

# MEMAHAMI ETIKA DAN BIAS UNTUK MEMBANGUN TRUSTWORTHY AI

**Bambang Riyanto Trilaksono**

Sekolah Teknik Elektro dan Informatika-ITB

Center for Artificial Intelligence-ITB

Riset.ai



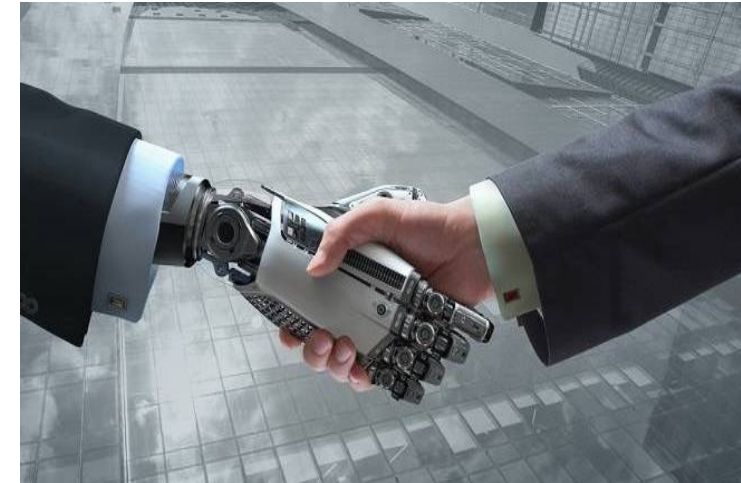
# Human vs AI

## Human Intelligence

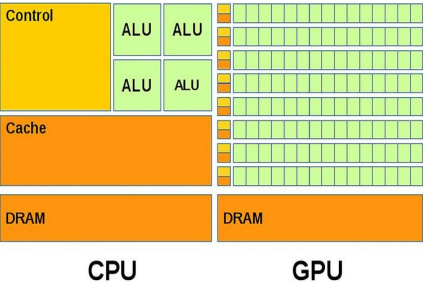
- Intuition, common sense, judgement, creativity, beliefs
- The ability to demonstrate their intelligence by communicating effectively
- Plausible reasoning, critical thinking

## Artificial Intelligence

- Ability to simulate human behavior and cognitive processes
- Capture and preserve human expertise
- Fast response –the ability to comprehend large amounts of data quickly



```
1 (function main)
2 {
3   main()
4 }
5
6 (script src="https://cdn.jsdelivr.net/npm/@tensorflow/tfjs@1.0.0")
7
8 (body)
9   @tensorflow/tfjs@1.0.0
10  (script)
11    main()
12    main()
13    main()
14    main()
15    main()
16    main()
17    main()
18    main()
19    main()
20    main()
21    main()
22    main()
23    main()
24    main()
25    main()
26    main()
27    main()
28    main()
29    main()
30    main()
31    main()
32    main()
33    main()
34    main()
35    main()
36    main()
37    main()
38    main()
39    main()
40    main()
41    main()
42    main()
43    main()
44    main()
45    main()
46    main()
47    main()
48    main()
49    main()
50    main()
51    main()
52    main()
53    main()
54    main()
55    main()
56    main()
57    main()
58    main()
59    main()
60    main()
61    main()
62    main()
63    main()
64    main()
65    main()
66    main()
67    main()
68    main()
69    main()
70    main()
71    main()
72    main()
73    main()
74    main()
75    main()
76    main()
77    main()
78    main()
79    main()
80    main()
81    main()
82    main()
83    main()
84    main()
85    main()
86    main()
87    main()
88    main()
89    main()
90    main()
91    main()
92    main()
93    main()
94    main()
95    main()
96    main()
97    main()
98    main()
99    main()
100   main()
101 }
```



