3000788 Intro to Comp Molec Biol

Lecture 30: Deep learning in life sciences

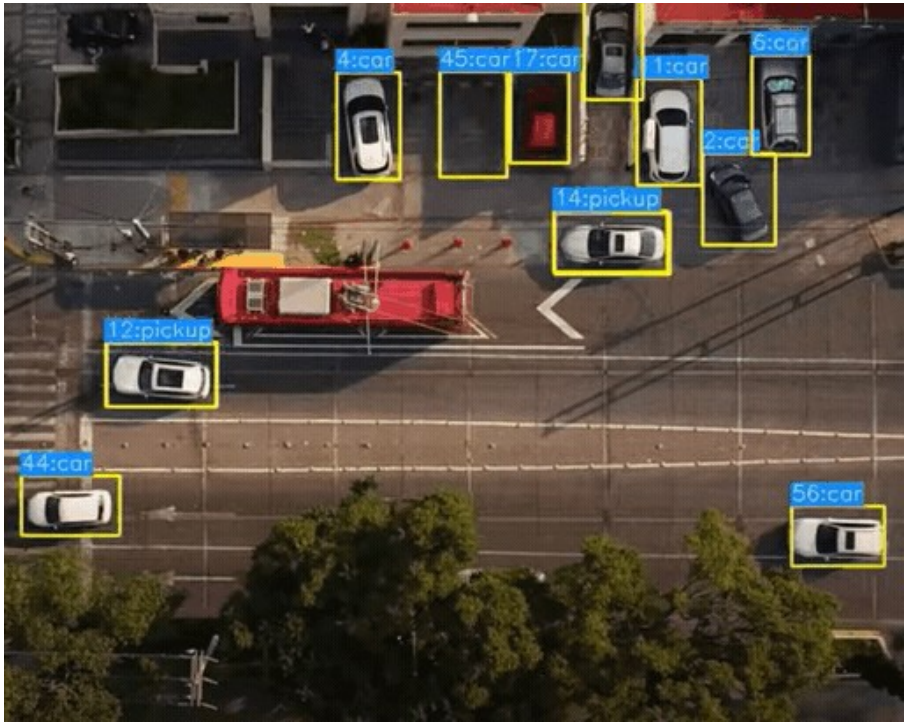
November 30, 2023

Sira Sriswasdi, PhD

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

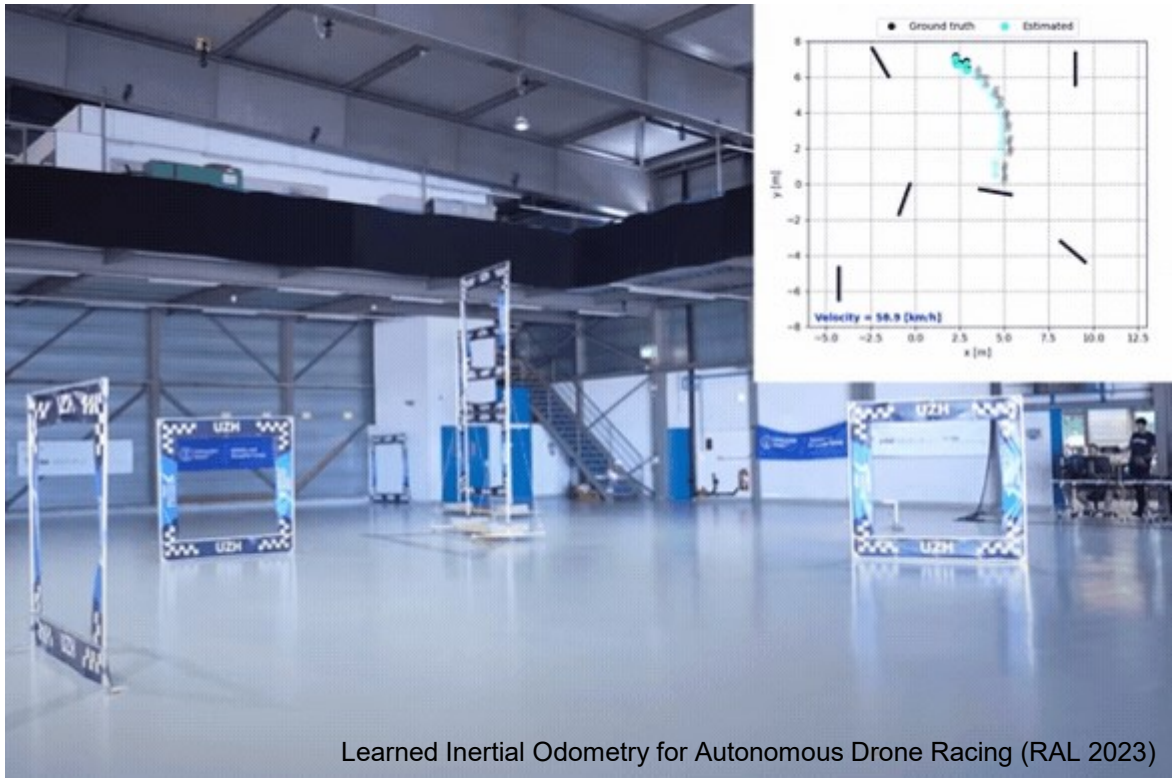
Today’s deep learning

