3000788 Intro to Comp Molec Biol

Lecture 23: Introduction to machine learning and AI

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

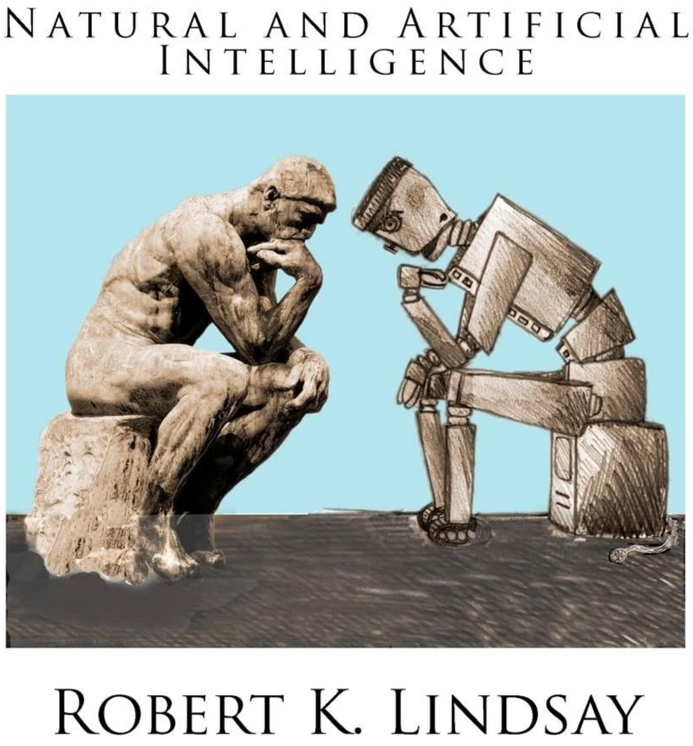
Today’s agenda

- What is machine learning and AI
- Evolution of AI: from rule-based to generative model

Natural vs artificial intelligence

- AI is a result of computer algorithms mimicking the natural learning process

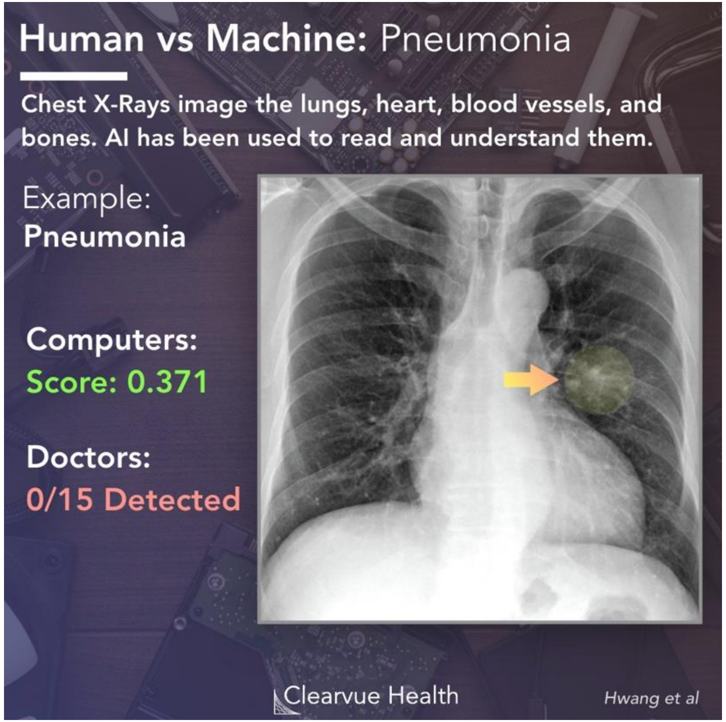
- **Signs of intelligence**
    - Memorization
    - Pattern recognition and generalization
    - Learning from trials and errors
    - Ability to create
    - Reasoning with cause and effect

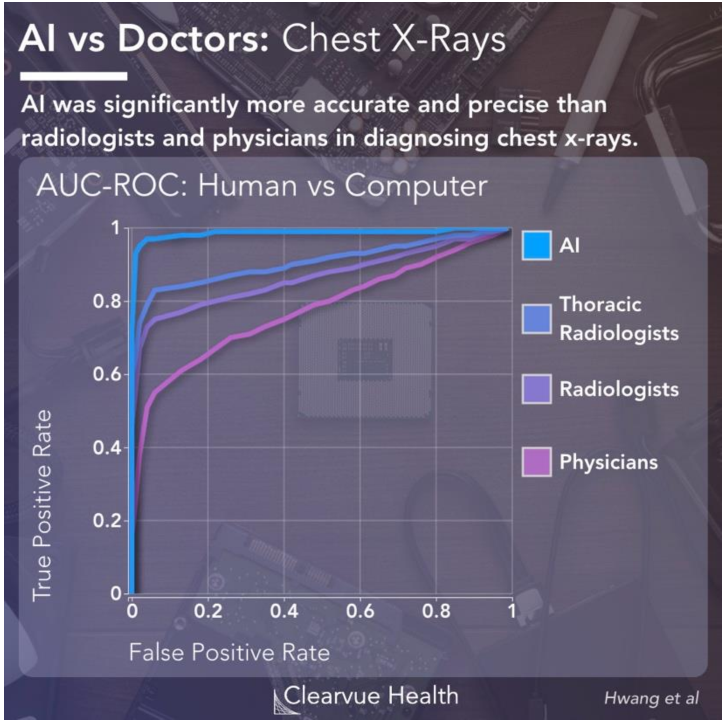


Memorization and information compression

https://www.youtube.com/watch?v=P18EdAKuC1U

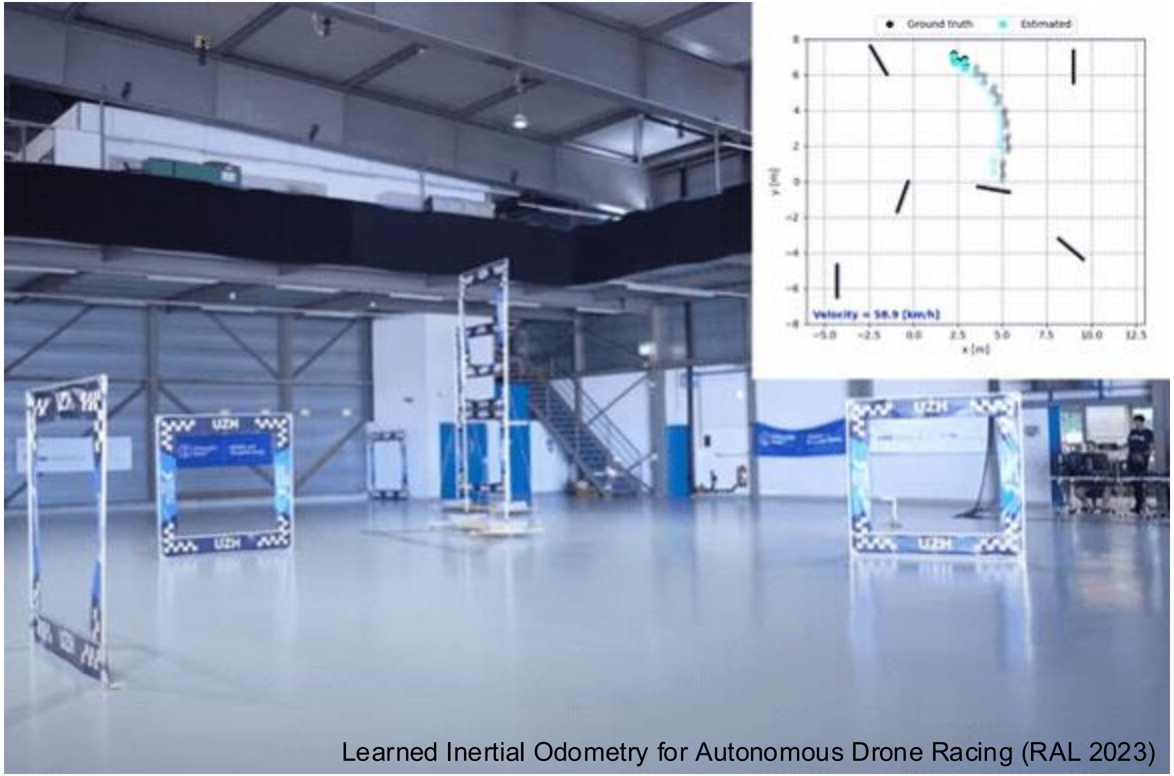
Pattern recognition





https://www.clearvuehealth.com/b/ai-radiology-xray/

Trial and error



Learned Inertial Odometry for Autonomous Drone Racing (RAL 2023)

Google DeepMind's AlphaGo computer beats top player Lee Sedol for third time to sweep competition

Machine learning (ML)

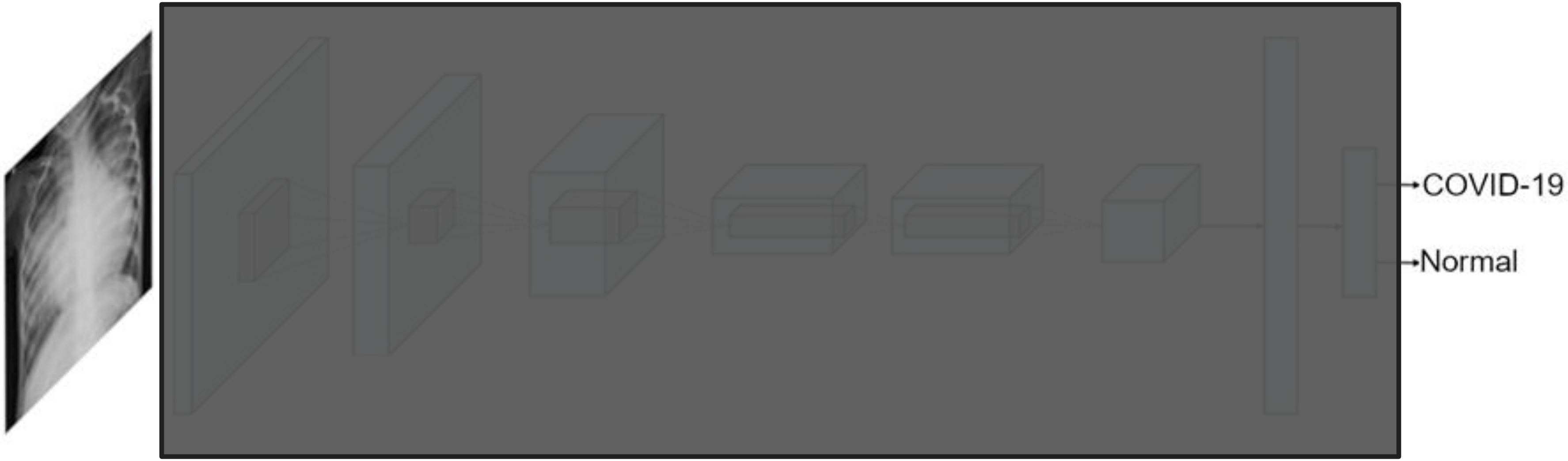
The engine behind modern AI

AI 1.0: Hand-crafted algorithms

https://www.cloudaccess.net/fresh/blog/entry/2018/12/07/wait-what-google-isn-t-the-only-search-engine-on-the-web.html