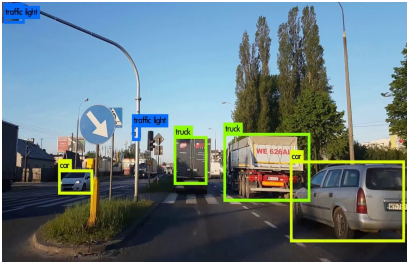
Algoritma makin baik

CPU & GPU makin kuat

Data makin berlimpah

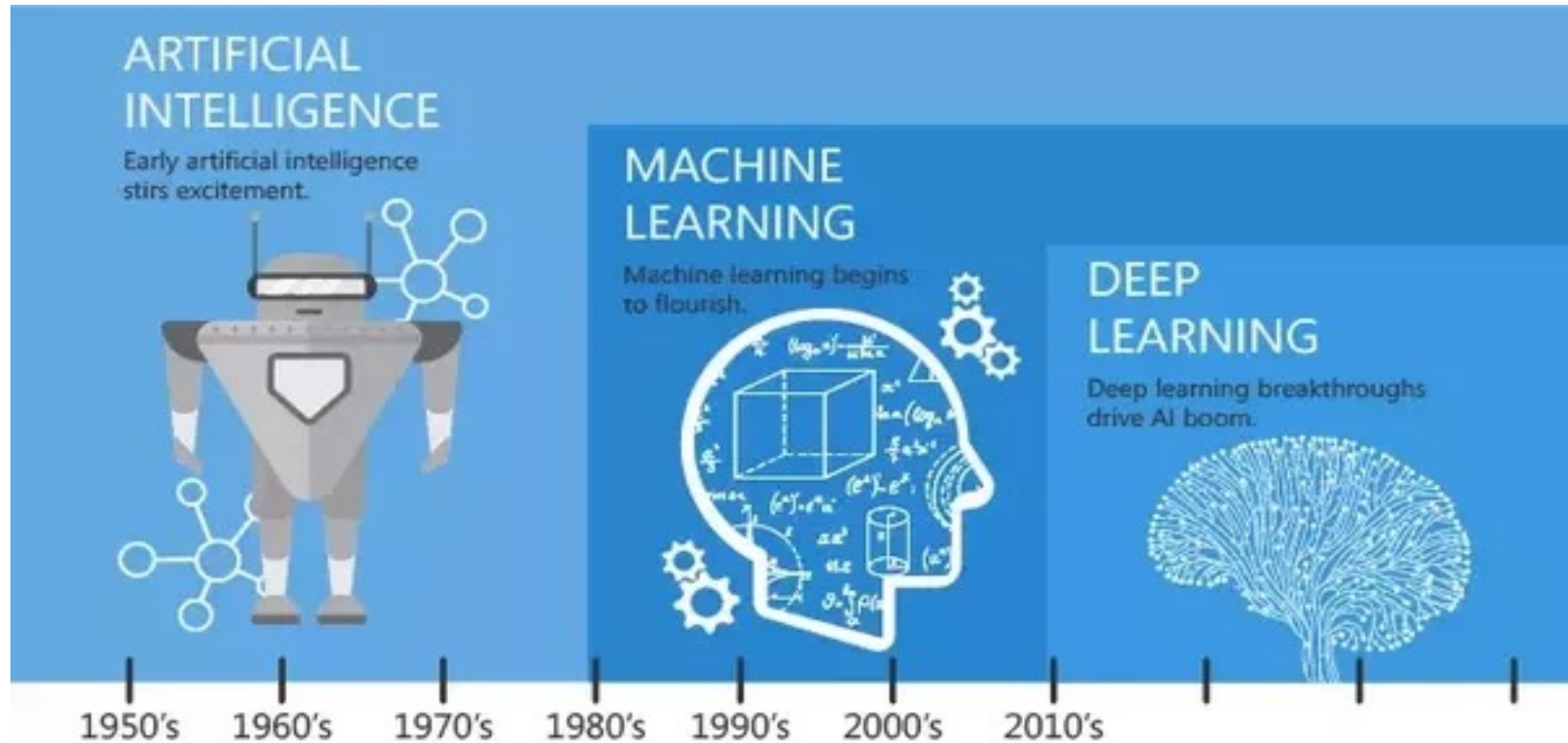
Aplikasi makin beragam

Perkembangan AI

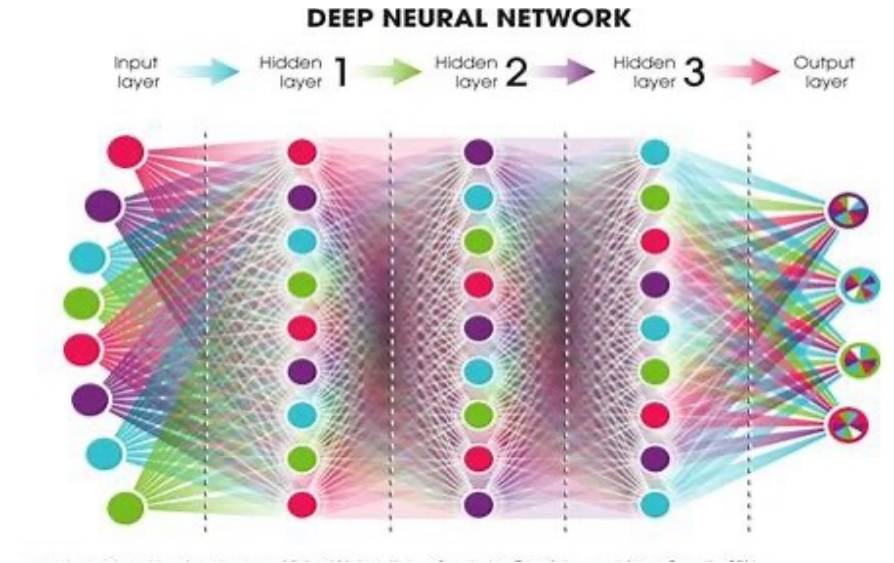
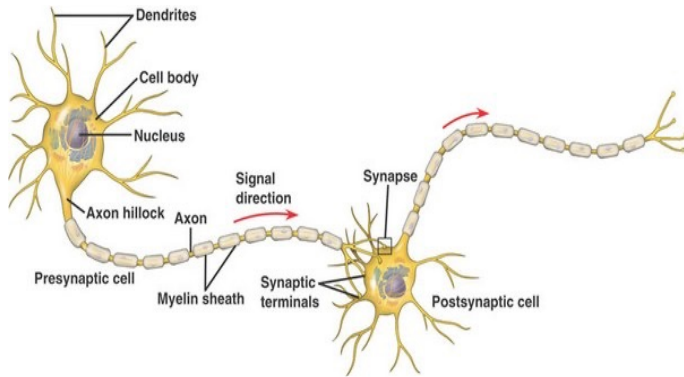
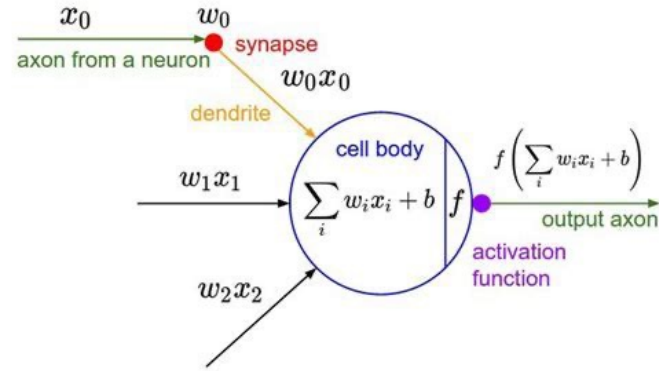


Mengapa Teknologi AI Semakin Berkembang?

# AI, Machine Learning, Deep Learning



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.



