4ou38jF8tuqo84pVE8AE_a9Qpi7V9CYaAhjCEALw_wcB

Harvesting & Processing Quality Assurance

Implement vision-based best practices to ensure quality and safety of produce with accuracy and efficiency.

- Detect real-time hazards such as foreign objects from production environments, equipment or inputs due to human handling
- Monitor process & safety adherence
- Automate alerts to potential risks & contaminants
- Create visual data records for traceability & remediation

https://plainsight.ai/industries/agriculture/?utm_term=smart%20agriculture&utm_campaign=Ag+Test&utm_source=adwords&utm_medium=ppc&hsa_acc=1094431452&hsa_cam=17450374732&hsa_grp=145057913984&hsa_ad=607275160697&hsa_src=g&hsa_tgt=kwd-335721596098&hsa_kw=smart%20agriculture&hsa_mt=p&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwuaiXBhCCARIsAKZLt3msR8utFAIejhlG4aG-zgfN4ou38jF8tuqo84pVE8AE_a9Qpi7V9CYaAhjCEALw_wcB



Fruits & Vegetables Ripeness Detection

Automate the detection and evaluation of fruit and vegetable maturation to enhance genetics, growing, and QA/QC processes.

- Inspect & measure levels/grades of vision-derivable characteristics including: Size, shape, firmness color, bruising, brix (fruit flavor, sweetness, etc), overall condition
- Monitor variations to determine seasonal, environmental, varietal, or management factors
- Assist genetic selection for high quality & longer shelf life

[10] https://plainsight.ai/industries/agriculture/?utm_term=smart%20agriculture&utm_campaign=Ag+Test&utm_source=adwords&utm_medium=ppc&hsa_acc=1094431452&hsa_cam=17450374732&hsa_grp=145057913984&hsa_ad=607275160697&hsa_src=g&hsa_tgt=kwd-335721596098&hsa_kw=smart%20agriculture&hsa_mt=p&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwuaiXBhCCARIsAKZLt3msR8utFAIejhlG4aG-zgfN4ou38jF8tuqo84pVE8AE_a9Qpi7V9CYaAhjCEALw_wcB

Livestock Health Monitoring

Understand the health of individual animals and livestock populations with continuous visual data-driven insights. Predict and intervene with early detection and accurate response.

- Monitor & track livestock feeding & care
- Evaluate animal health by gait & size
- Identify cattle for health data via facial recognition, and combine with biometrics, & genetic data
- Detect visual indications of disease with monitoring for skin, behavioral abnormalities

[10] https://plainsight.ai/industries/agriculture/?utm_term=smart%20agriculture&utm_campaign=Ag+Test&utm_source=adwords&utm_medium=ppc&hsa_acc=1094431452&hsa_cam=17450374732&hsa_grp=145057913984&hsa_ad=607275160697&hsa_src=g&hsa_tgt=kwd-335721596098&hsa_kw=smart%20agriculture&hsa_mt=p&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwuaiXBhCCARIsAKZLt3msR8utFAIejhlG4aG-zgfN4ou38jF8tuqo84pVE8AE_a9Qpi7V9CYaAhjCEALw_wcB