Real-time object recognition



Real Time Object Detection and Tracking with YOLO-NAS and DeepSORT on Custom Dataset

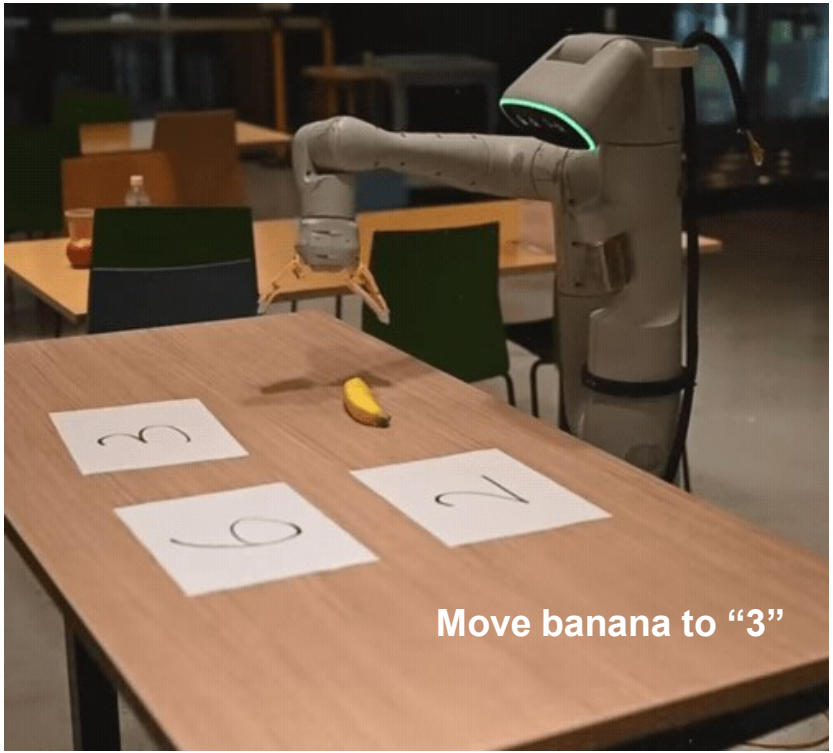
Interaction with real-world environment



Learned Inertial Odometry for Autonomous Drone Racing (RAL 2023)