https://shapeshed.com/photoshop-101-the-magic-wand-tool/

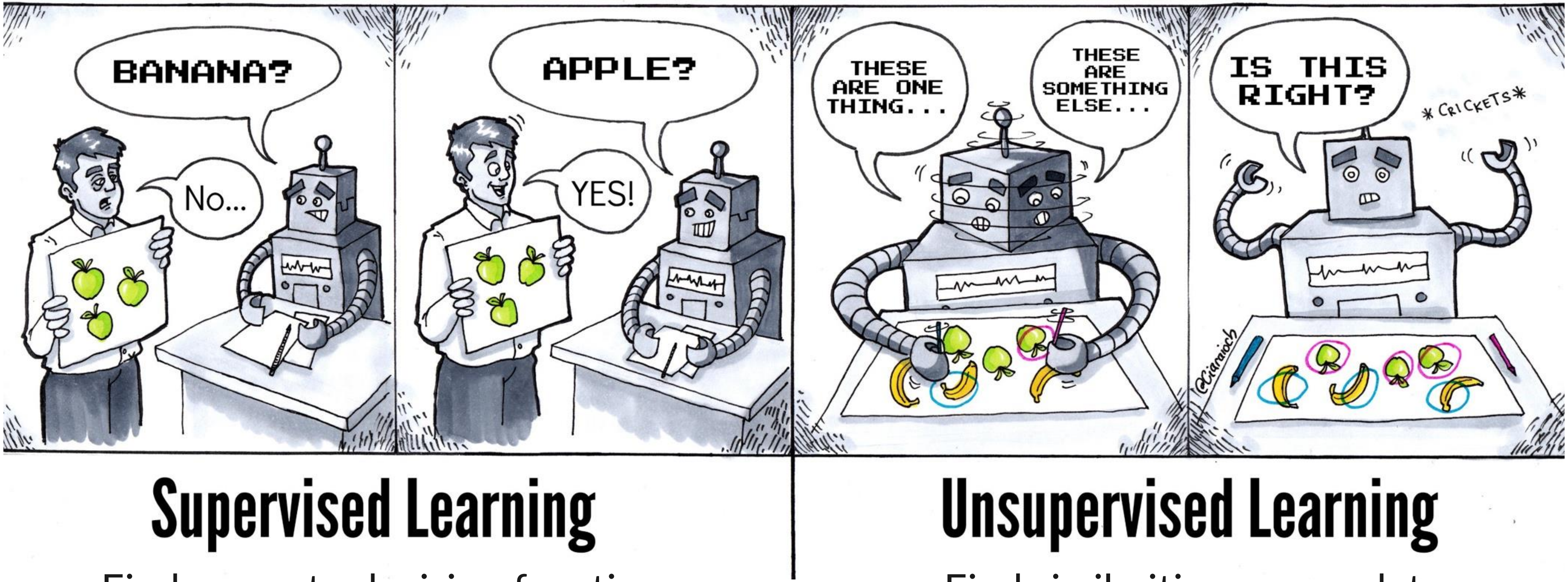
Making computer optimizes itself



Source: Cortes and Sanchez. IEEE Latin America Transaction (2021)

- The relationship is too complex for human to define
- Instead, human provides the data (*x*, *y*) and let the computer do the fitting

Machine learning paradigms



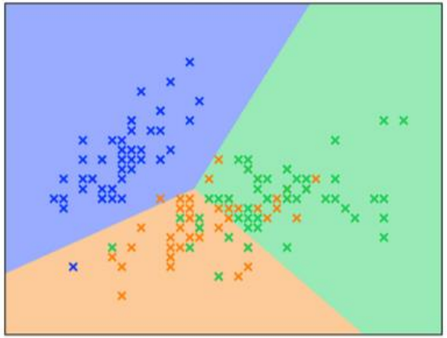
Find accurate decision functions

Find similarities among data

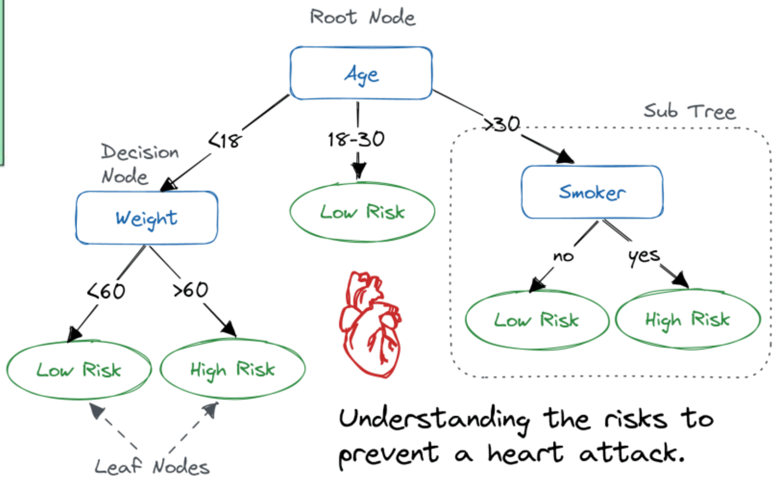
https://x.com/athena_schools/status/1063013435779223553/photo/1

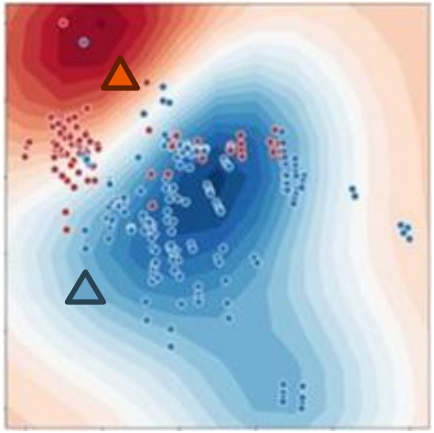
AI 2.0: Classical ML models

Linear: Weighted score = (input$_1$ x w$_1$) + … + (input$_n$ x w$_n$)



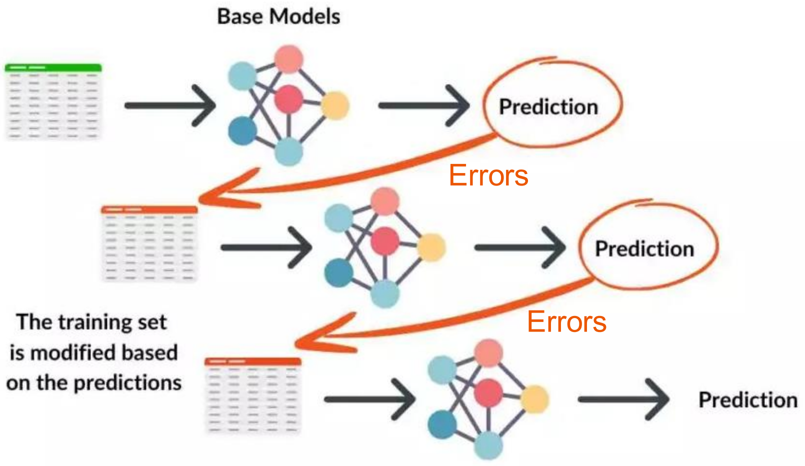
Tree: Collection of decisions





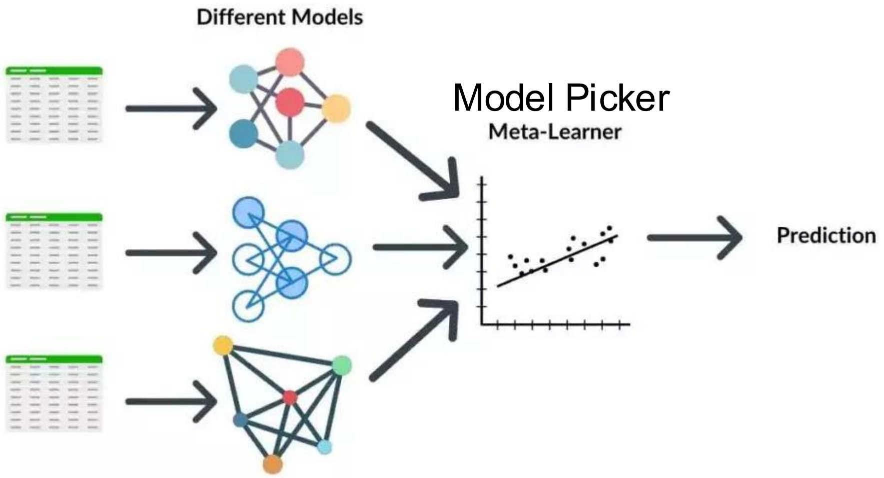
Nearest neighbors: Predict using similarity to past observations

Enhancement with ensemble approaches



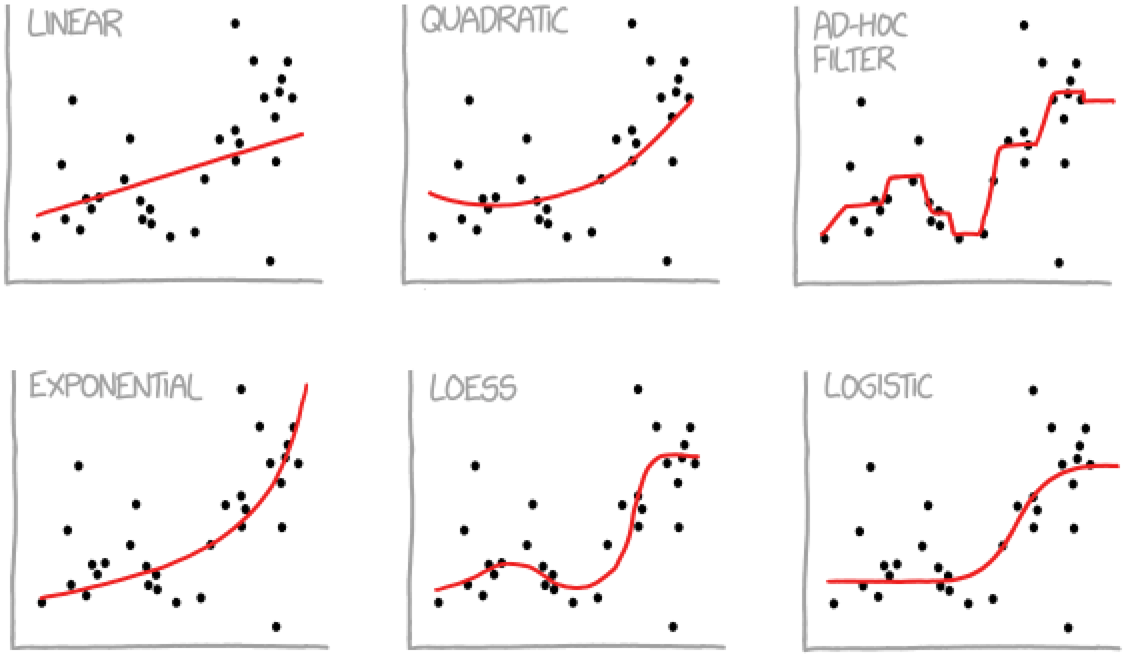
Boosting: Iterative improvement with additional models

Stacking: Combine multiple models with different capabilities



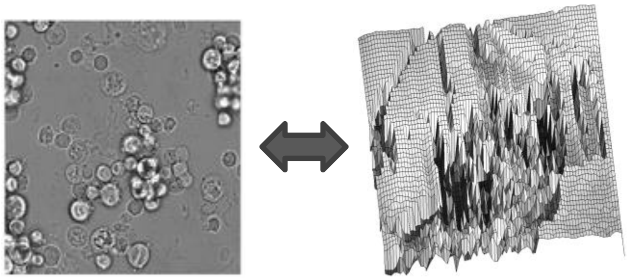
https://spotintelligence.com/2024/03/18/bagging-boosting-stacking/