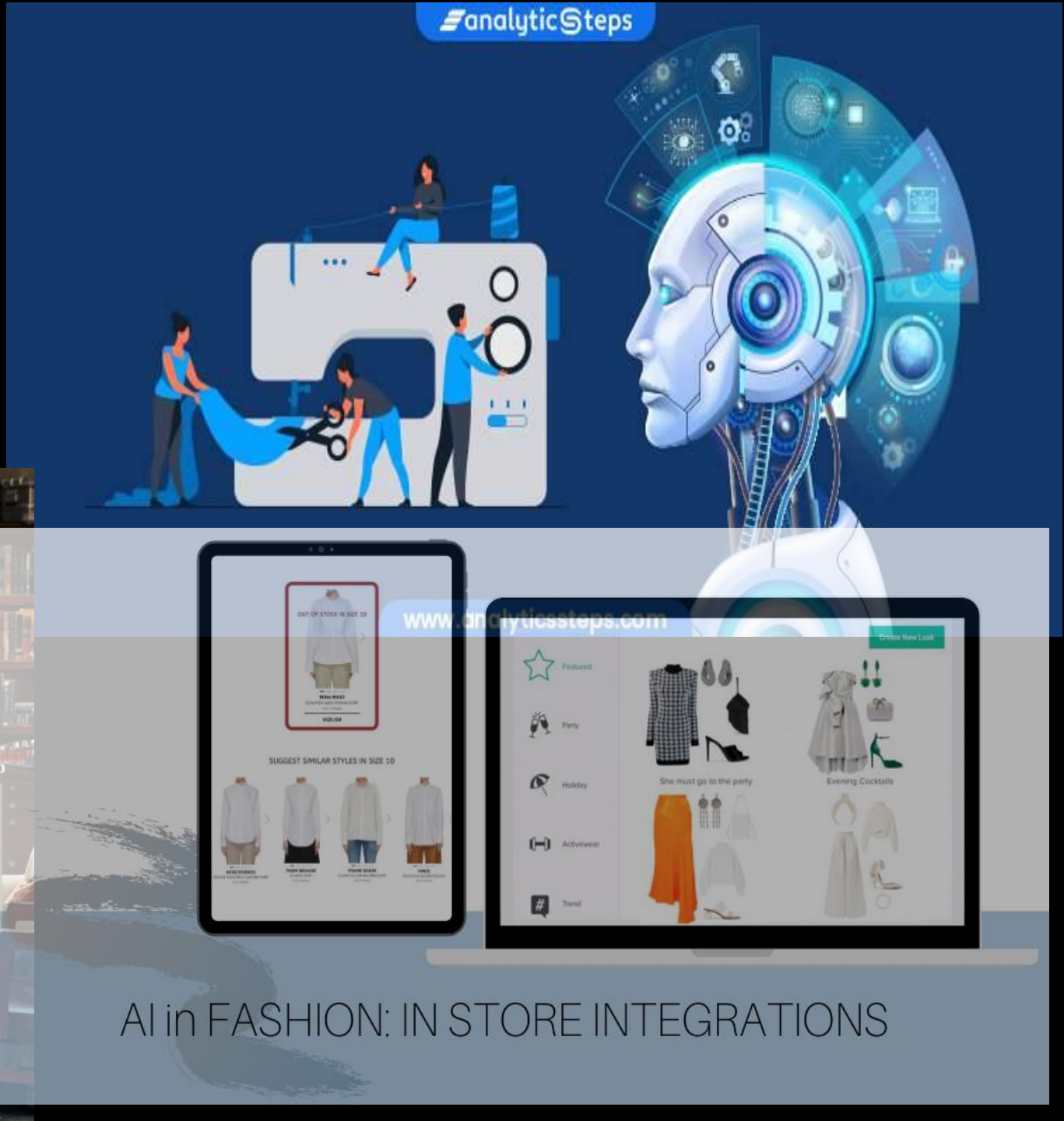
Crop Management & Risk Recognition

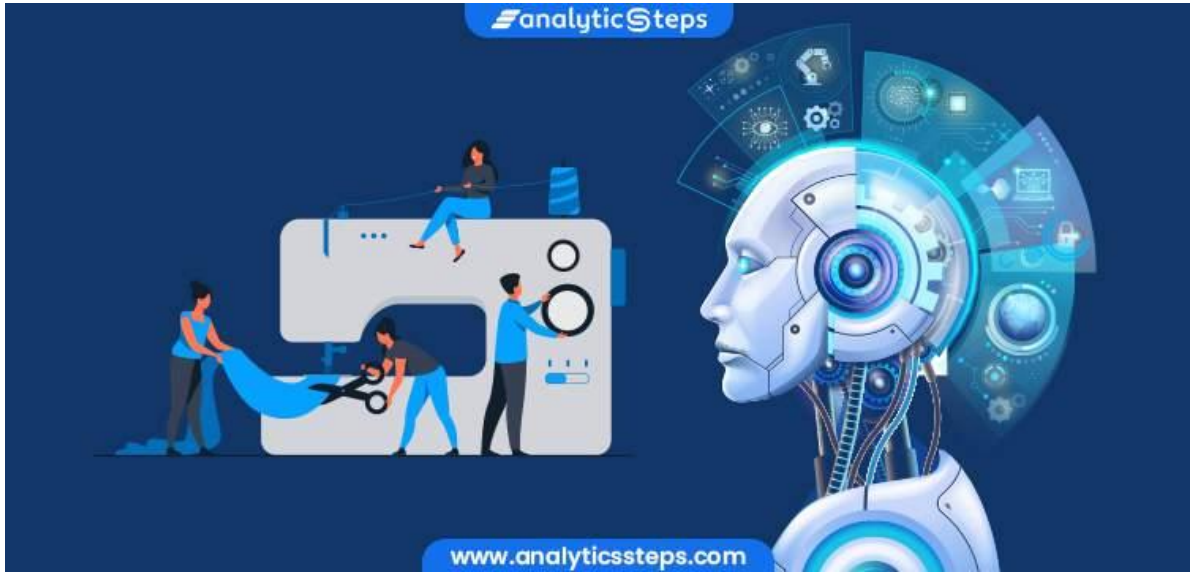
Improve knowledge and productivity with vision AI-based accuracy to solve problems before large-scale damage is done, ensuring consistent high-quality operations and high-yield production.