# Neuron & Artificial Neural Networks Model

# Convolutional Neural Networks



Convolutional networks + FC networks



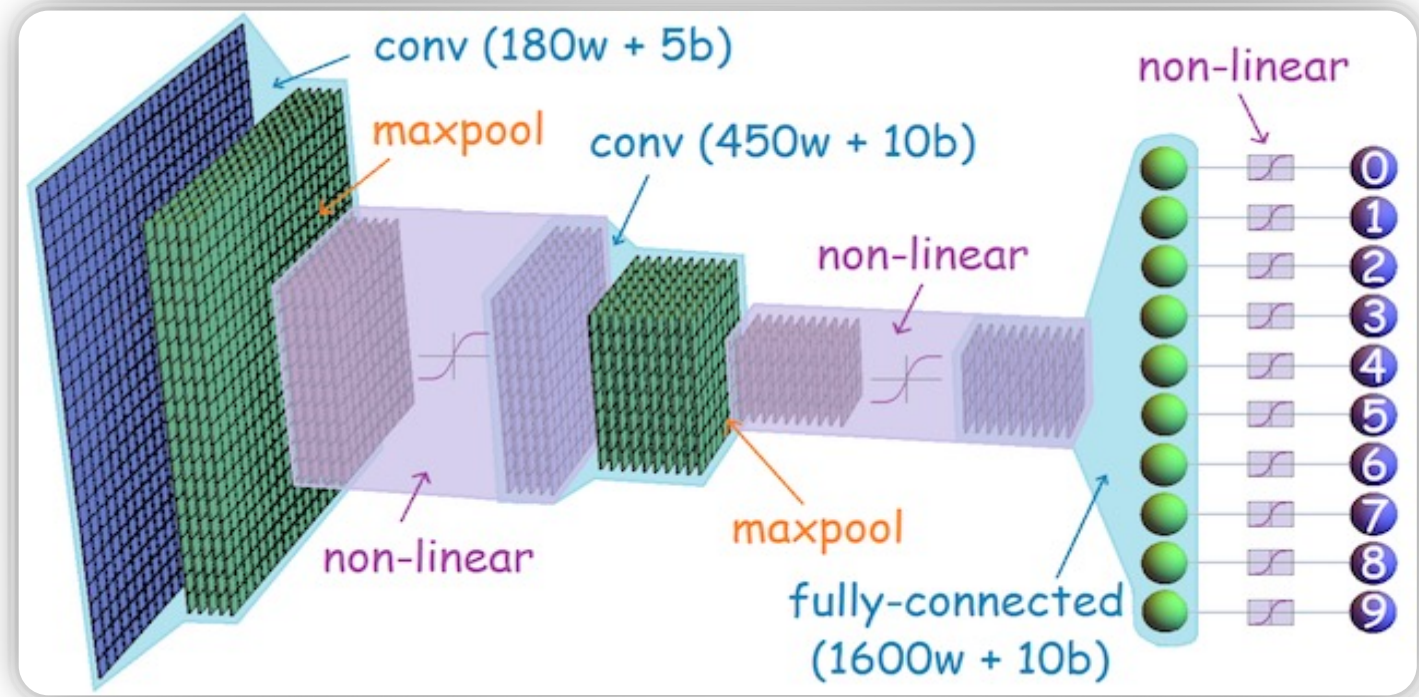
Mengklasifikasi image



Ekstraksi fitur secara otomatis

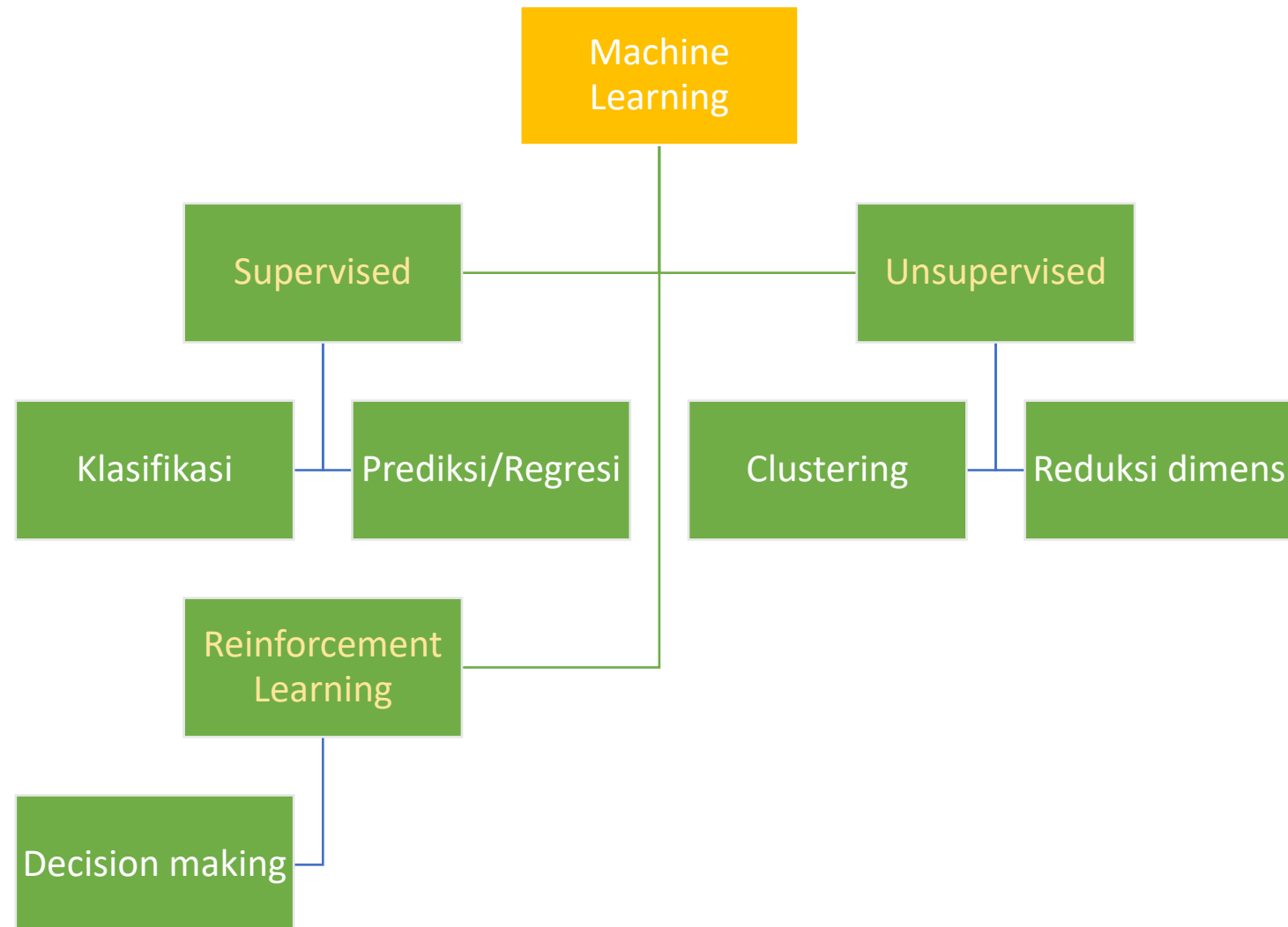


Forward Computation & Backpropagation

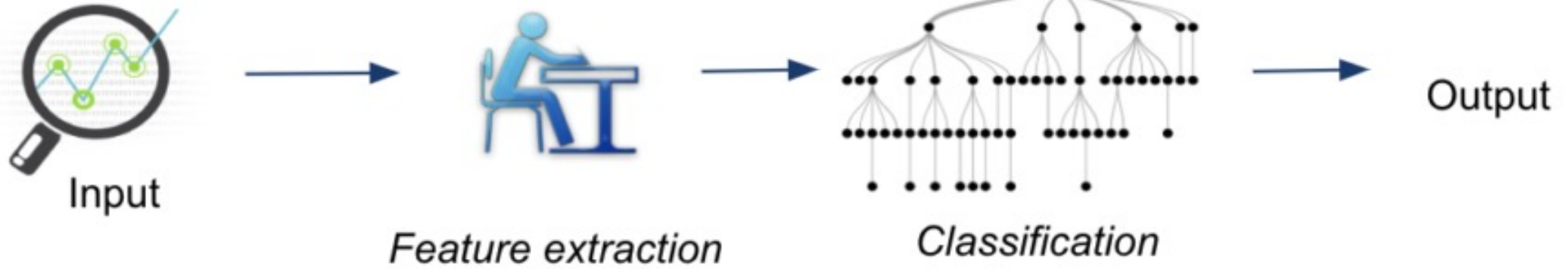


Deep Learning Architecture

# Machine Learning & Deep Learning Task

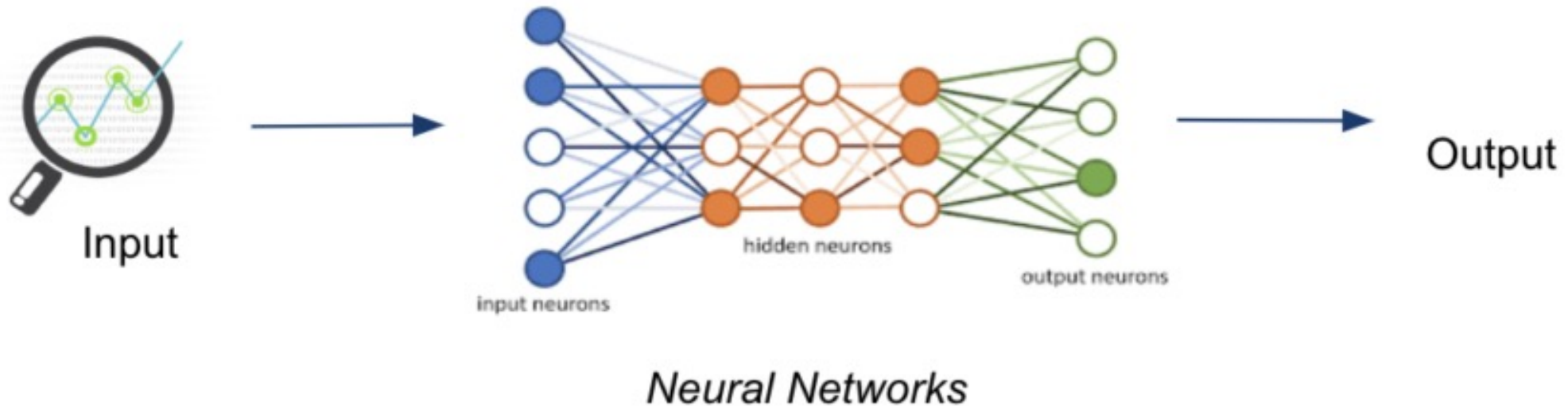


## Machine Learning



Traditional machine learning uses hand-crafted features, which is tedious and costly to develop.