Limitation of classical ML



https://www.xkcd.com/2048/

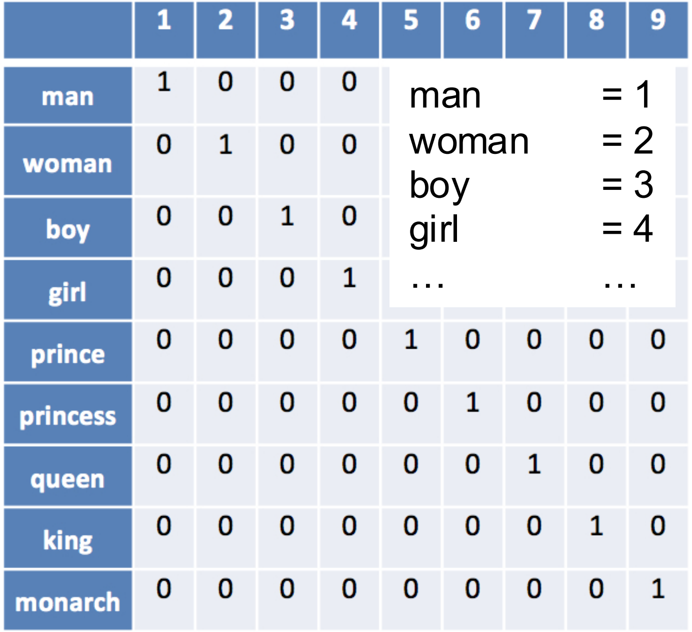
- Unable to fit complex relationships
- Unable to handle raw data like text of image



Representation learning

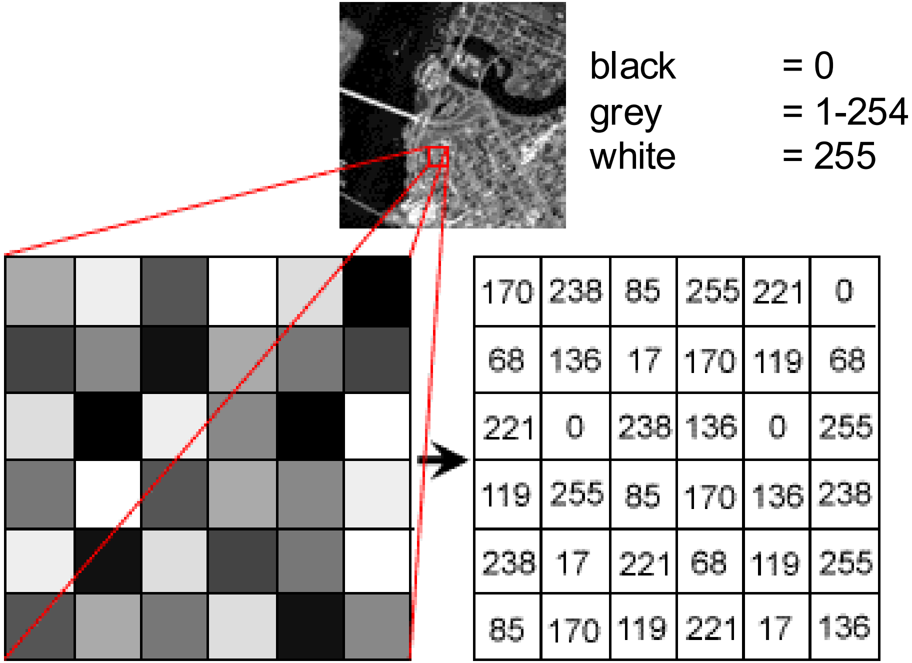
Training computer to parse complex data

Naïve data representation is not useful for prediction



	1	2	3	4	5	6	7	8	9
man	1	0	0	0					
woman	0	1	0	0					
boy	0	0	1	0					
girl	0	0	0	1					
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Image from hackermoon.com



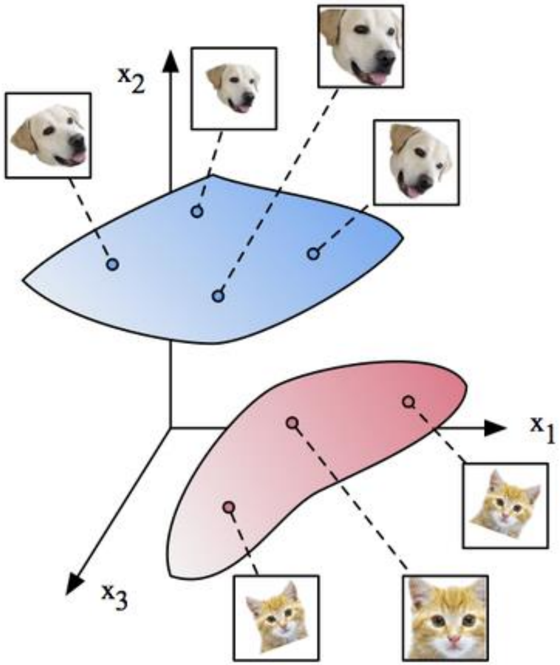
170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

Image from naushardsblog.wordpress.com

What is a (good) representation of the data?

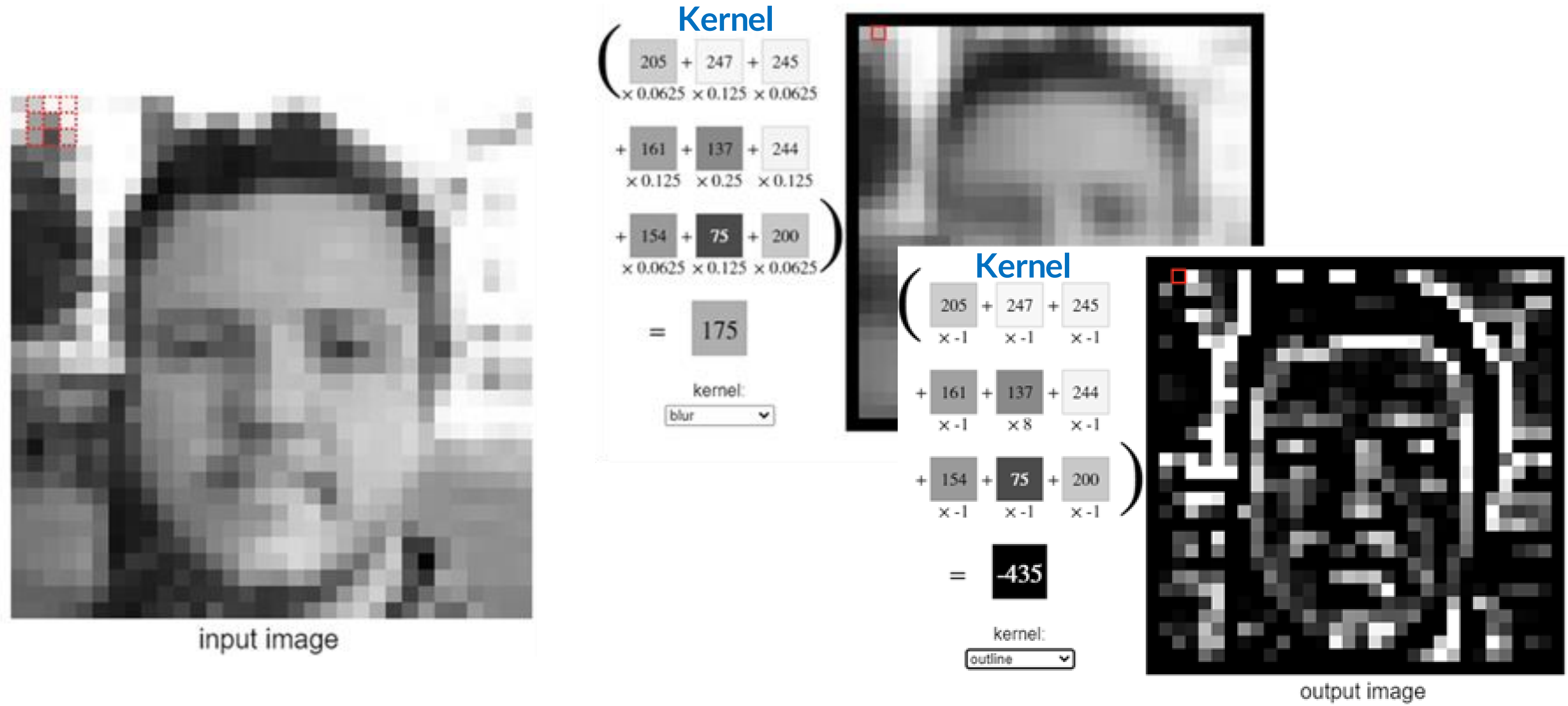
- Numerical mapping of the raw data that are informative of the **key characteristics**

- **Key characteristics** have different meaning based on your purpose for the data
  - Classification / regression
  - Reconstruction
  - Dimensionality reduction

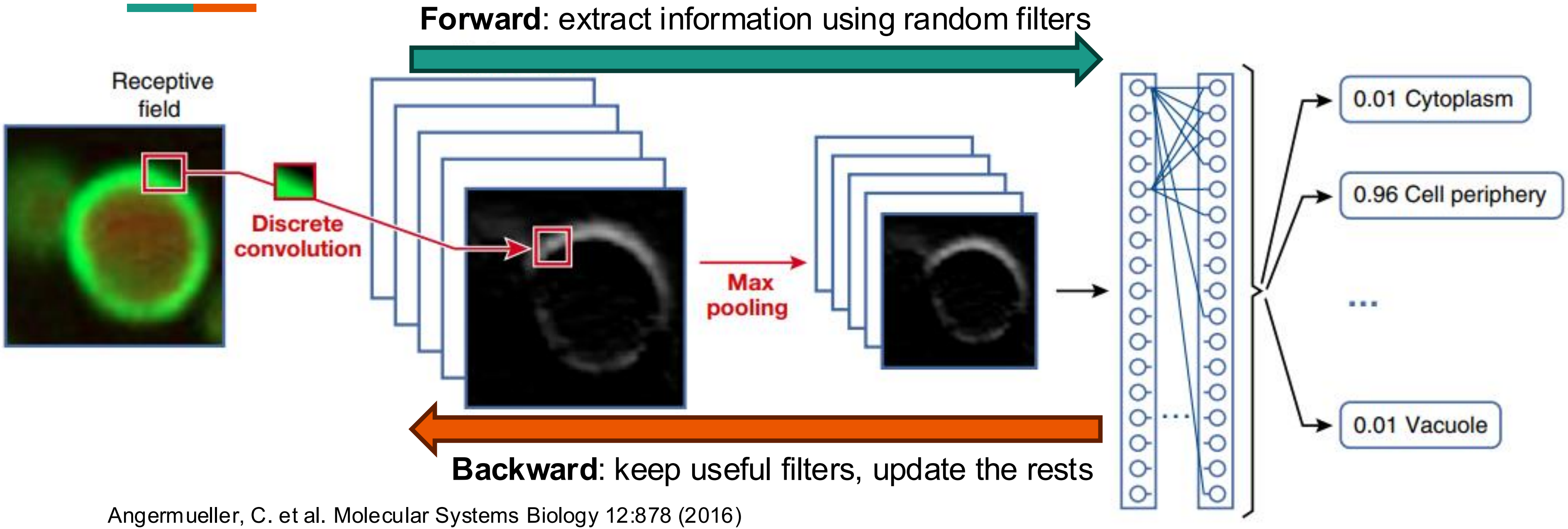


Chung, S. et al. “Classification and Geometry of General Perceptual Manifolds”