- Inventory plant population & measure volume
- Gauge crop health & productivity
- Capture plant color, shape, & texture
- Automate early detection & measure plant stress, disease, damage
- Identify weeds, invasive species, & pests

[10] https://plainsight.ai/industries/agriculture/?utm_term=smart%20agriculture&utm_campaign=Ag+Test&utm_source=adwords&utm_medium=ppc&hsa_acc=1094431452&hsa_cam=17450374732&hsa_grp=145057913984&hsa_ad=607275160697&hsa_src=g&hsa_tgt=kwd-335721596098&hsa_kw=smart%20agriculture&hsa_mt=p&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjwuaiXBhCCARIsAKZLt3msR8utFAIejhlG4aG-zgfN4ou38jF8tuqo84pVE8AE_a9Qpi7V9CYaAhjCEALw_wcB

AI in Fashion Industry

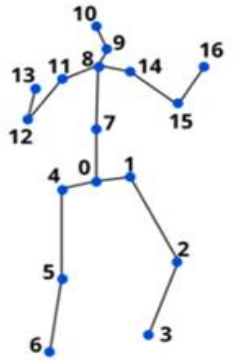




AnalyticSteps

3d Keypoints and their specification

0 — Bottom torso
1 — Left hip
2 — Left knee
3 — Left foot
4 — Right hip
5 — Right knee
6 — Right foot
7 — Center torso
8 — Upper torso