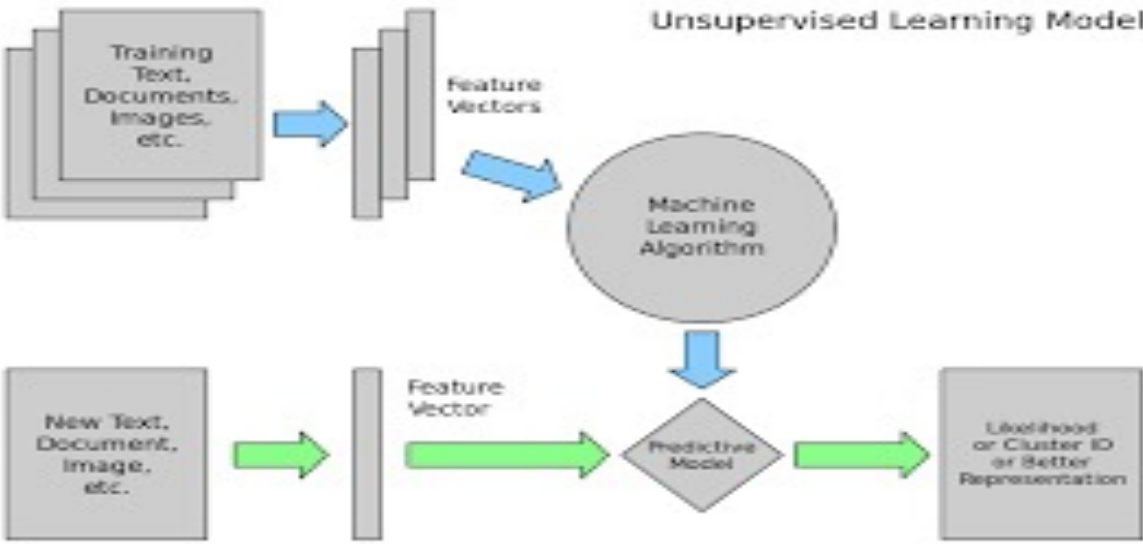
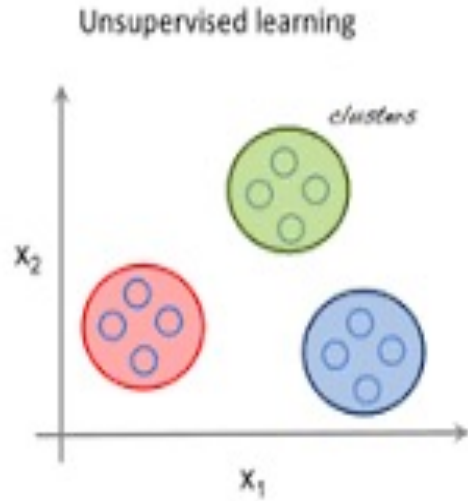
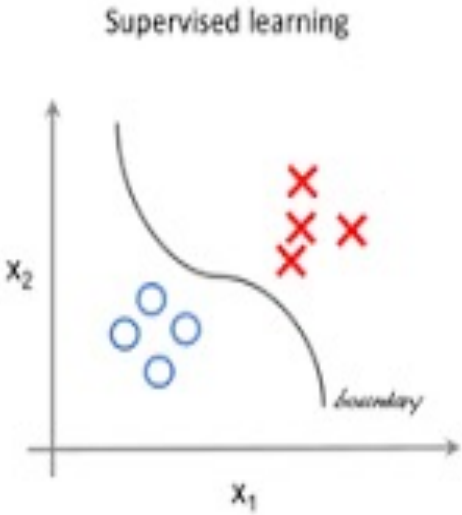
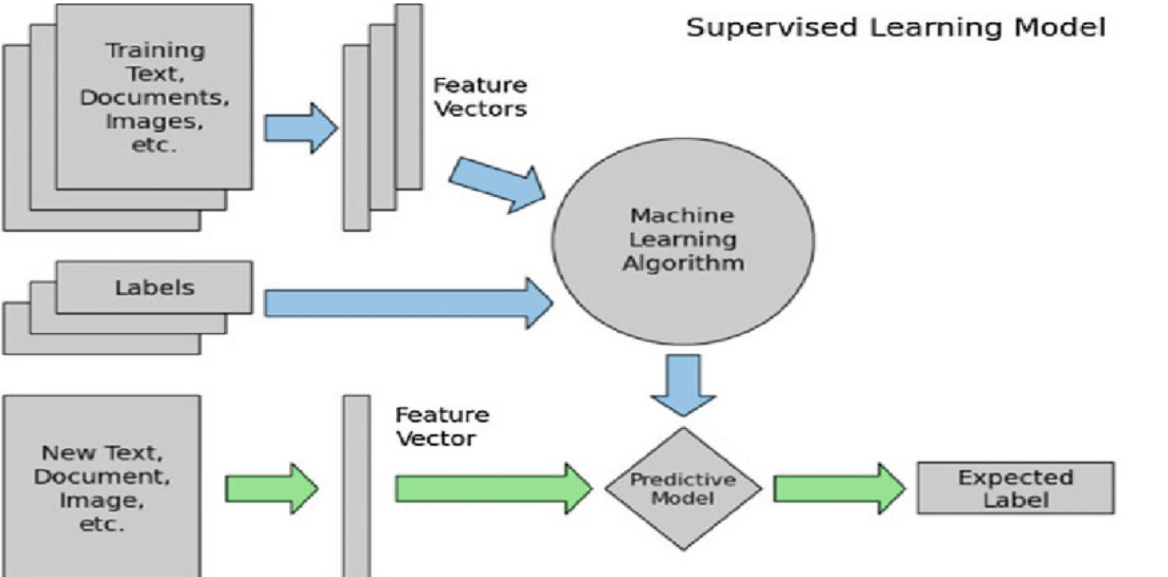
## Deep Learning



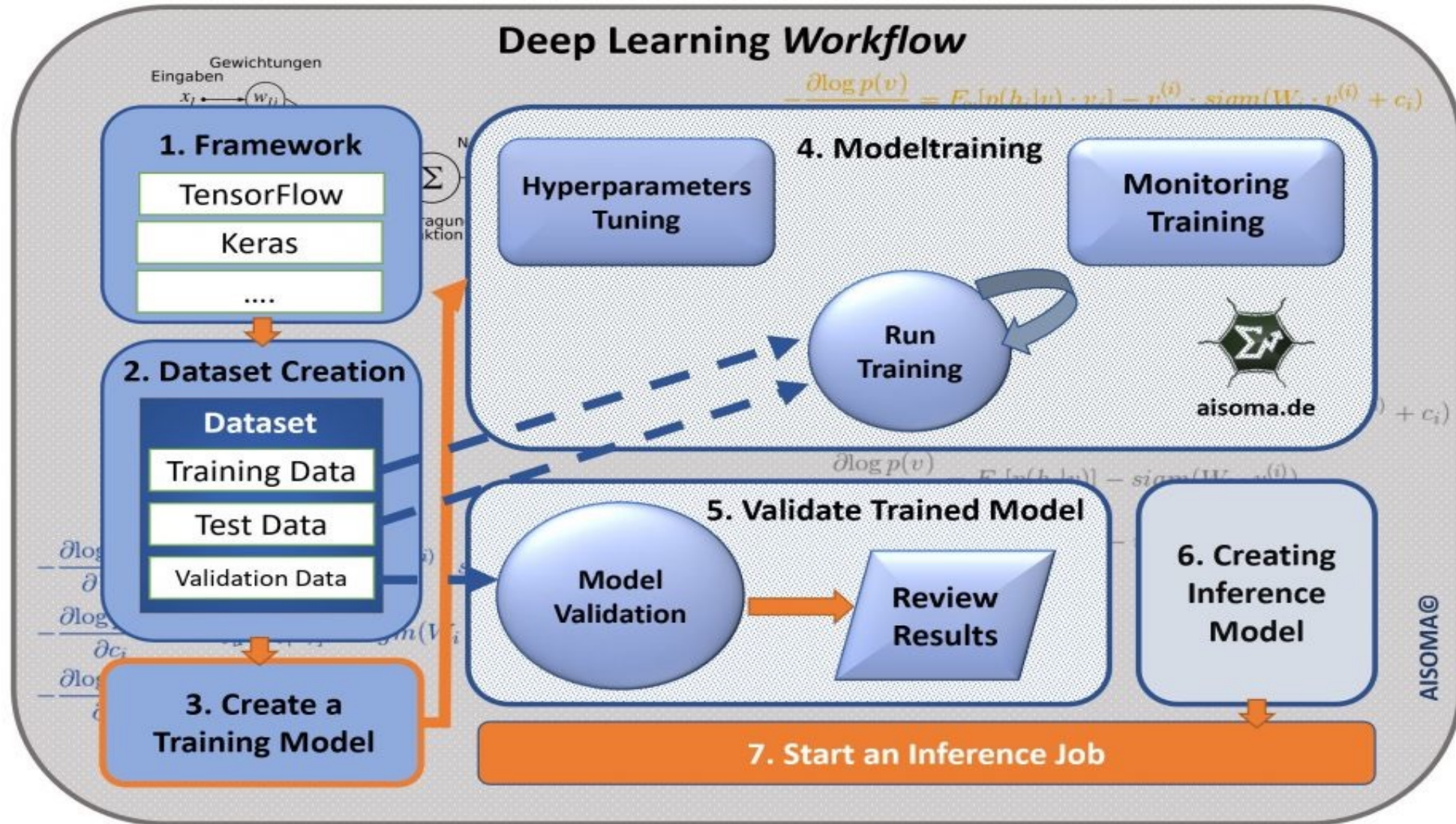
Deep learning learns hierarchical representation from the data itself, and scales with more data.



# Machine Learning Workflow



# Deep Learning Workflow




AISOMA©

# AI APPLICATIONS

Image Classification    Object Detection

**COMPUTER VISION**



This panel illustrates computer vision applications. It features icons for image classification (cat, dog, leaf) and object detection (car, person). The central image shows a street scene with green bounding boxes around pedestrians. Below this, there are images of an iris being scanned and two people walking in a greenhouse.

Voice Recognition    Language Translation

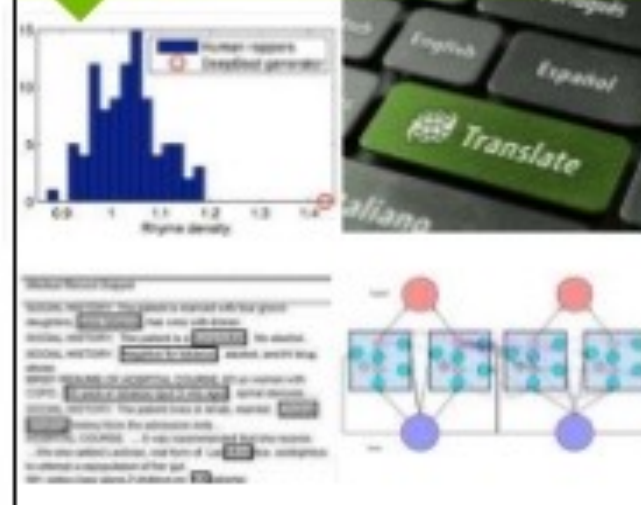
**SPEECH & AUDIO**



This panel illustrates speech and audio applications. It features icons for voice recognition (microphone) and language translation (waveform). The central image shows a person smiling while talking on a blue smartphone, with a spectrogram overlaid on the right side.