Information extraction from image with kernels



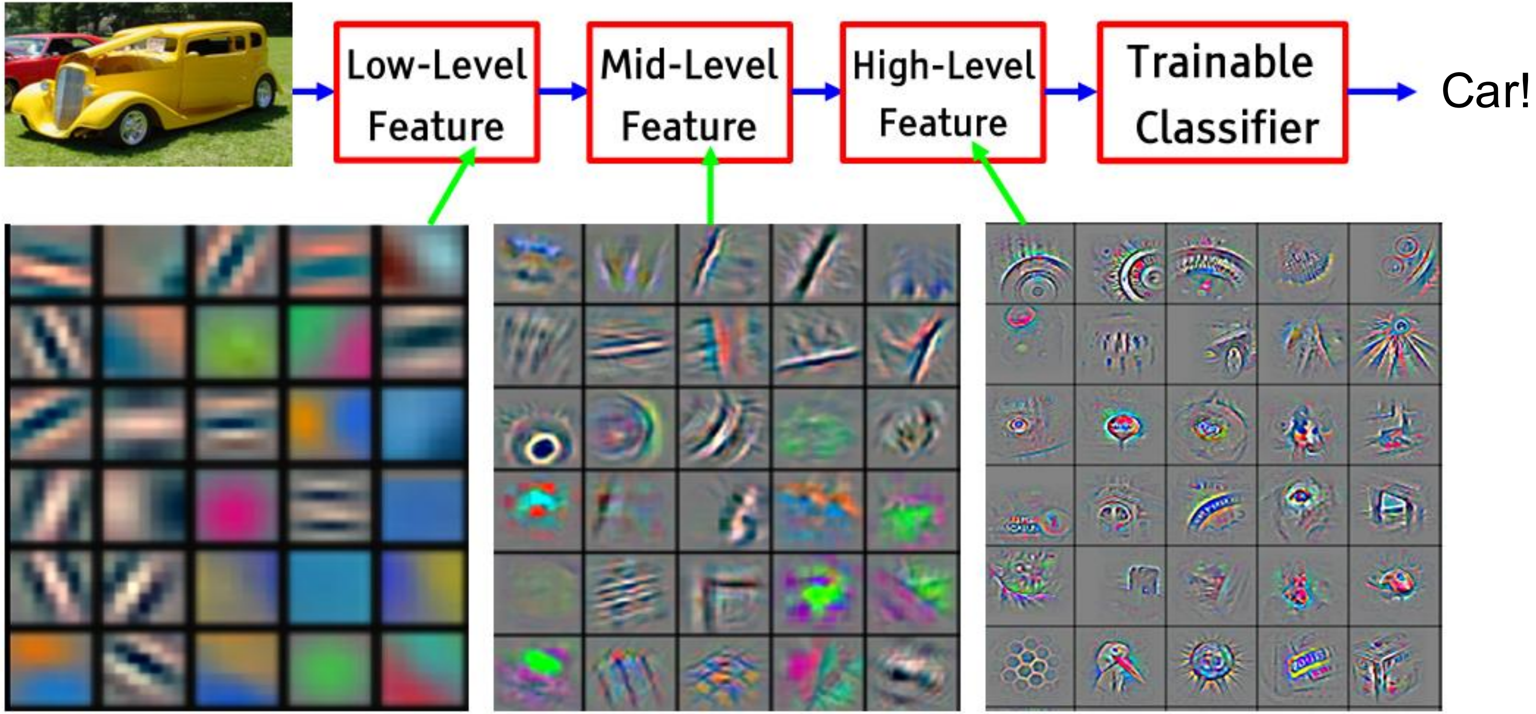
AI 3.0: Data-driven representation learning



Angermueller, C. et al. Molecular Systems Biology 12:878 (2016)

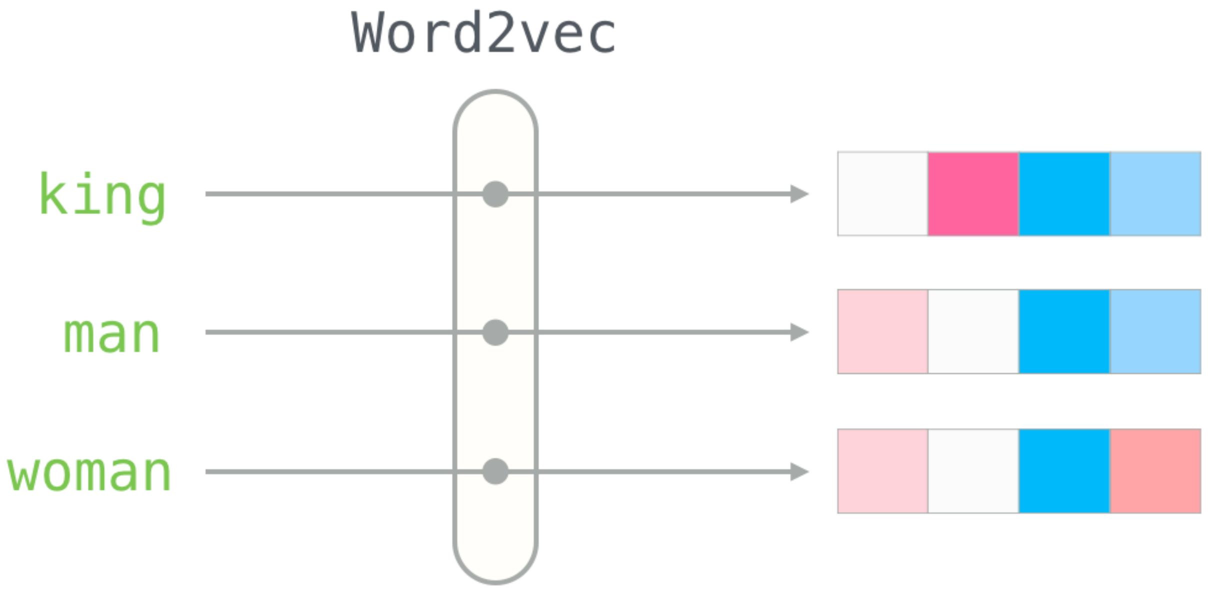
- Finding the right kernels is too difficult for human
- Instead, human provides the data (x, y) and let the computer do the fitting

Convolution kernels create image representation



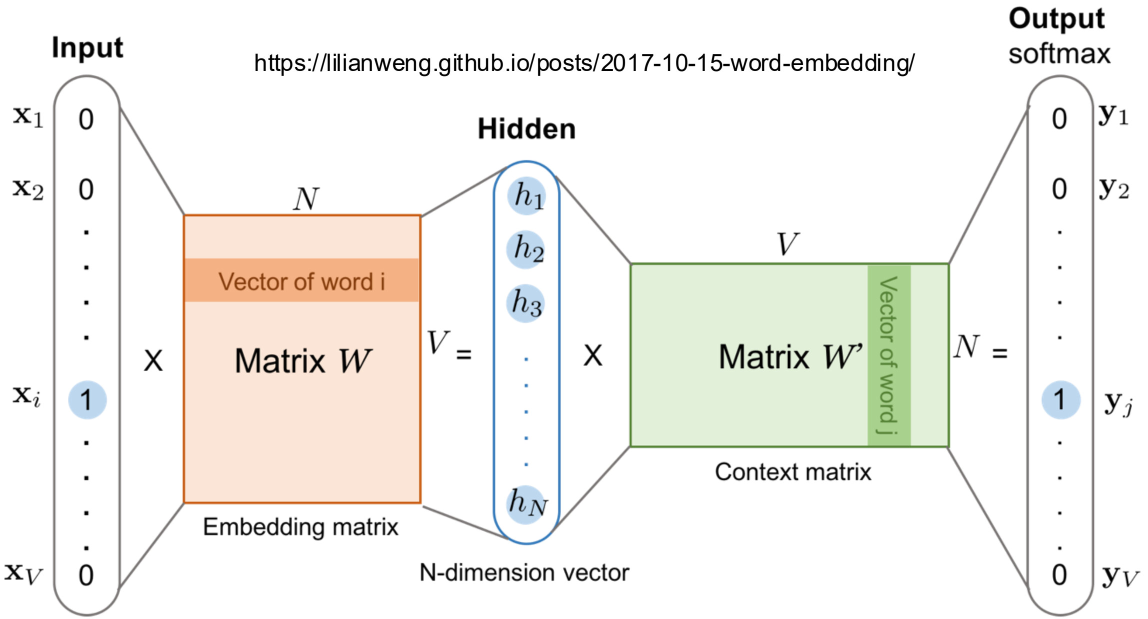
Hinton G, Bengio, Y, and LeCun, Y. NIPS2015

Word embedding



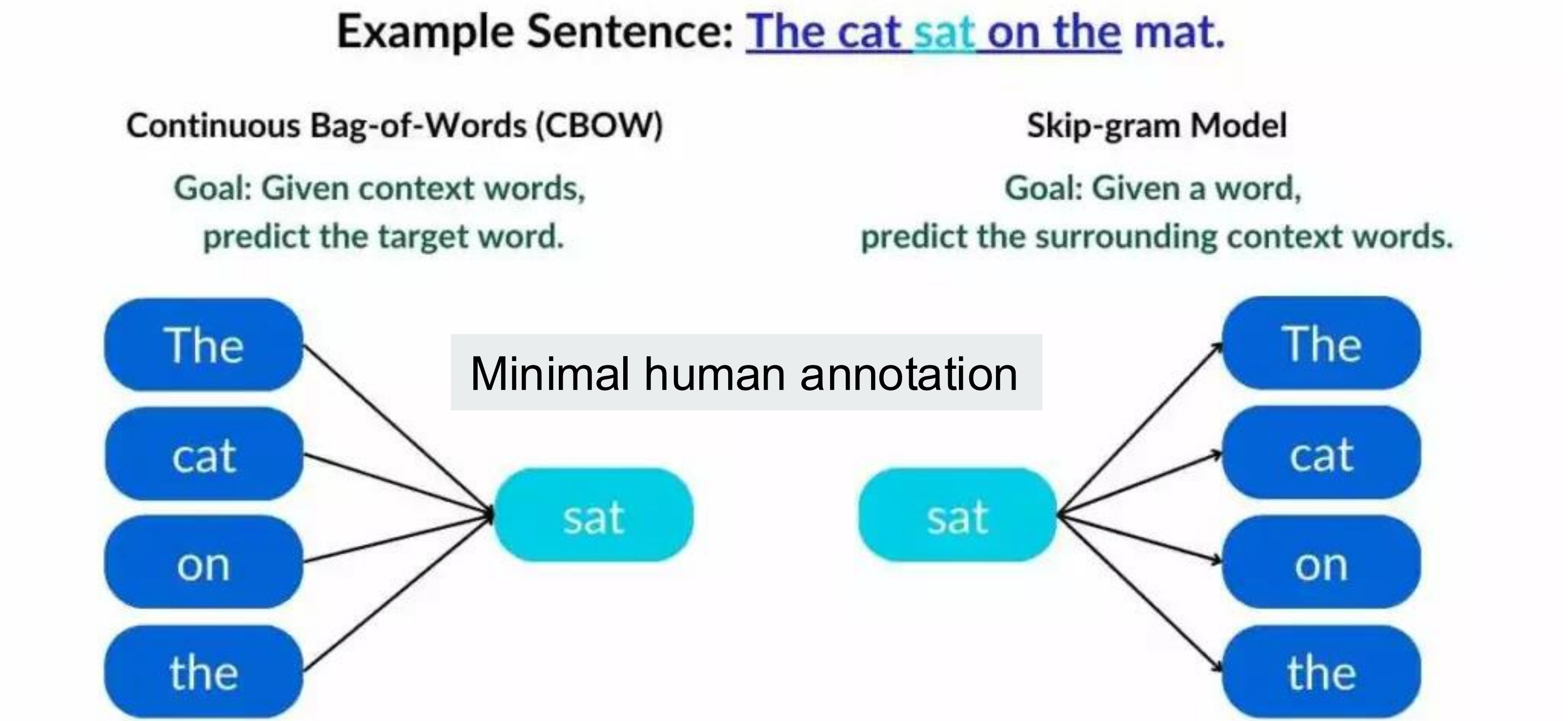
- Transformation of a word into a useful vector that captures its meaning and other characteristics. But how?