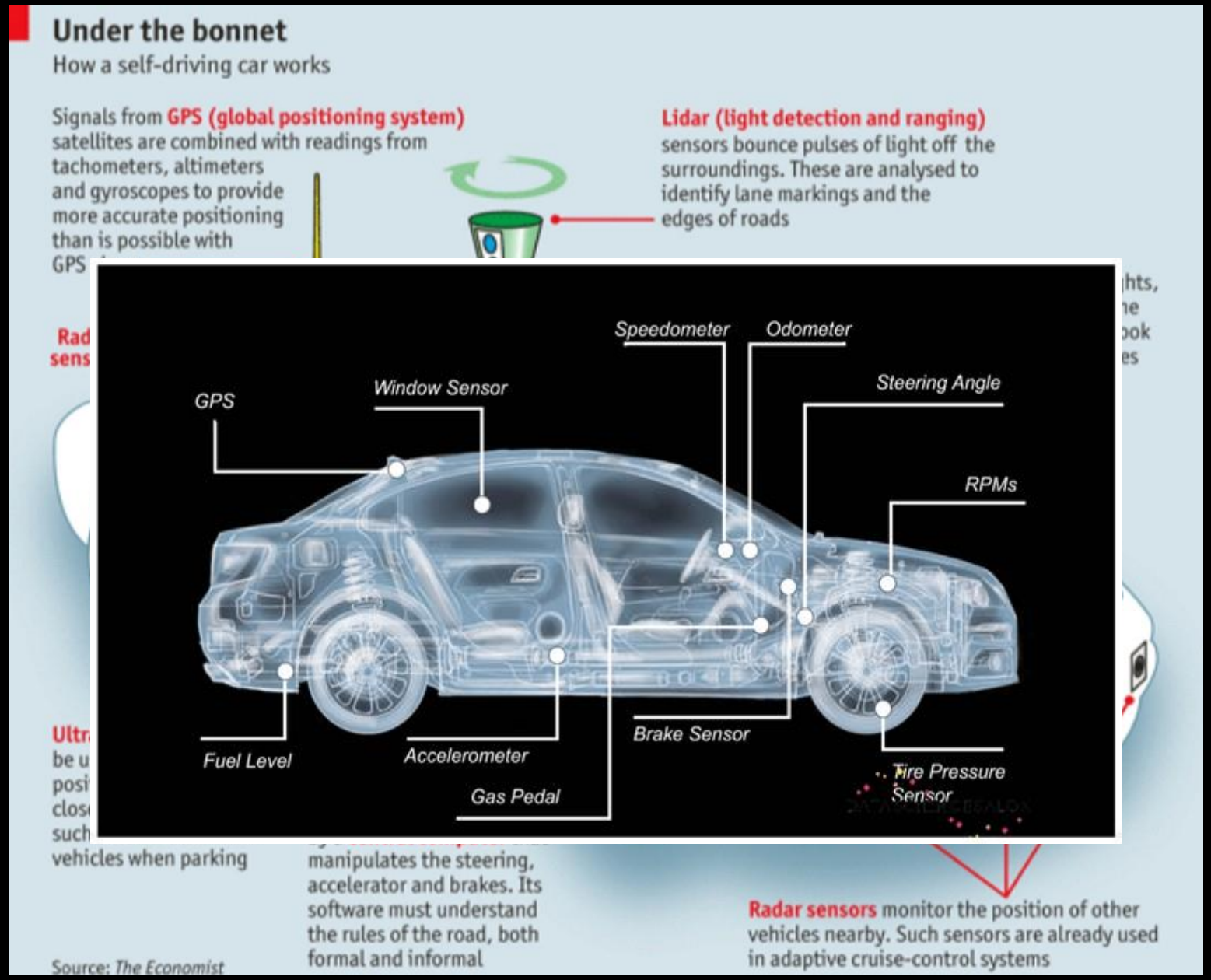


9 — Neck base
10 — Center head
11 — Right shoulder
12 — Right elbow
13 — Right hand
14 — Left shoulder
15 — Left elbow
16 — Left hand

www.analyticssteps.com

CV in Sports

CV in smart cars



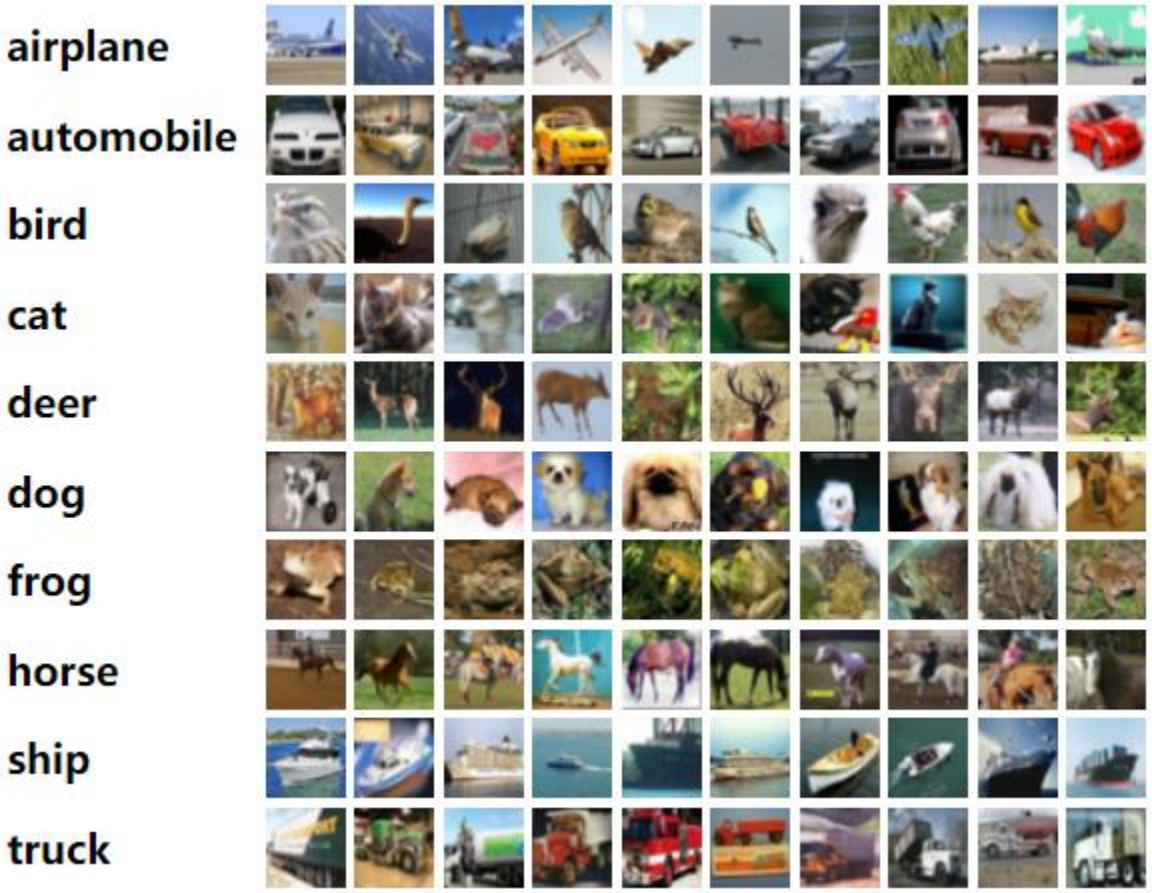
[13] https://medium.com/@jadesolaadedokun/use-case-for-data-science-in-self-driving-cars-9b0ca92f16e2