Recommendation Engines    Sentiment Analysis

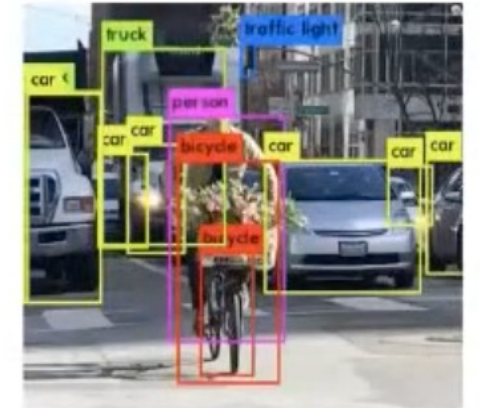
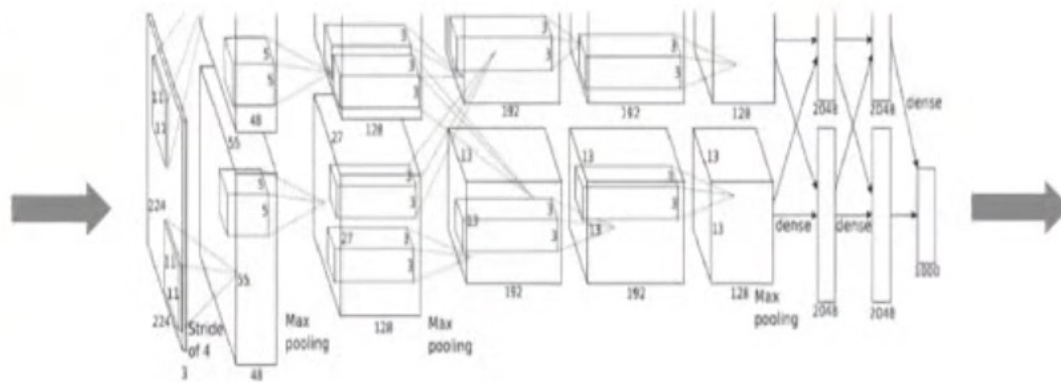
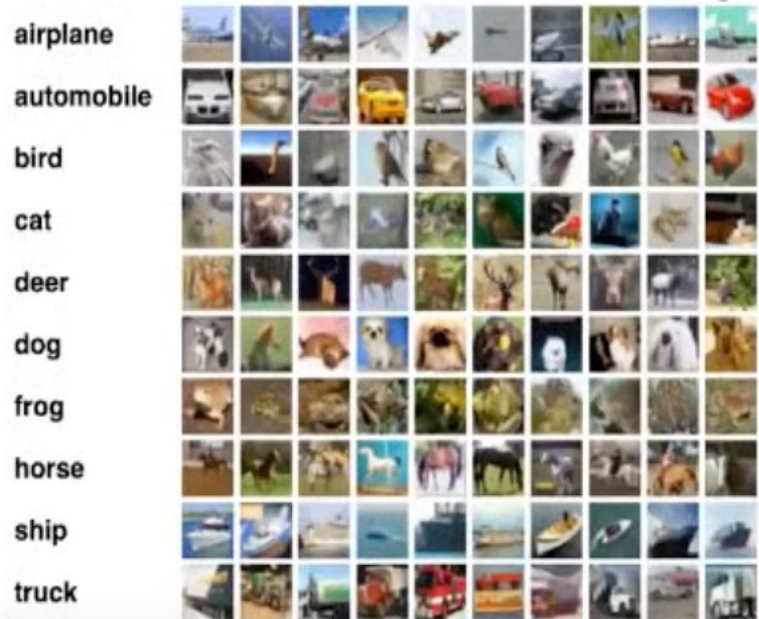
**NATURAL LANGUAGE PROCESSING**



This panel illustrates natural language processing applications. It features icons for recommendation engines (thumbs up, thumbs down) and sentiment analysis (smiley and frowny faces). The central image is a collage including a histogram of 'Movie density', a keyboard with a green 'Translate' key, a snippet of text with highlighted words, and a diagram of a neural network.