How to learn a good word embedding?



- Map each word to an *n*-dimensional vector (representation)
- Use these representations to make prediction, but what to predict?

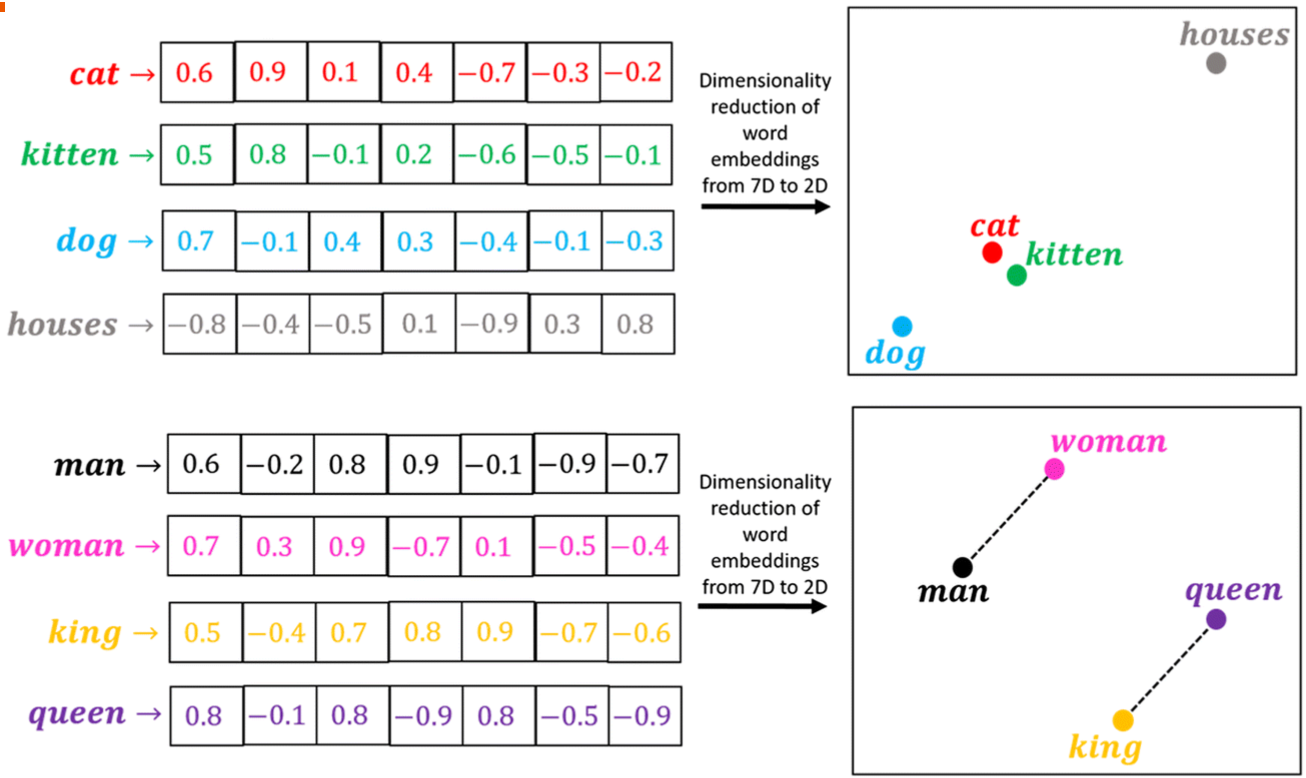
Predict the central word or surrounding words



Minimal human annotation

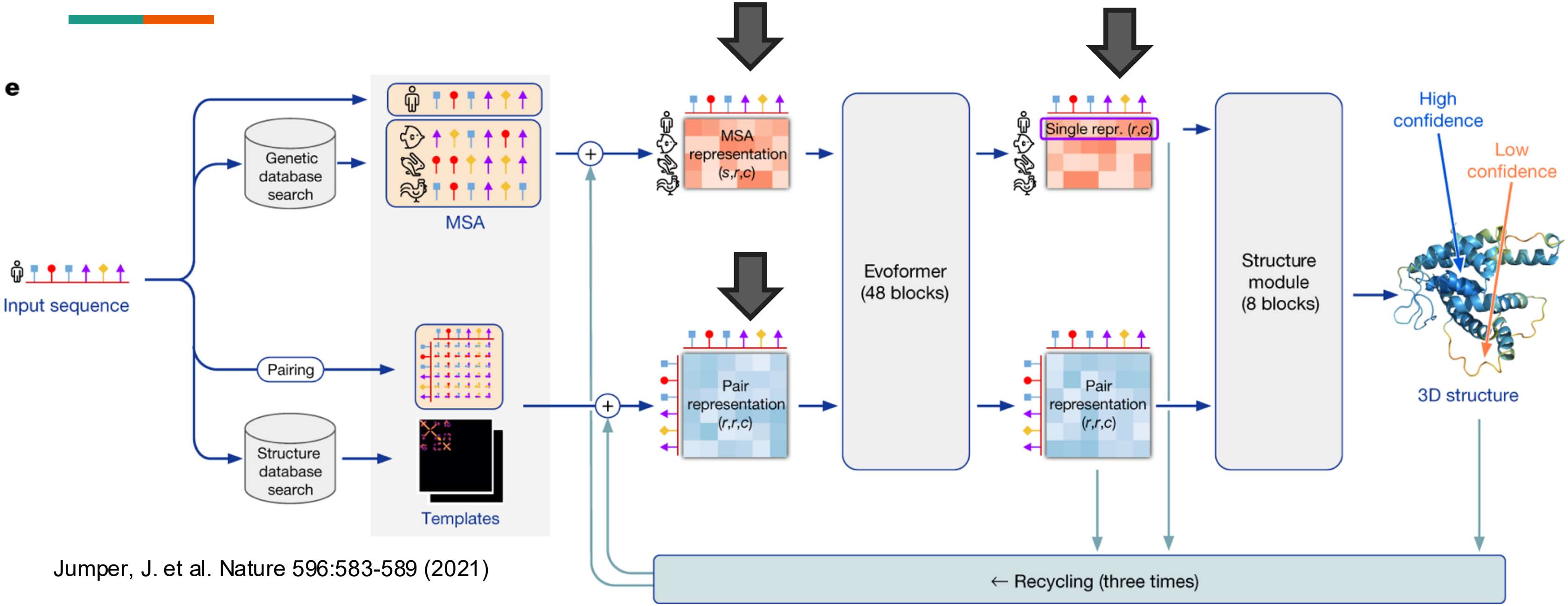
https://spotintelligence.com/2023/12/05/fasttext/

Meaningful word embedding



https://swatimeena989.medium.com/training-word2vec-using-gensim-14433890e8e4

AlphaFold v2 architecture



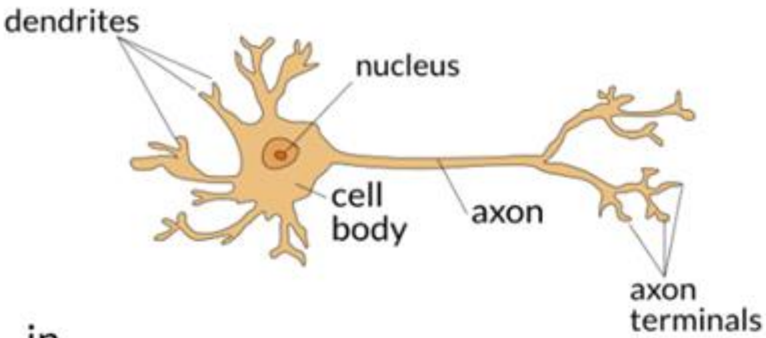
Jumper, J. et al. Nature 596:583-589 (2021)

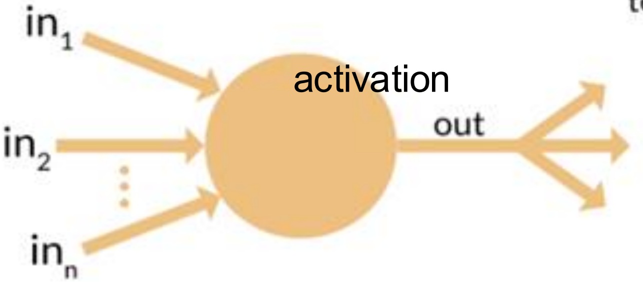
- Multiple representations of different biological concepts are trained to be computationally compatible inside the model

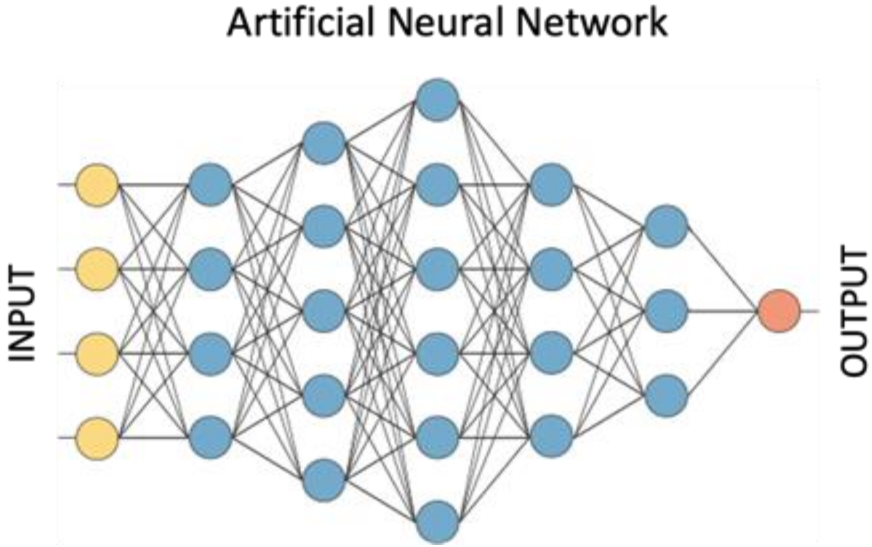
The rise of artificial neural network

An 80-year journey

Inspired by biology

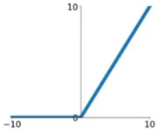




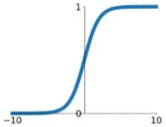


out = f($w_1$in$_1$ + $w_2$in$_2$ + … + $w_n$in$_n$)
f( ) is an activation function