Computer Vision aspects

Image classification

- Image classification is a kind of biologically primary ability of human visual perception system.
- It has been an active task and plays a crucial role in the field of computer vision, which aims to automatically classify images into pre-defined classes.
- Traditionally, classification models can perform well only on **small datasets** such as CIFAR-10 [11] and MNIST [12].
- The great-leap-forward development of image classification occurred when the **large-scale image dataset** "ImageNet" was created by **Feifei L**i in 2009 [6].
- It was almost the same time when the well-known deep learning technologies started to show great performance in classification and stepped onto the stage of computer vision.

The **CIFAR-10 dataset** consists of **60000** 32x32 colour images in **10** classes, with **6000** images per class. There are **50000** training images and **10000** test images.



https://www.cs.toronto.edu/~kriz/cifar.html

The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class.

The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

The CIFAR-100 dataset

This dataset is just like the CIFAR-10, except it has **100 classes** containing **600 images each**. There are **500 training images** and **100 testing images per class**. The 100 classes in the CIFAR-100 are grouped into **20 superclasses**. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

The **MNIST database** (Modified National Institute of Standards and Technology database[1]) is a large database of handwritten digits that is commonly used for training various image processing systems

The MNIST database contains 60,000 training images and 10,000 testing images.[8] Half of the training set and half of the test set were taken from NIST's training dataset, while the other half of the training set and the other half of the test set were taken from NIST's testing dataset

The **ImageNet dataset** contains over a **million images** with labels and bounding boxes.
The dataset was created based on the Wordnet hierarchy. Every important concept in WordNet is called a “synonym set” or “synset”.
The majority of synsets in ImageNet are nouns (80,000+) and there are more than 100,000 synsets in total. Each image in the dataset is annotated by humans and is quality controlled.

ImageNet is one of the most benchmarked datasets in the history of computer vision.

This dataset spans 1000 object classes and contains **1,281,167** training images, **50,000 validation images** and **100,000 test images**.

https://datasets.activeloop.ai/docs/ml/datasets/imagenet-dataset/
https://www.image-net.org/

Computer Vision Aspects

Image classification

- Before the explosion of deep learning methods, research works put lots of efforts in designing scale-invariant features:

  (e.g. SIFT [13], HOG [14], GIST [15]), feature representations (e.g. Bag-of-Features [16], Fisher Kernel [17]) and classifiers (e.g. SVM [18]) for image classification [19,20].

- However, these manually crafted features work against objects in natural images with complicated background, variant color, texture, illumination and ever-changing poses and view factors.

- **At the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, AlexNet [21] won the first prize by a significant margin over the second place that was based on SIFT and Fisher Vectors (FVs) [20].**

*It demonstrates that the **classification model** based on **deep CNN** performs much more robustly than other **conventional methods** in the presence of **large-scale variations**.*

Global image descriptor (GIST) **Scale-invariant feature transform (SIFT) Histogram of Oriented Gradients**

Computer Vision Aspects

Image classification

- A typical deep CNN model consists of several convolution layers followed by activation functions and pooling layers, and several fully connected layers before prediction. It comes into deep structure to facilitate filtering mechanisms by performing convolutions in multi-scale feature maps, leading to highly abstract and discriminative features.