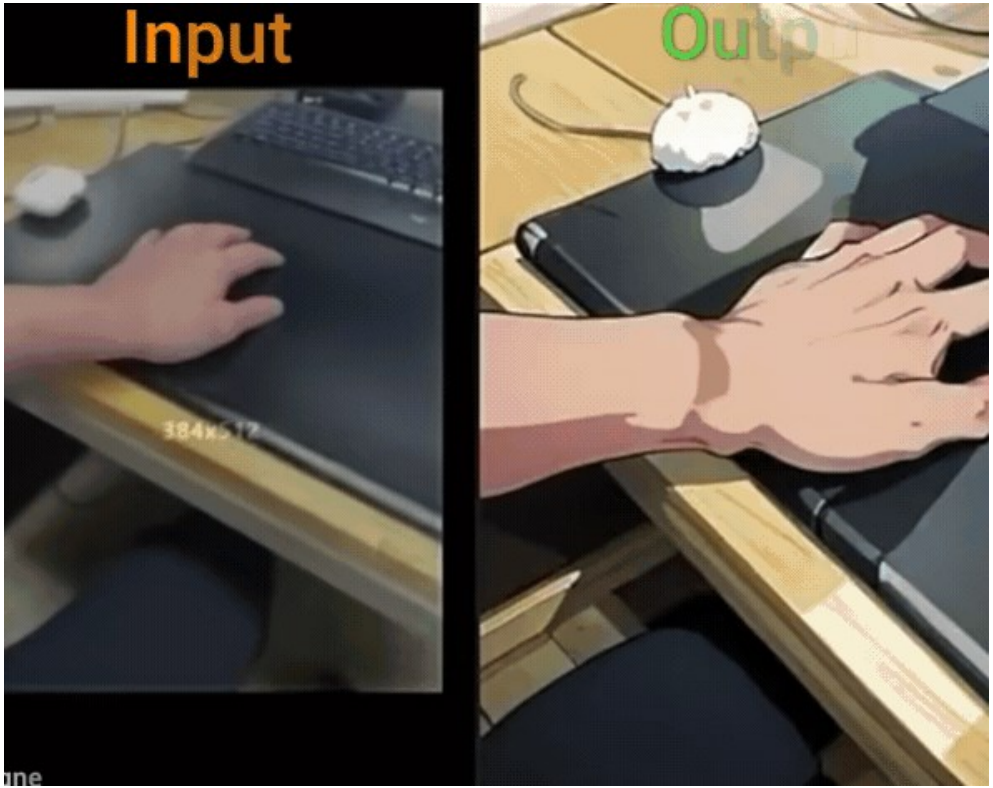
Google DeepMind's AlphaGo computer beats top player Lee Sedol for third time to sweep competition

Text to action: Google RT-2



Art and deepfake

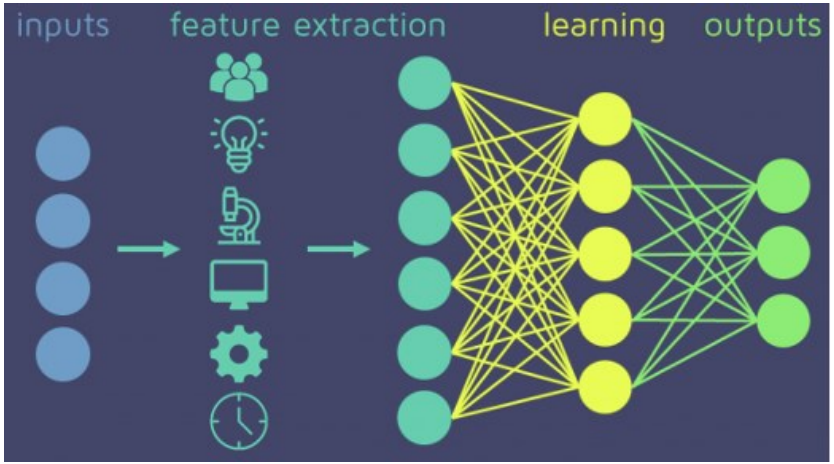
Talking head anime: https://github.com/pkhungurn

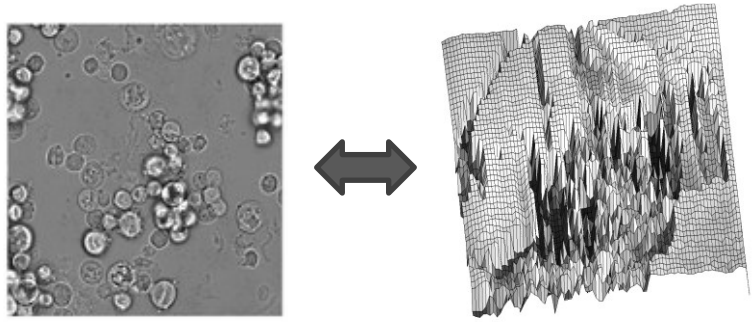


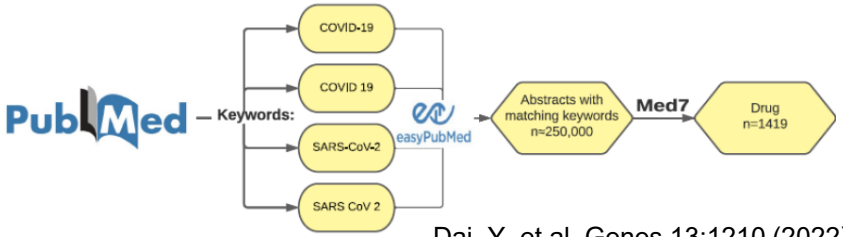
Stable Diffusion: 8

How did the magic happen?