ReLU
$\max(0, x)$



Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$



Activation function

Sigmoid

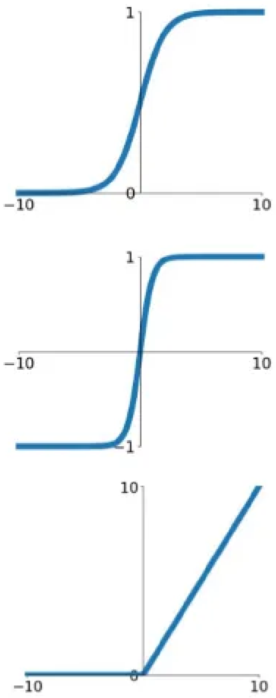
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

tanh

$$\tanh(x)$$

ReLU

$$\max(0, x)$$



- Simple, non-linear function
- Mimic the activation of a biological neuron
- Without non-linear activation function, ANN is just a linear regression

https://medium.com/@shrutijadon/survey-on-activation-functions-for-deep-learning-9689331ba092

Universal approximation theorem (Cybenko 1989)

Universal Approximation Theorem: Fix a continuous function $\sigma : \mathbb{R} \to \mathbb{R}$ (activation function) and positive integers d, D. The function σ is not a polynomial if and only if, for every continuous function $f : \mathbb{R}^d \to \mathbb{R}^D$ (target function), every compact subset K of $\mathbb{R}^d$, and every $\epsilon > 0$ there exists a continuous function $f_\epsilon : \mathbb{R}^d \to \mathbb{R}^D$ (the layer output) with representation

$$f_\epsilon = W_2 \circ \sigma \circ W_1,$$

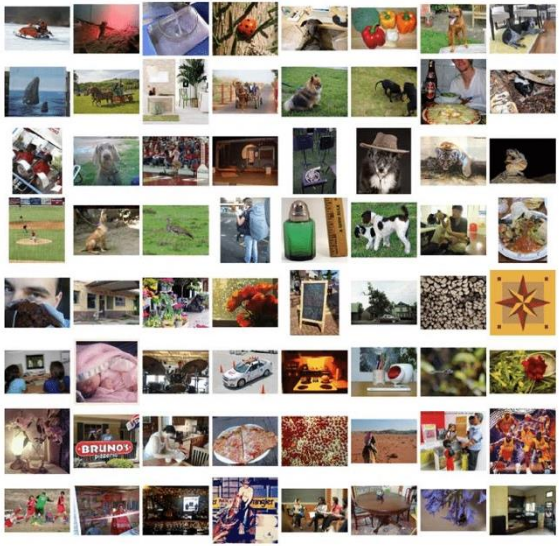
where W_2, W_1 are composable affine maps and $\circ$ denotes component-wise composition, such that the approximation bound

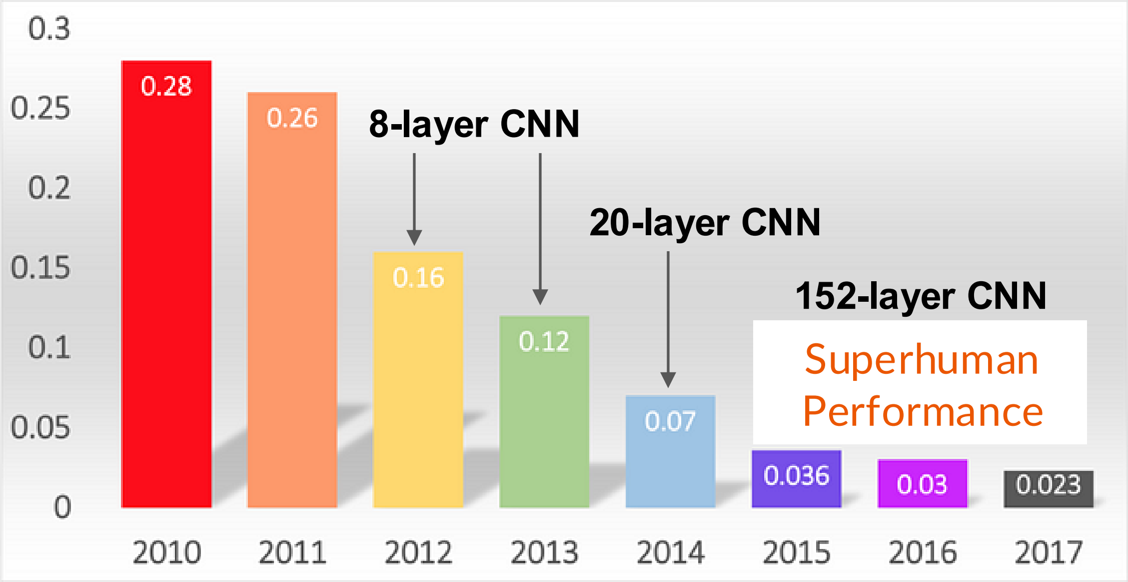
$$\sup_{x \in K} \|f(x) - f_\epsilon(x)\| < \varepsilon$$

holds for any ϵ arbitrarily small (distance from f to f_ϵ can be infinitely small).

- ANN with just one layer of neurons and a non-polynomial activation function **can capture any continuous mathematical relationship**

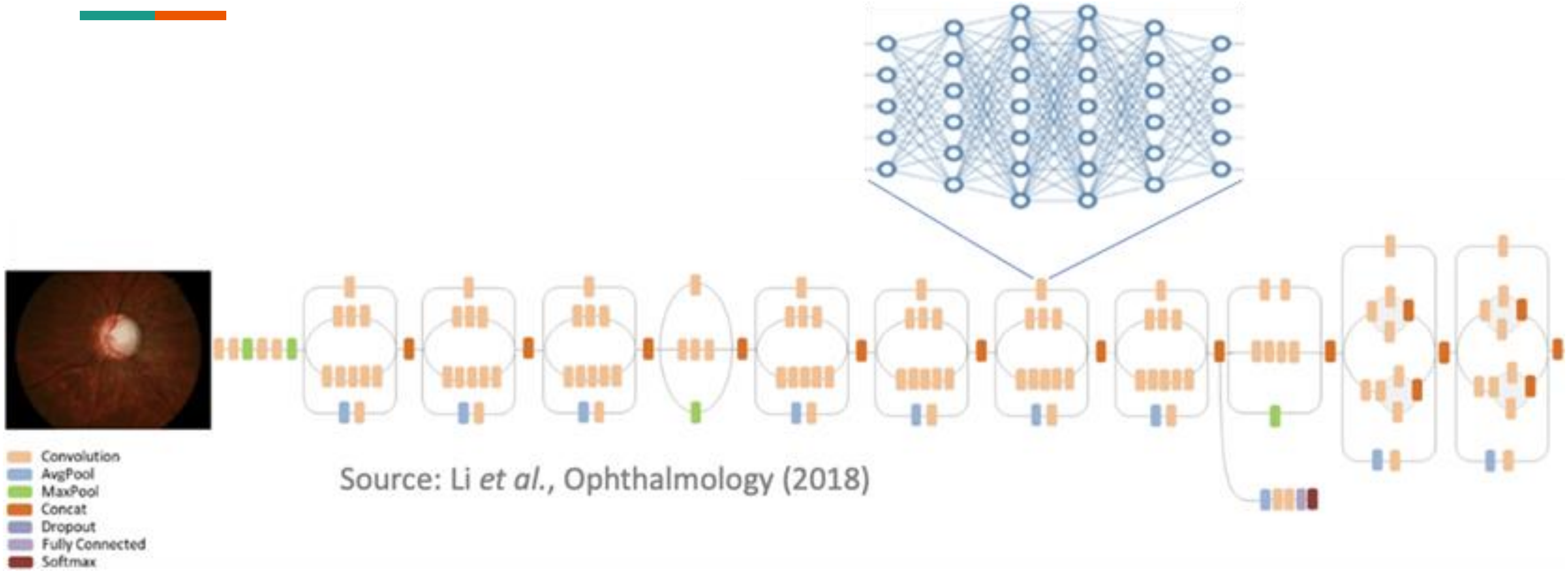
ImageNet: a showcase of ANN on real-world data





- Emergence of neural network capability due to **data** (internet and digital technology) and **computing resources** (GPU)

Deep learning is ML for deep neural network



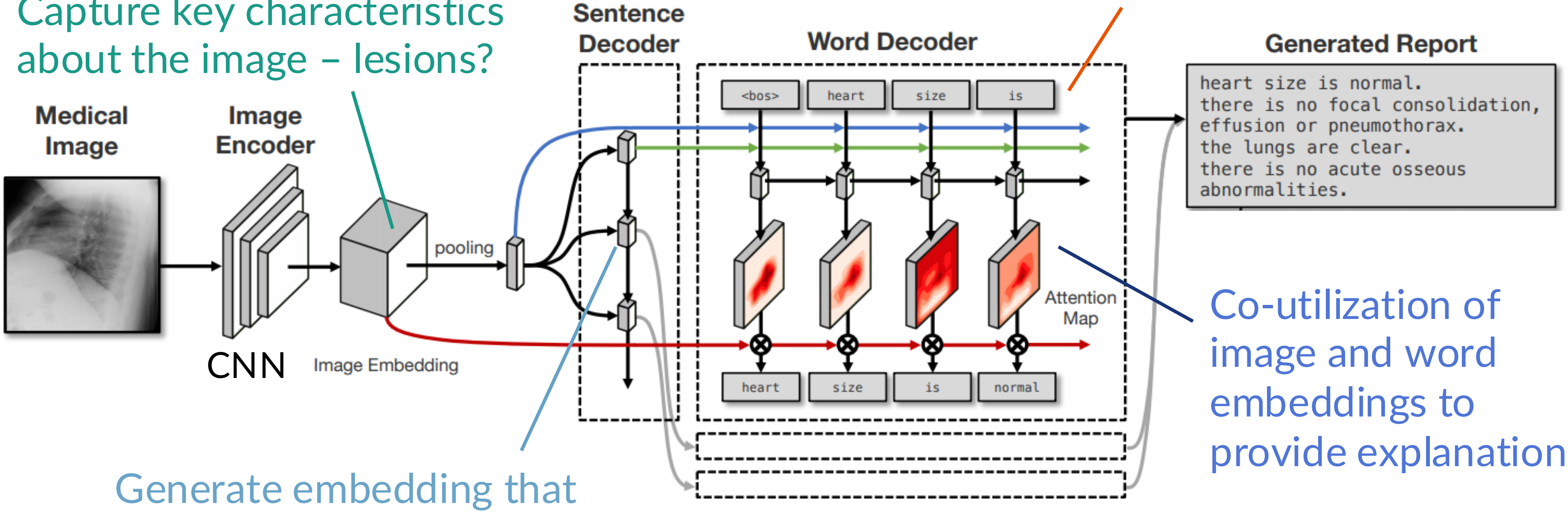
Source: Li *et al.*, Ophthalmology (2018)

- Explosion in dataset and model sizes
- Techniques for guiding the model to learn embedding

AI 3.0 learns and utilizes embedding

Sequentially generate words that follow the learned language and defined topic

Capture key characteristics about the image – lesions?