- AlexNet has 8 convolution layers, 3 pooling layers and 3 fully connected layers, with a total of 60 million parameters. It successfully uses ReLU as the activation function instead of sigmoid. Furthermore, data augmentation and dropout are widely used today as efficient learning strategies. AlexNet is hence known as the foundation work of modern deep CNN.

Computer Vision Aspects

Image classification

- Inspired by AlexNet, VGGNet [22] and GoogleNet [23] focus on designing deeper networks to further improve accuracy. They were the runner-up and winner of ILSVRC in 2014 respectively.

- By repeatedly stacking 3 x 3 convolutional kernels and 2 x 2 maximum pooling layers, **VGGNet** successfully constructs a convolutional neural network of 16–19 layers.

- Although deeper networks offer better accuracy, simply increasing the number of layers cannot continuously improve accuracy because of **vanishing/exploding gradient** information during network training.

- **ResNet [24**], which makes another great progress of deep network structure, proposes to use a shortcut connection between residual blocks to make full use of information from previous layers and keep the gradients during backward propagation.

By using this residual block, ResNet successfully trains very deep networks with up to 152 layers and was the winner of ILSVRC in 2015.

Computer Vision Aspects

Image classification

- Following the idea of ResNet, DenseNet [25] establishes connections between all previous layers and the current layer. It concatenates and, therefore, reuses the features from all previous layers.
- DenseNet presents with great advantage in classification accuracy on ImageNet.

Summary of different CNN models on ImageNet classification task.

Model	Time	Accuracy	Num. of Parameters	Num. of FLOPs	Num. of Layers
AlexNet [21]	2012	57.2%	60 M	720 M	8
VGGNet [22]	2014	71.5%	138 M	15,300 M	16
GoogleNet [23]	2014	69.8%	6.8 M	1,500 M	22
ResNet [24]	2015	78.6%	55 M	2,300 M	152
DenseNet [25]	2017	79.2%	25.6 M	1,150 M	190
SENet [26]	2017	82.7%	145.8 M	42,300 M	–
NASNet [27]	2018	82.7%	88.9 M	23,800 M	–
SqueezeNet [29]	2016	57.5%	1.2 M	833 M	–
MobileNet [30]	2017	70.6%	4.2 M	569 M	28
ShuffleNet [31]	2018	73.7%	4.7 M	524 M	–
ShiftNet-A [32]	2018	70.1%	4.1 M	1,400 M	–
FE-Net [33]	2019	75.0%	5.9 M	563 M	–

Computer Vision Aspects

Image classification

- By using ResNet or DenseNet as the major backbone structure, researchers focus on improving the functionality of neural network blocks. SENet [26], which was the winner of **ILSVRC 2017**, proposes a "squeeze-and-excitation" (SE) unit by taking channel relationship into account.
- It learns to recalibrate channel-wise feature maps by explicitly modeling the interdependencies among channels, which is consequently exploited to enhance informative channels and suppress other useless channels.
- Despite the high classification performance of the aforementioned CNN models, appropriately designing the optimal network structure often requires significant engineering work.
- NASNet [27] studies a paradigm to learn the optimal convolutional architecture based on training data. It adopts a neural architecture search (NAS) framework derived from reinforcement learning [28].
- In addition, it designs a new search space to enable network mapping from a proxy dataset (e.g. CIFAR-10) to ImageNet, and a regularization technique for generalization purpose. Offering less computational complexity than SENet, NASNet achieves the state-of-the-art accuracy on ImageNet

Computer Vision Aspects

Image classification

- Many real-life classification applications, such as robotics, autonomous driving, smartphone, etc., the classification task is highly constrained by the computational resources that are available.
- The problem thus becomes to pursue the optimal accuracy subject to a limited computational budget (i.e. memory and/or MFLOPs).

- Therefore, a set of lightweight networks such as **SqueezeNet [29], MobileNet [30], ShuffleNet [31], ShiftNet [32] and FE-Net [33]** start a wave.

Computer Vision Aspects

Image classification

- Many real-life classification applications, such as robotics, autonom
  classification task is highly constrained by the computational resou
- The problem thus becomes to pursue the optimal accuracy subject t
  memory and/or MFLOPs).