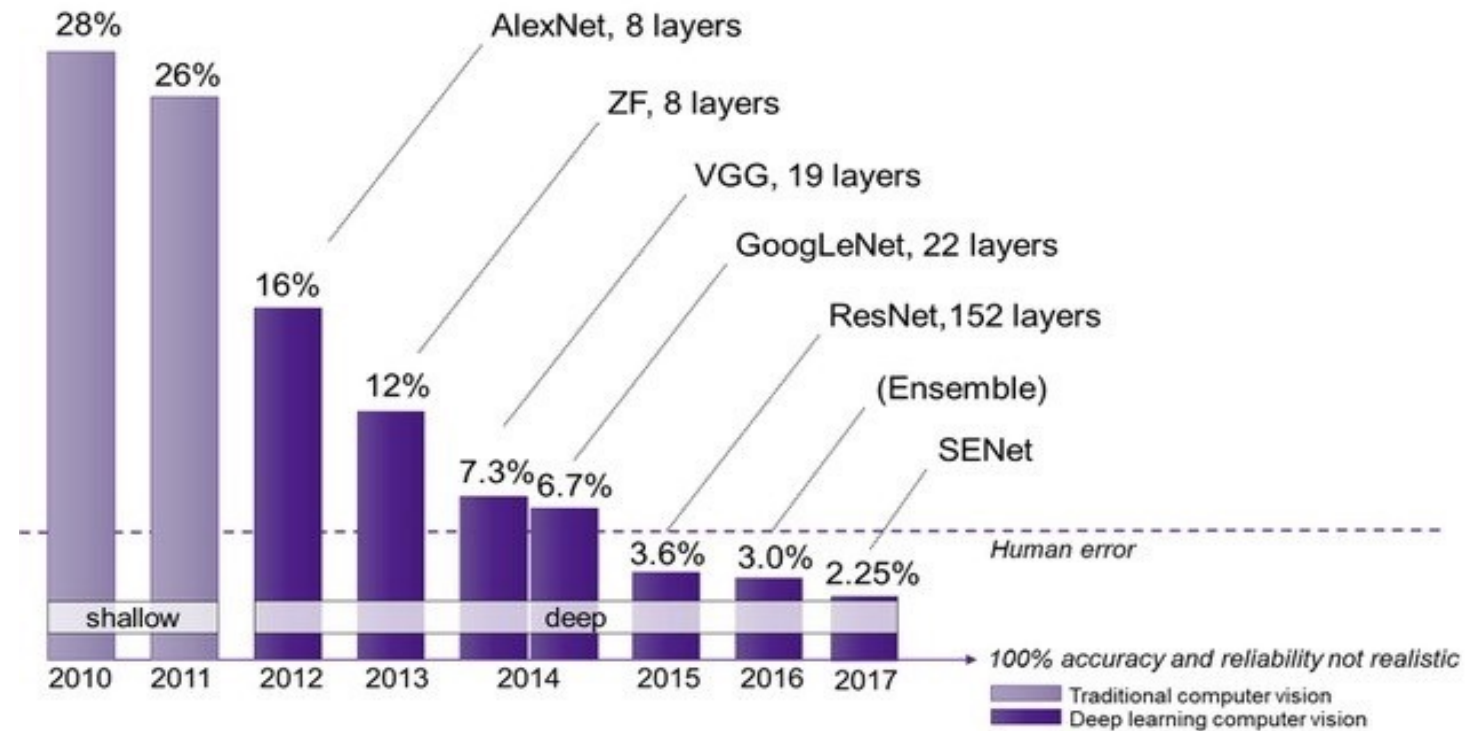
Big Data Analytics

# Supervised (Deep) Learning

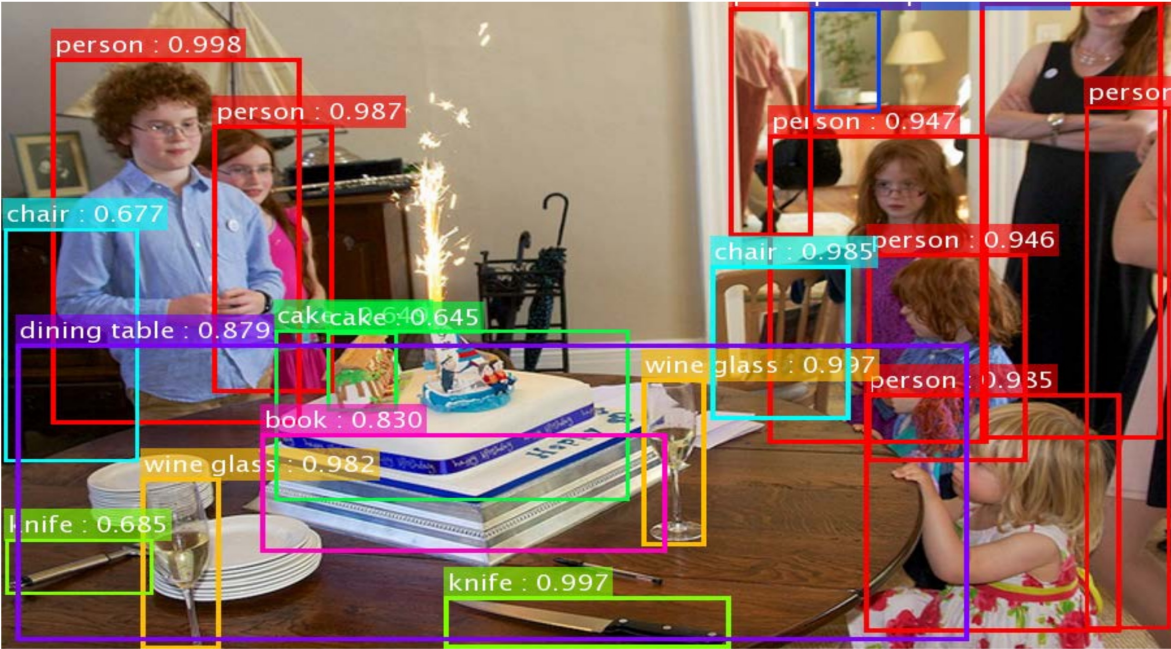
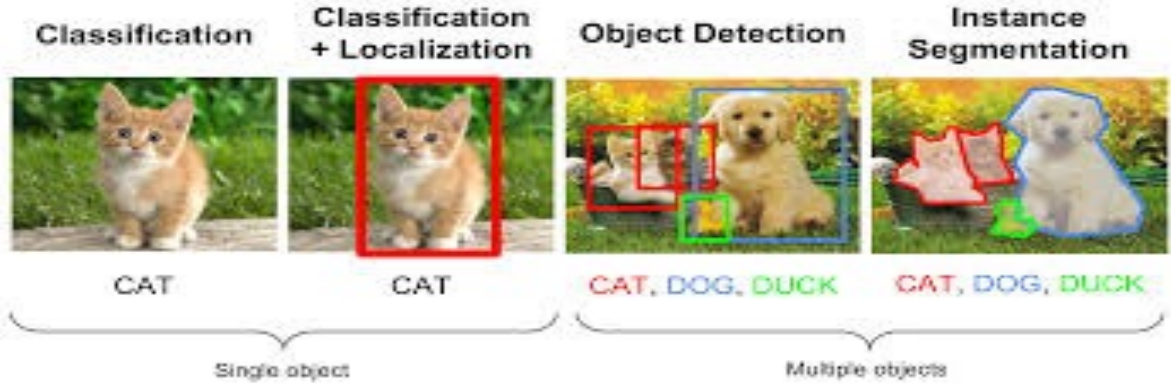


- Predefine the set of visual concepts to be learned
- Collect diverse and large number of examples for each of them
- Train a deep model for several GPU hours / days

# Kompetisi ImageNet



# Aplikasi AI untuk Computer Vision



Describes without errors	Describes with minor errors	Somewhat related to the image
 <p>A person riding a motorcycle on a dirt road.</p>	 <p>Two dogs play in the grass.</p>	 <p>A skateboarder does a trick on a ramp.</p>
 <p>A group of young people playing a game of frisbee.</p>	 <p>Two hockey players are fighting over the puck.</p>	 <p>A little girl in a pink hat is blowing bubbles.</p>

## AI & Etika

Bias

- Is AI fair?

Liability

- Who is responsible for AI?

Human interaction

- Will we stop talking to one another?

Employment

- Is AI getting rid of jobs?

Wealth

- Who benefits from AI?

Power & control

- Who decides how to deploy AI?

Robot rights

- Can AI suffer?

## Contoh Bias pada AI

### Amazon recruiting tool

- Aplikasi rekrutmen yg mereview resume applicant dan memberikan rating
- Bias thd gender (wanita)

### Healthcare risk prediction

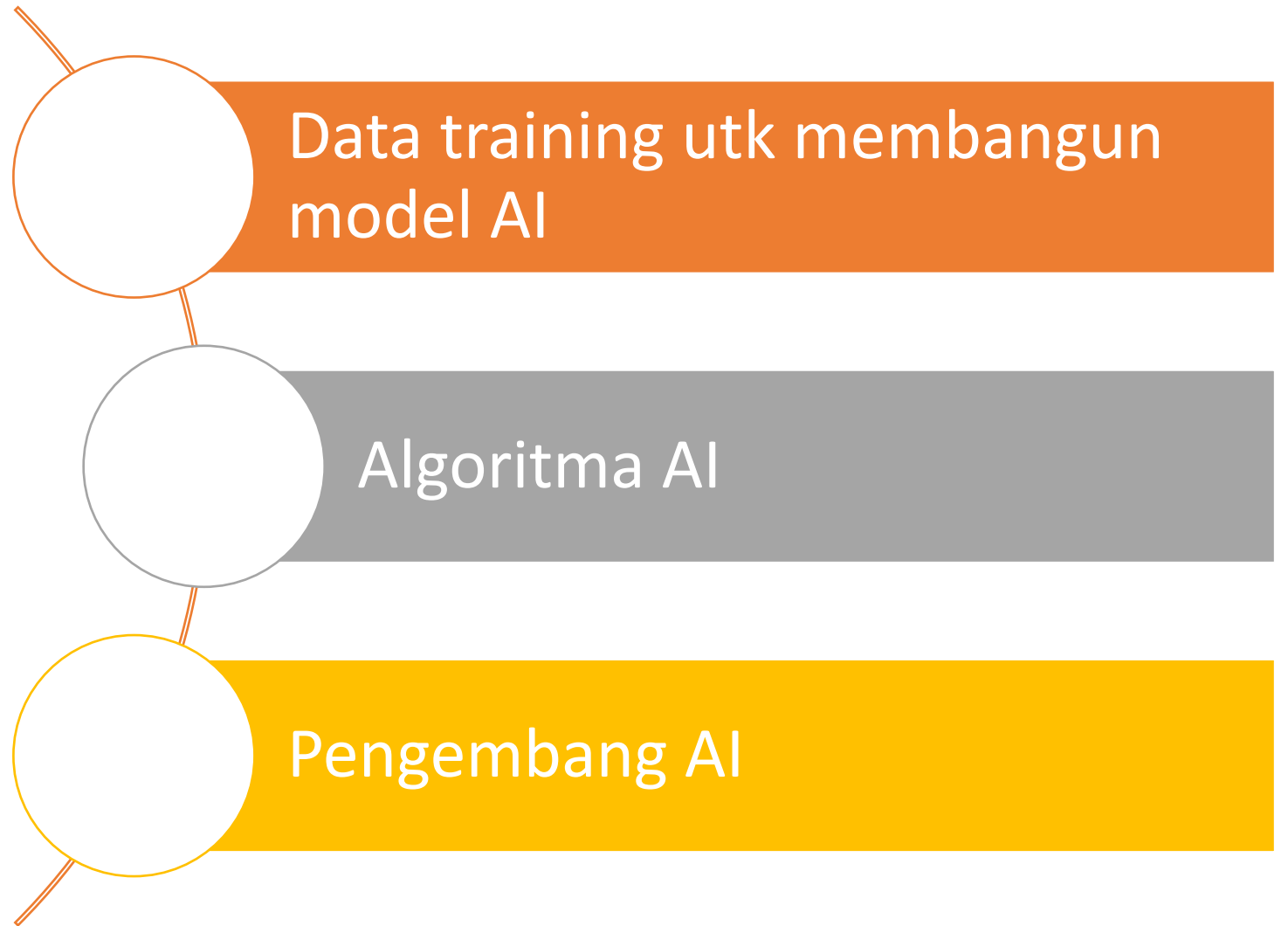
- Memprediksi pasien yg memerlukan layanan kesehatan ekstra
- Memprioritaskan pasien kulit putih dibandingkan kulit hitam