Co-utilization of image and word embeddings to provide explanation

Generate embedding that define a sentence's topic

Recurrent Neural Network (RNN)

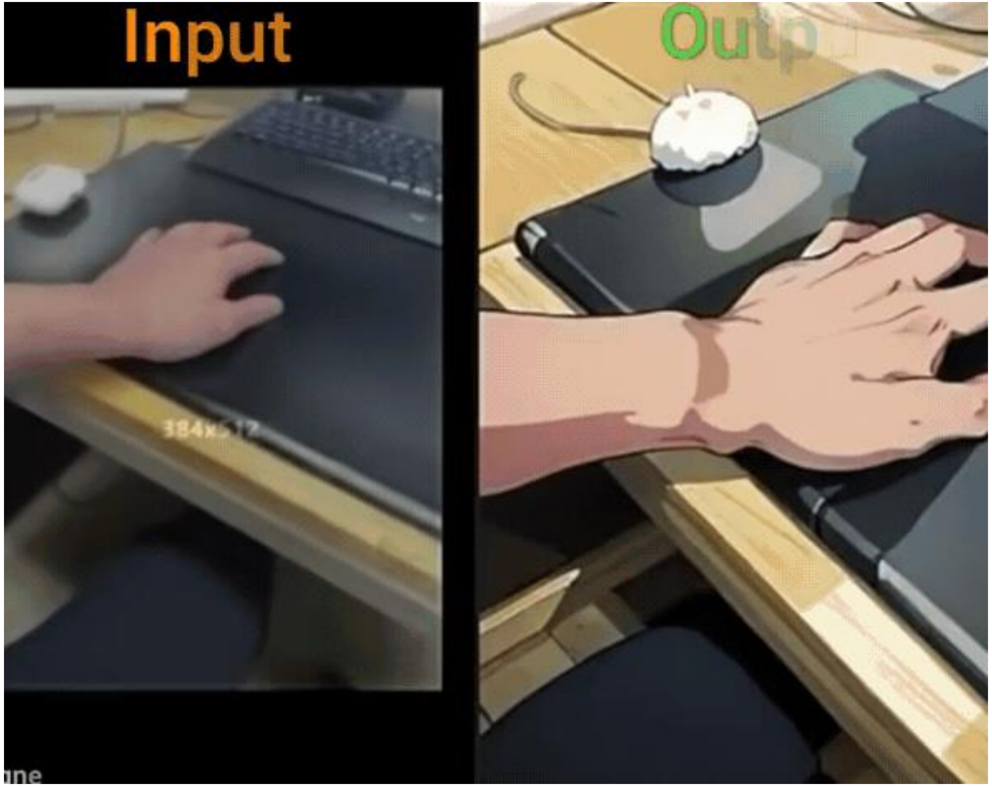
Liu et al. Clinically Accurate Chest X-Ray Report Generation (2019)

Linking visual and text embeddings with robot control

AI 4.0: Generative and foundational models

Going beyond prediction

Talking head anime: https://github.com/pkhungurn



Stable Diffusion: 8

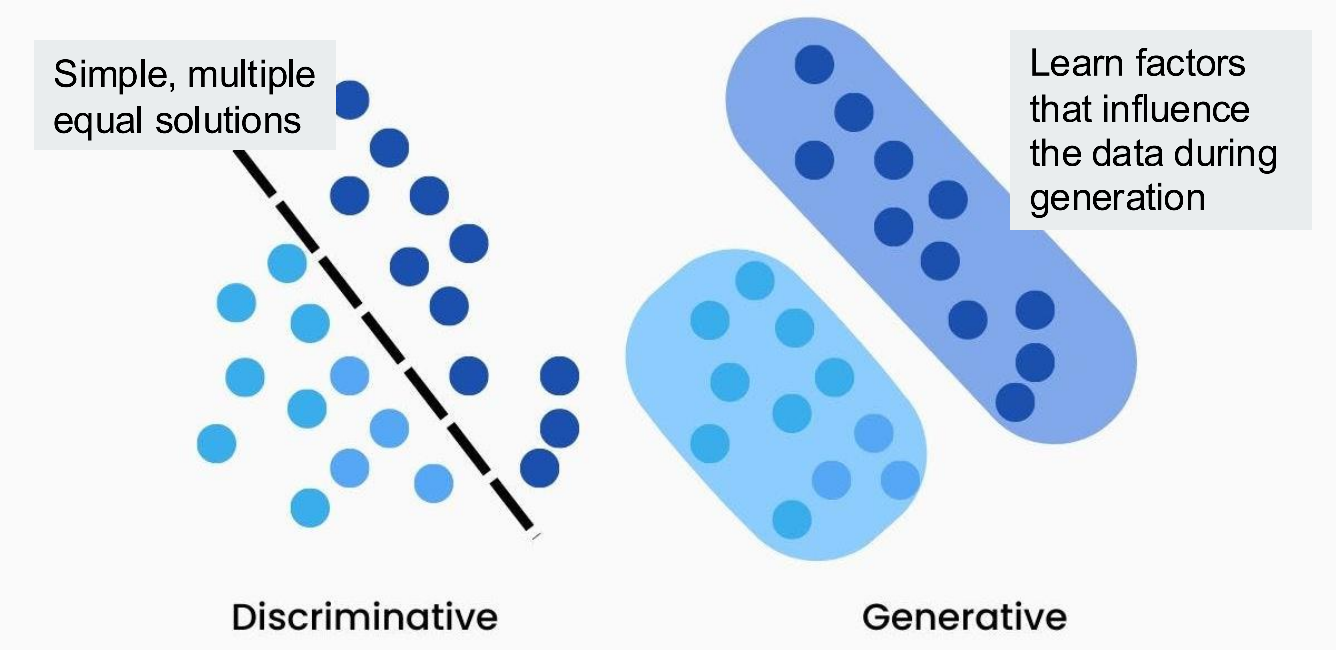
The rise of large language model

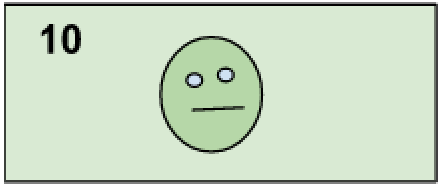
Upon examination of the chest X-ray image:

- The lungs are predominantly clear without any obvious consolidations, masses, or pneumothoraces.
- The cardiac silhouette appears within normal size limits.
- The bony thorax, including the ribs and clavicles, appears intact without any visible fractures.
- The diaphragm and costophrenic angles are well visualized and appear normal.
- There is no visible mediastinal widening or significant lymphadenopathy.

https://xrayinterpreter.com/resource/how-to-use-chatgpt-to-interpret-xrays

Importance of generative approach



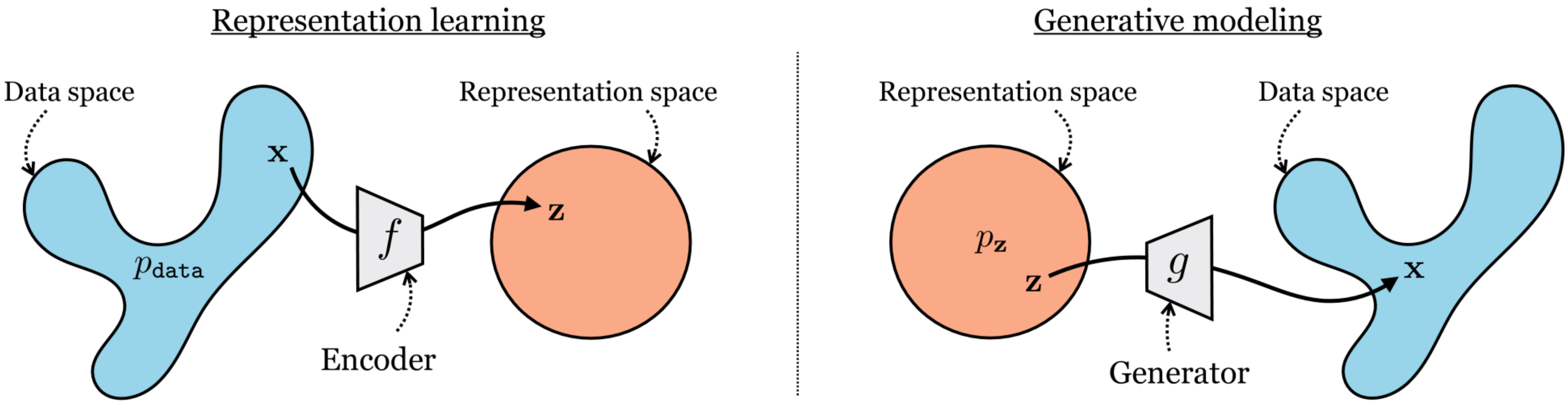


vs

https://www.turing.com/kb/generative-models-vs-discriminative-models-for-deep-learning

- It takes much more understanding to generate **realistic** data

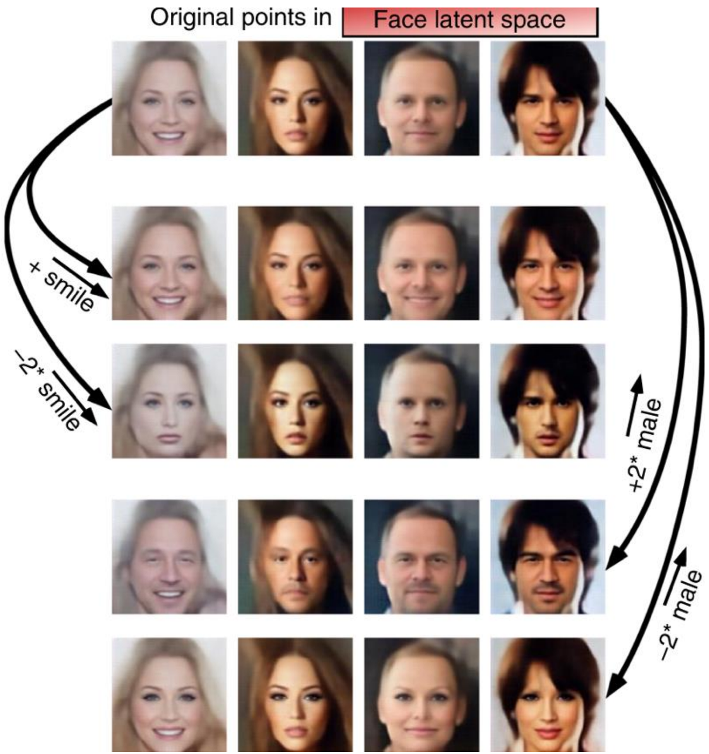
Reversed representation learning



https://visionbook.mit.edu/generative_modeling_and_rep_learning.html

- Generative model can be thought of as reversed representation learning
  - Compress vs decompress
- How do we train, or “guide”, the generative process?