- Therefore, a set of lightweight networks such as **SqueezeNet [29], MobileNet [30], ShuffleNet [31], ShiftNet [32] and FE-Net [33]** start a wave.

SqueezeNet substitutes most 3 x 3 filters by 1 1 filters and cuts down the numbers of input channels for 3 x 3 filters to reduce the network complexity.

To maximize the accuracy with a limited number of network parameters, it delays the down-sampling operation to avoid information loss in early layers.

SqueezeNet is 50 smaller than AlexNet. If combined with deep compression [34], it can even be reduced to be 510 smaller than AlexNet

Computer Vision Aspects

Video classification

- The great success of deep learning in image domain has stimulated a variety of techniques to learn robust feature representations for **video classification**, where the semantic contents such as human actions [38] or complex events [39] are automatically categorized.

- **Early works** often treat a video clip as a collection of frames. Video classification is implemented by aggregating frame-level CNN features by averaging or encoding [40]. Standard classifiers, such as SVM, are finally used for recognition [41,42].

- In contrast to the **frame-level classification methods**, there are a number of other approaches applying end-to-end CNN models to learn the hidden spatio-temporal patterns in video. For example, the C3D features [43] are derived from a deep 3-D convolutional network trained on the large-scale UCF101 dataset.

- A **two-stream** approach [44] is proposed to factorize the learning problem of video representation into spatial and temporal cues separately.

- Specifically, a spatial CNN is adopted to model the appearance information from RGB frames, while a temporal CNN is used to learn the motion information from the dense optical flow among adjacent frames

Computer Vision Aspects

Video classification

- Two-stream approach only depicts movements within a short time window and fails to consider the temporal order of different frames, several recurrent connection models for sequential data, including recurrent neural networks (RNNs) and long short-term memory (LSTM) models, are leveraged to model the temporal dynamics for videos.

- In [45], two two-layer LSTM networks are trained with features from the two-stream approach for action recognition.

- In [46], the LSTM model and CNN model are combined to jointly learn spatial-temporal cues for video classification.

- In [47,48], attention mechanism is introduced for convolutional LSTM models to discover relevant spatio-temporal volumes for video classification.

Computer Vision Aspects

Object Detection

- Object detection, which is to determine and locate the object instances either from a large number of predefined categories in natural images or for a given particular object (e.g., Donald Trump's face, the distorted area in an image, etc.), is another important and challenging task in computer vision.

- Object detection and image classification share a similar technical challenge: both of them must handle a large number of highly variable objects.

- However, object detection is more difficult than image classification, as it must identify the accurate localization of the object of interest.

Computer Vision Aspects

Object Detection

- The **image classification** describes the image, while **object detection** aims to detect the location of a set of target objects.

- The **detection** task consists of **two sub-tasks**, one is the category information and probability of the target, and it is also a classification task.

- The other is the specific location of the target by utilizing bounding boxes with labels, which is a positioning task.

Computer Vision Aspects

Object Detection

- Historically, most research efforts have focused on detecting a single category of given objects such as pedestrian [14,49] and face

[50] by designing a set of appropriate features (e.g. HOG [14,49], Harr-like [50], LBP [51], etc.).

- In these works, objects are detected by using a set of **predefined feature templates** matching with each location in the image or

feature pyramids. Standard classifiers such as SVM [14,49] and Adaboost [50] are often used for this purpose.

- In order to build a general-purpose, robust object detection system, research community has started to develop large-scale, multi-

class datasets in recent years. **Pascal-VOC 2007 [52] with 20 classes** and **MS-COCO [53] with 80 object categories are two iconic**

object detection datasets.

In these two datasets, detection results are evaluated by two possible metrics: (i) Average Precision (AP) by counting the correctly detected bounding boxes for which the overlap ratio exceeds 0.5, and (ii) mean Average Precision (mAP) by averaging the AP values associated with different thresholds of the overlap ratio.

Computer Vision Aspects

Object Detection

- The current mainstream methods are mainly divided into a **one stage approach** (e.g., SSD, YOLO) and a **two-stage approach** (e.g., RCNN series).
- The **two-stage** approach firstly generates a sparse set of the bounding box from the image. It then makes corrections based on the bounding box region to improve the final detection results.
- The **single-stage** approach directly calculates the image and generates detection results.
- **The single-stage detection speed is faster, but the detection accuracy is lower. In contrast, the two-stage approach is completely the opposite.**

Computer Vision Aspects

Object Detection

Single – Stage

- YOLO Redmon et al. (2016) proposed YOLO that frames object detection problem as a **regression problem** instead of classification.