### Face recognition

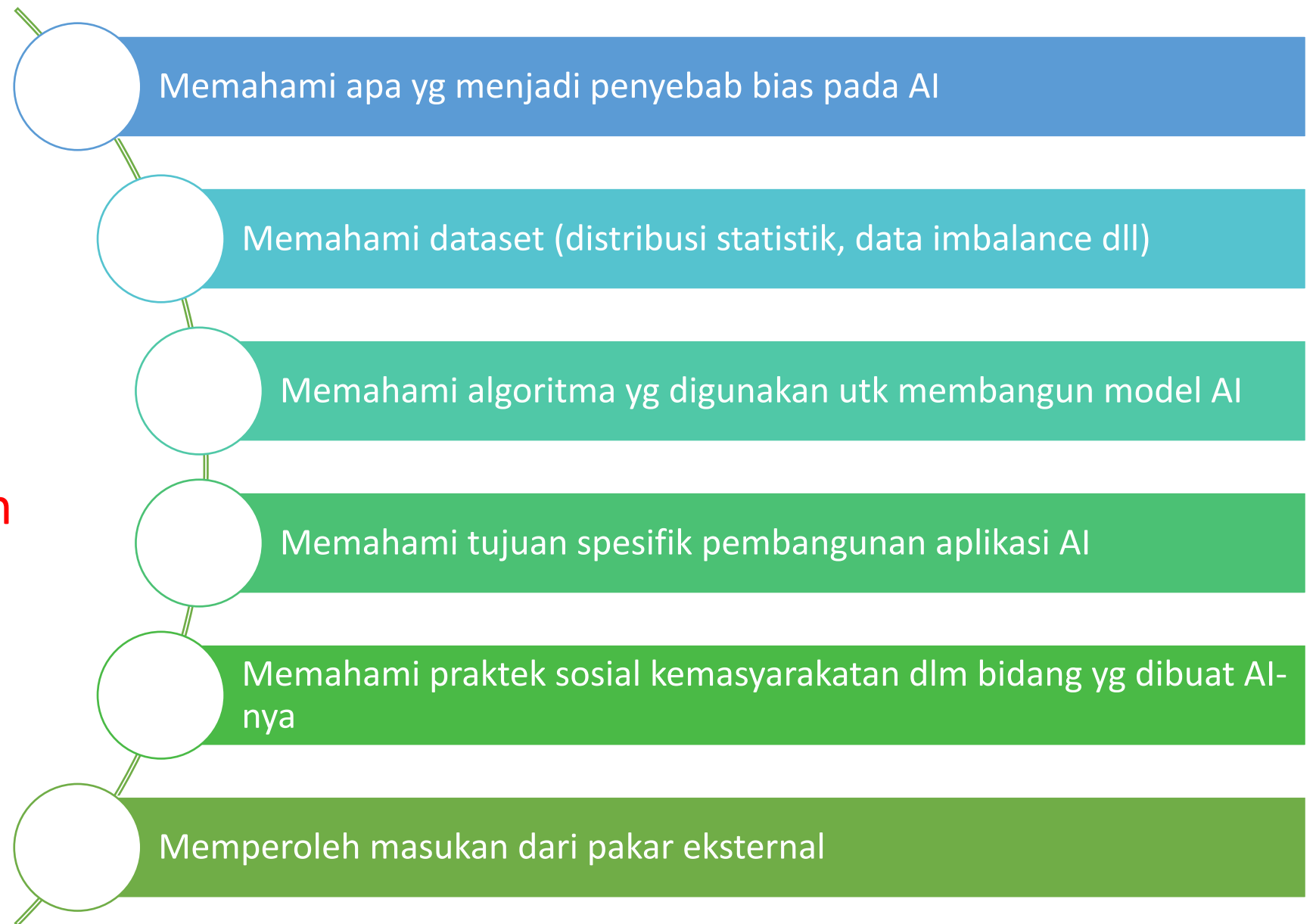
- Temuan ACLU & MIT : akurat utk pria kulit putih, tdk akurat pada wanita kulit hitam



Apa yg menjadi penyebab terjadinya bias pada AI?



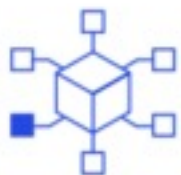
Bagaimana cara mengatasi atau memecahkan masalah bias pada AI?



# Minimizing bias will be critical if artificial intelligence is to reach its potential and increase people's trust in the systems.

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider

1



Be aware of contexts in which AI can help correct for bias and those in which there is high risk for AI to exacerbate bias

2



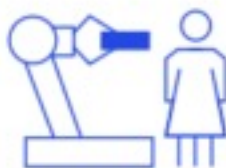
Establish processes and practices to test for and mitigate bias in AI systems

3



Engage in fact-based conversations about potential biases in human decisions

4



Fully explore how humans and machines can best work together

5



Invest more in bias research, make more data available for research (while respecting privacy), and adopt a multidisciplinary approach

6



Invest more in diversifying the AI field itself

Tools utk  
Mengurangi  
Bias

AI Fairness  
360 (IBM)

- Menguji bias pada data dan model
- Terbatas utk klasifikasi biner

IBM Watson  
OpenScale

- Memeriksa bias secara real-time ketika AI membuat prediksi

Google  
What-If Tool

- Menguji performance
- Menganalisis atribut data yg berbeda
- Visualisasi

# Trustworthy AI

**Robust** : tahan thd serangan (attack)

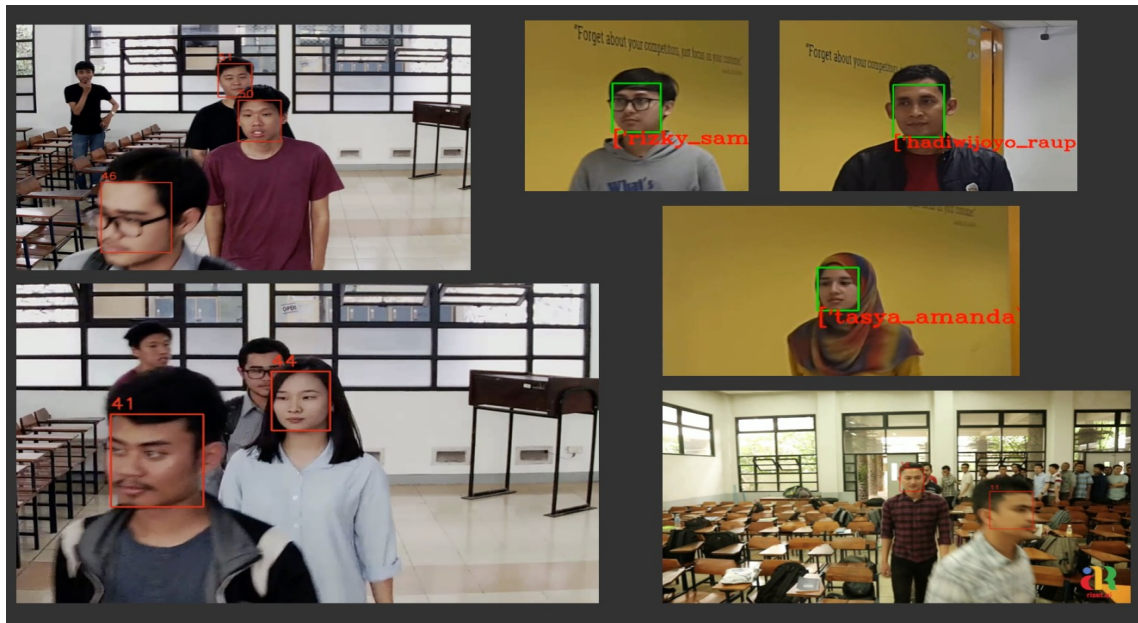
**Fair** : berkeadilan (tidak bias)