Interpretable generative process



Van Rullen, R. and Reddy, L. Communications Biology 2:193 (2019)

- **Assumption**: Generative model helps us understand the underlying mechanisms and the factors that generated the data

- Disentangle factors from observation

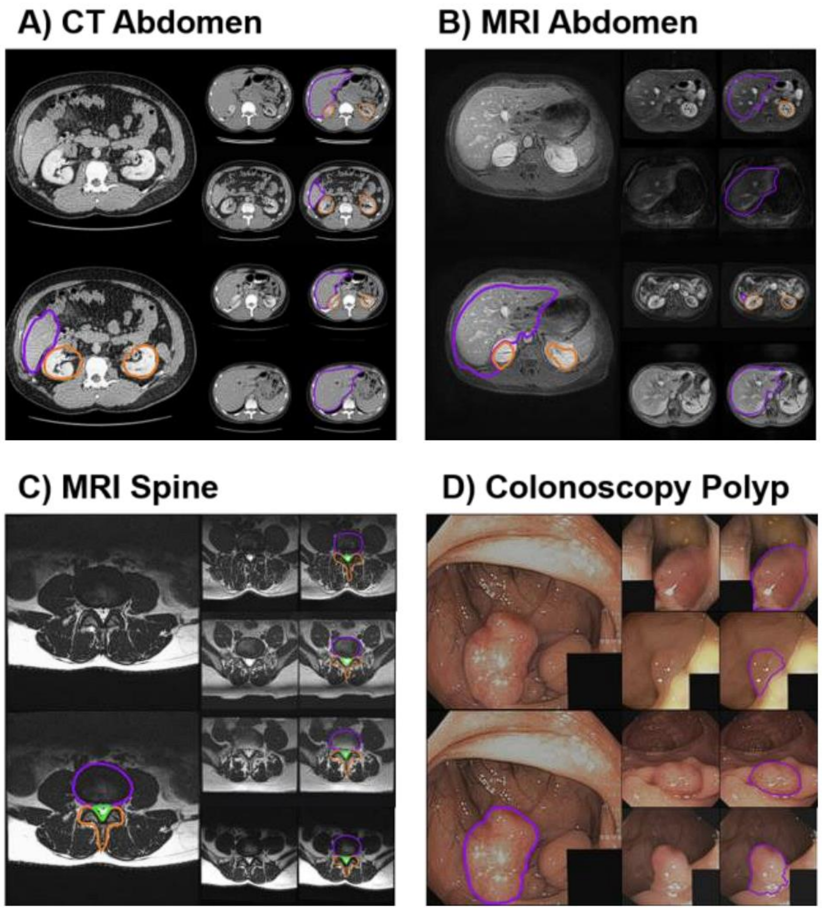
- Counterfactual “what-if” analysis

RadImageGAN

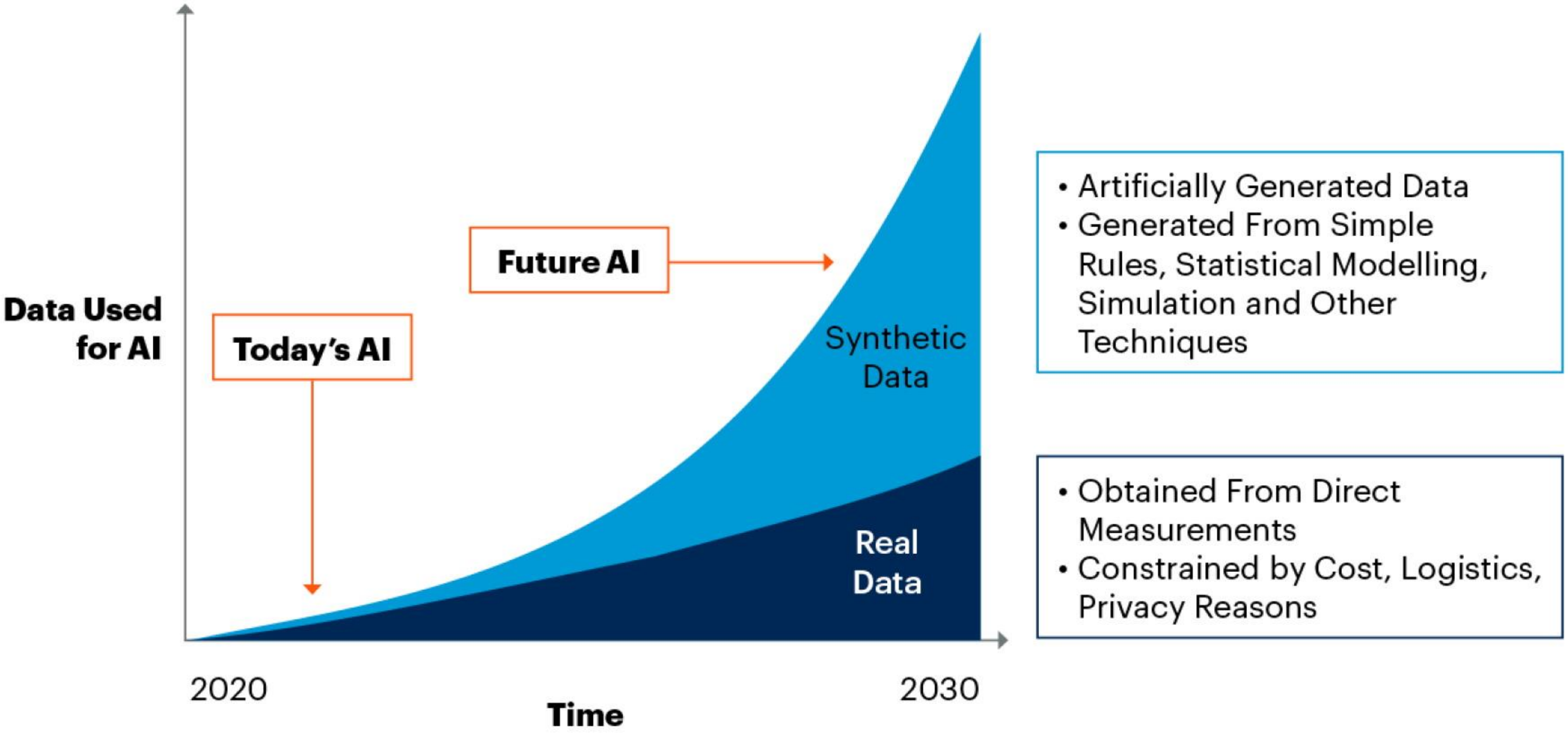
- StyleGAN on radiological images from 100,000 patients

- 12 anatomical regions and 130 pathological classes

- Can generate both images and labeled masks to train future AI models

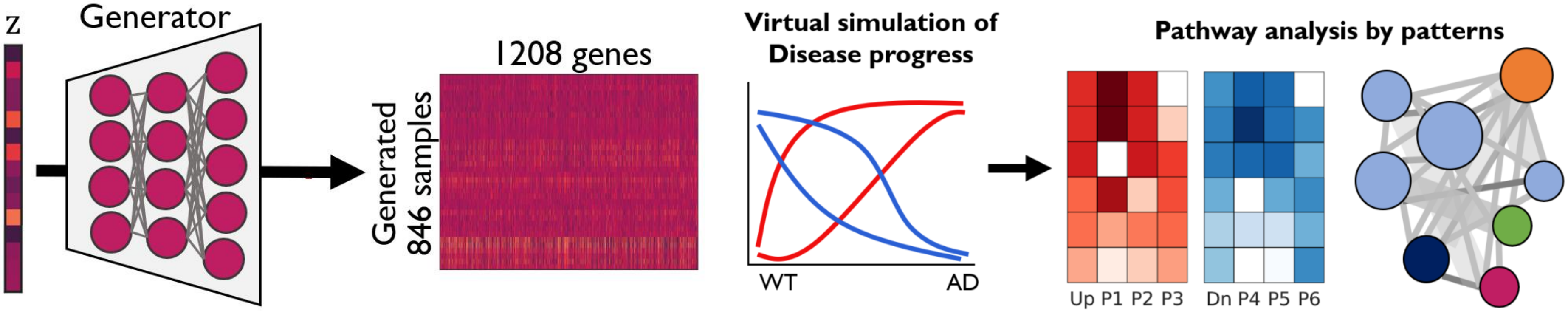


By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



Source: Gartner

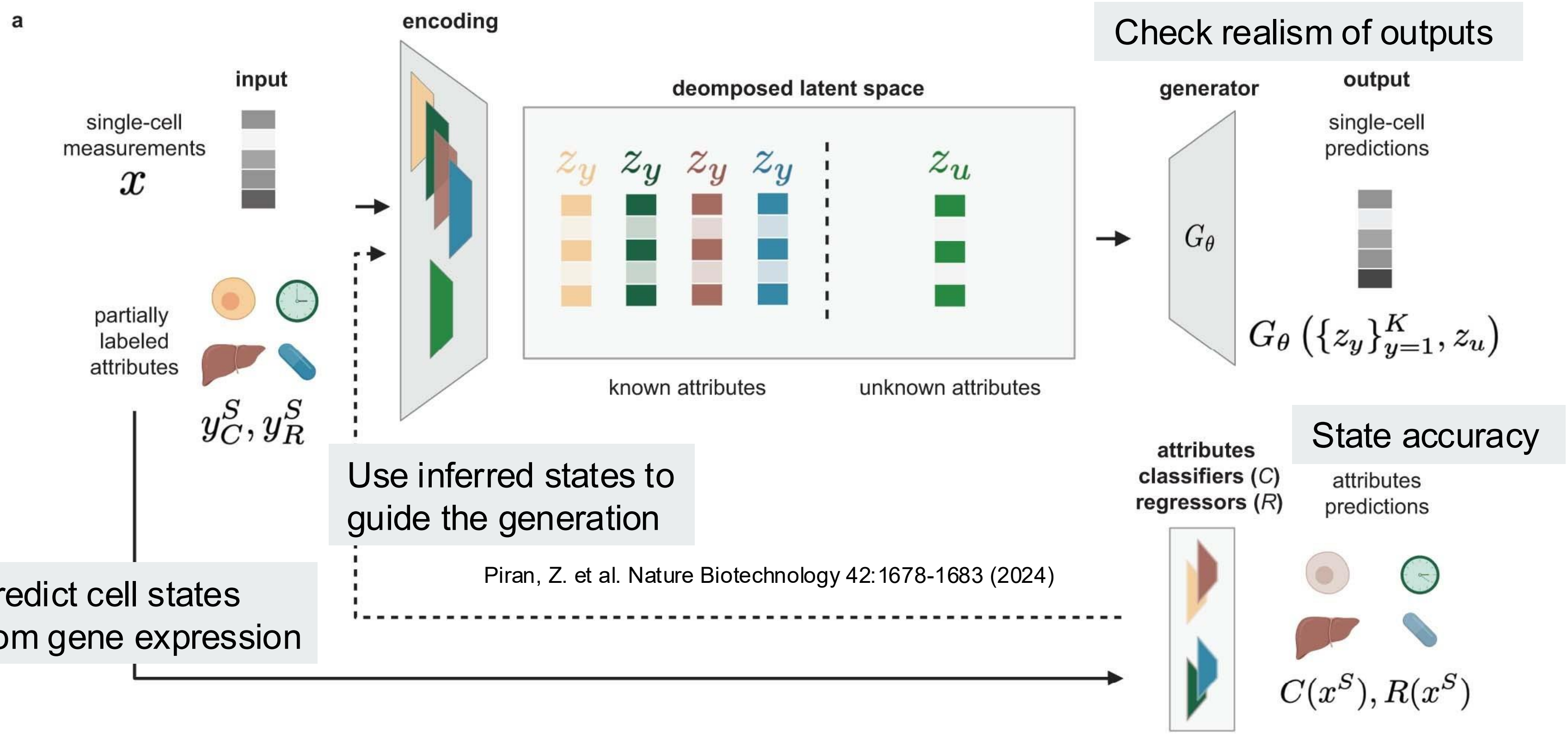
Knowledge from synthetic data



Park, J. et al. PLoS Computational Biology 16:e1008099 (2020)

- Train a generative model with data from small-scale experiment
- Simulate time-course transcriptomics data
- Analyze simulated data to gain insights into disease progression
- Validate biomarkers in external datasets

Disentangle factors influencing gene expression



Piran, Z. et al. Nature Biotechnology 42:1678-1683 (2024)

Summary

- Early AI were driven by human knowledge and fixed rules
- The emergence of digital computer and internet enabled data-driven, machine learning approach to AI
- New hardware and training techniques gave rise to the ability of AI to learn representations by themselves
- Learning through generation allows modern AI to mimic complex mechanisms behind the data and disentangle important factors

Any question?

- See you next time