- YOLO can achieve **45 frames per second**, and the fast version has higher efficiency. That is, 155 frames per second doubles mAP (mean of Average Precision) compare with other real-time systems.

- Note that YOLO still lags in critical detection systems in terms of accuracy.

Computer Vision Aspects

Object Detection

Single – Stage

- Redmon and Farhadi (2017) introduced YOLO9000 (also known as YOLOv2), which made various improvements on YOLO and can detect over 9000 object categories.

- Compare with YOLO, YOLOv2 **made the following changes**, including batch normalization, use high resolution training images, dimension cluster, and convolutional with anchor box, which means predicting offsets instead of coordinates of bounding boxes.

- YOLOv3 is later proposed by Redmon and Farhadi (2018). The backbone of YOLOv3 has evolved from **Darknet-19** in YOLOv2 to **Darknet-53**, which deepens the number of network layers and introducing the cross-layer add operation in ResNet

Computer Vision Aspects

Object Detection

Single – Stage

- Bochkovskiy et al. (2020) improved YOLOv4 significantly, such as weighted-residual-connection (WRC), cross-stage-partial-connections (SCP), cross mini-batch Normalization (CmBN), self-adversarial- training (SAT), and mish-activation.

- The result of YOLOv4 is 43.5% AP (Average Precision) (65.7% AP50) for the MS COCO dataset at a real-time speed of ~65 FPS on Tesla V100.

- PP-YOLO (Long et al., 2020) have got improvements based onYOLOv3 after YOLOv4, which uses **Resnet50-V** instead of **Darknet53** as a backbone.

Computer Vision Aspects

Object Detection

Single – Stage **Single Shot Multi Box Detector (SSD)**

- Liu et al. (2016) presented a method that eliminates the process of generating bounding boxes.

- Their method first processes **six feature maps**. Each of the anchor boxes on each feature map generates a different length of the anchor boxes on the original input. Therefore, it can function feature maps from different resolutions to handle the various size of objects.

- The detection speed is up to **59 FPS** when the input size is **300 × 300**. Changing the input size to **512 × 512** achieves **76.9% mAP** on **VOC 2007** dataset, which outperforms the critical detection algorithm, a faster R-CNN.

Computer Vision Aspects

Object Detection

Single – Stage **RetinaNet**

- Lin, Goyal et al. (2017) thought that the low accuracy of the onestage approach was caused by the class imbalance and propose a new structure, RetinaNet, using Focal Loss. RetinaNet used ResNet and Feature Pyramid Network (FPN) as the backbone.

- It used single-level target recognition with focal loss, which can apply a modulating term to the cross-entropy loss. This is for focusing learning on hard examples and down-weighting the numerous easy negatives.

- This structure reaches 39.1 mAP higher than 36.2 mAP that Faster R-CNN on FPN (Lin, Dollar et al., 2017) got based on the challenging COCO datasets.

Computer Vision Aspects

Object Detection

- R-CNN [54] was the first two-stage method among the earliest CNN-based generic object detection techniques.

- It adopts AlexNet to extract a fixed-length feature vector from each resized region proposal, which is the object candidate generated by selective search algorithm [55]. Each region is then classified by a set of category-specific linear SVMs.

- The method **shows significant improvement** in mAP over the traditional state-of-the-art DPM detector [49]. It is, however, not elegant and inefficient, due to its multistage complex pipeline and the redundant CNN feature extraction from numerous region proposals.

- To address the aforementioned challenge associated with computational complexity, Faster R-CNN [58] further proposes a Region Proposal Network

(RPN). It then integrates both RPN (for proposal generation) and Fast-RCNN (for region classification) into a unified, end-to-end network

structure.

RPN and Fast-RCNN share most of the convolution layers, and the features from the last shared layer are used for two separate tasks (i.e. proposal generation and region classification).

With this highly efficient architecture, Faster R-CNNachieves 6 FPS inference speed on a GPU and the state-of-the-art detection accuracy on Pascal-VOC 2007.

National University
of Computer & Emerging Sciences

Thank you
and
Question and Answer