**Responsible** : Akuntabel

**Transparent** : dapat menjelaskan

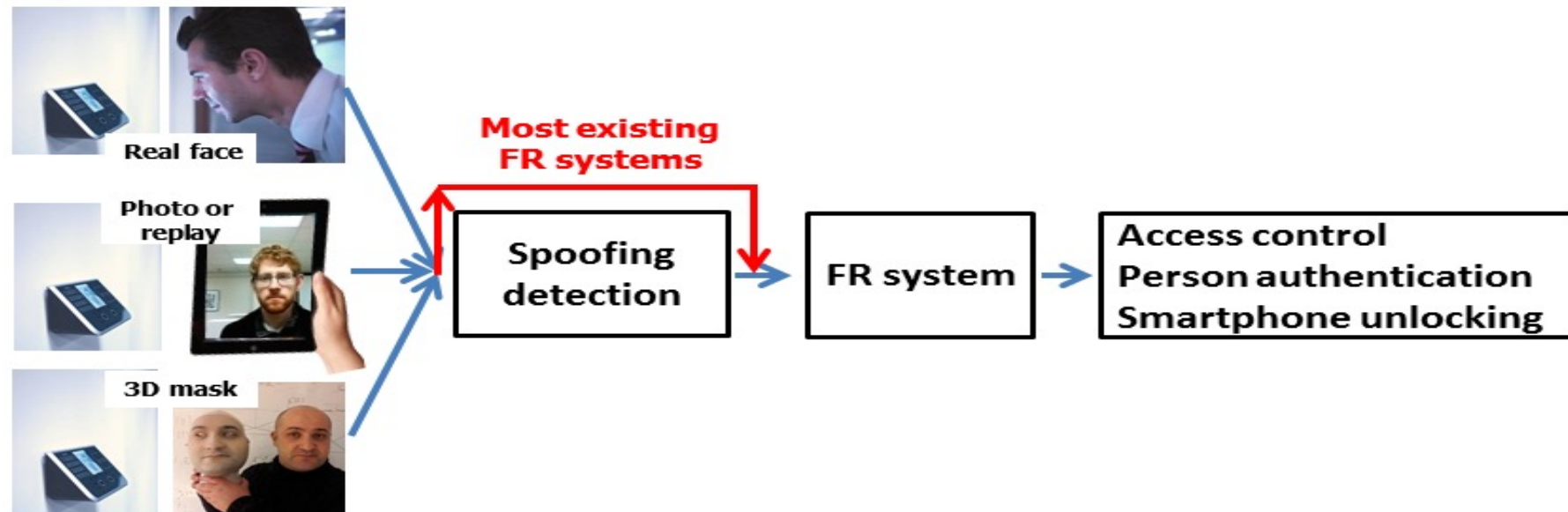
**Secure** : aman

**Privacy** : menjaga privasi



## Spooing--serangan pada pengenalan wajah

- Deteksi Liveness
- Aktif vs pasif



## The goal of explainable AI

### Today



Training  
Data



Learning  
Process



Learned  
Function



Output



User with  
a Task

### Tomorrow



Training  
Data



New Learning  
Process



Explainable  
Model



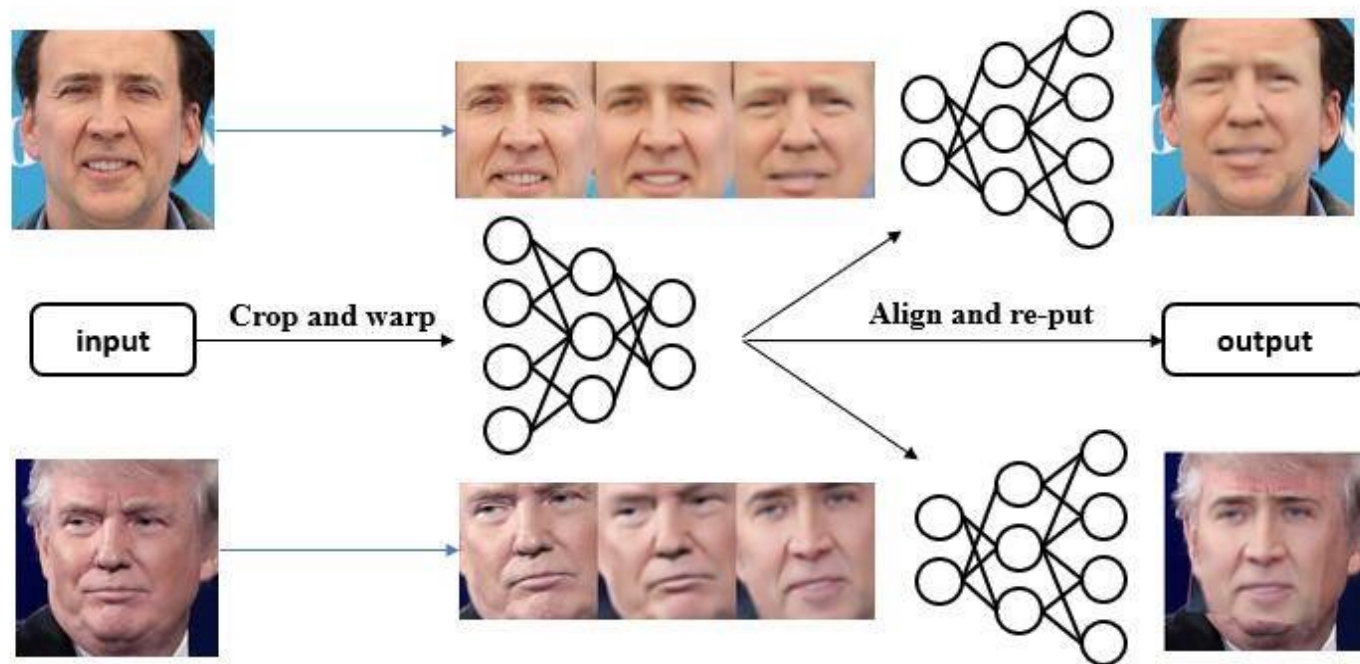
Explainable Interface



User with  
a Task

Explain-  
able AI

# Deepfake

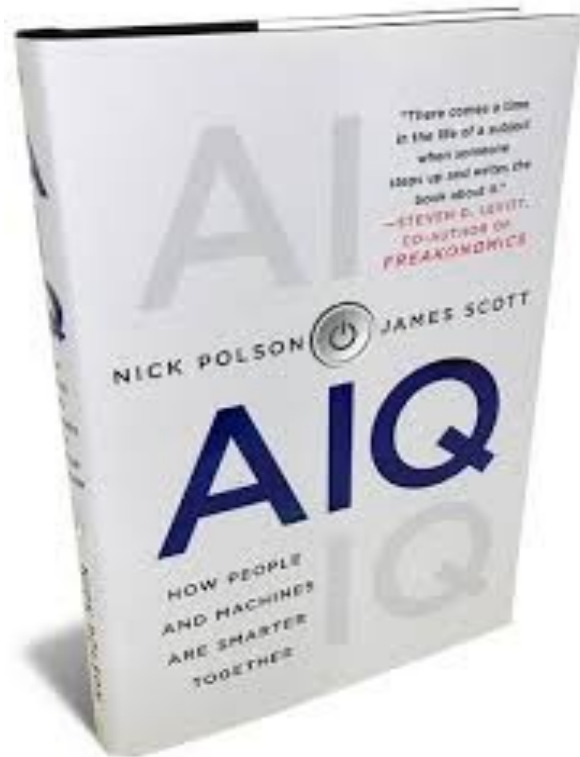


- Memiliki kemampuan utk membuat foto dan video (digabungkan dg suara) palsu yg seolah-olah benar/valid
- AI yg bertanggung-jawab?
- Microsoft Deepfake Detection Tool





Masih diperlukan riset yg mendalam, komprehensif dan multidisiplin untuk memahami **etika AI**, mengatasi **bias**, dan membangun **Trustworthy AI**



A revolution of intelligent machines, from self driving cars to smart digital assistants, is now remaking our world, just as the Industrial Revolution remade the world of the nineteenth century.

AI is not some science fiction droid from the future. It's right here, right now, and it's changing our lives at lightning-fast speed.

Artificial Intelligence Quotient : How People and Machines Are Smarter Together (Nick Polson & James Scott, St. Martins Press, 2018)

# TERIMAKASIH