Limitation of classical (non-deep) learning



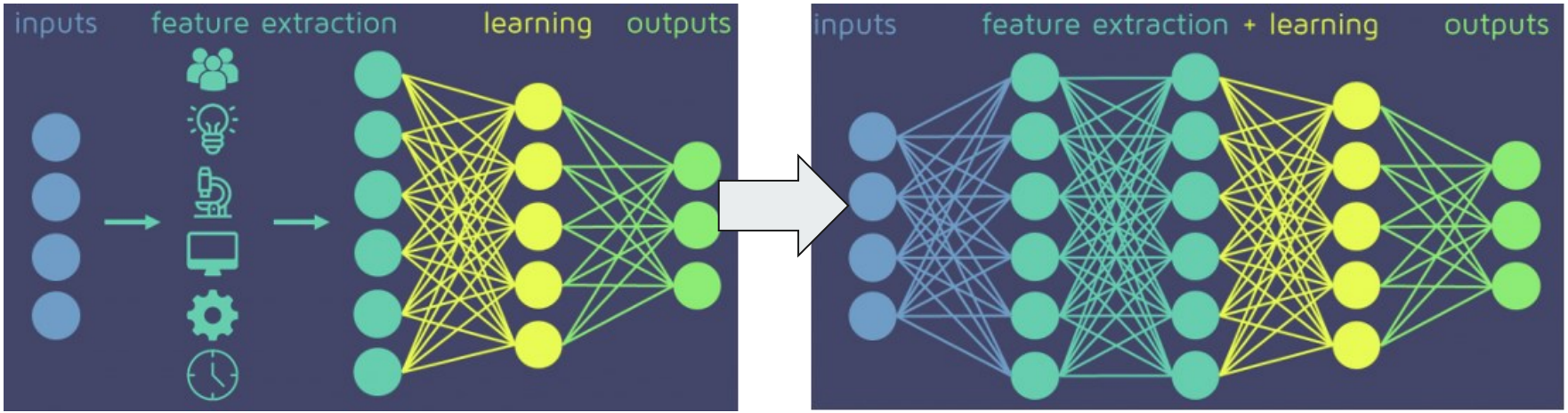




Dai, Y. et al. Genes 13:1210 (2022)

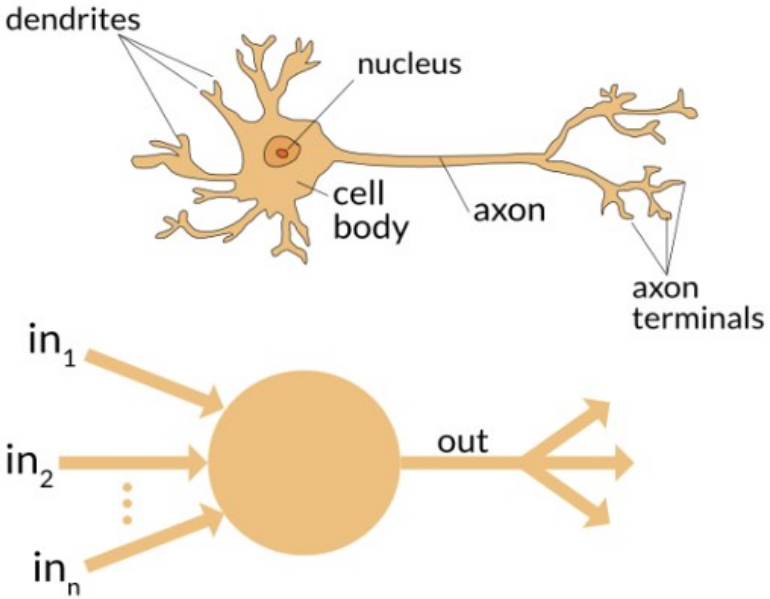
- Classical machine learning requires the input to be formatted and pre-processed by human

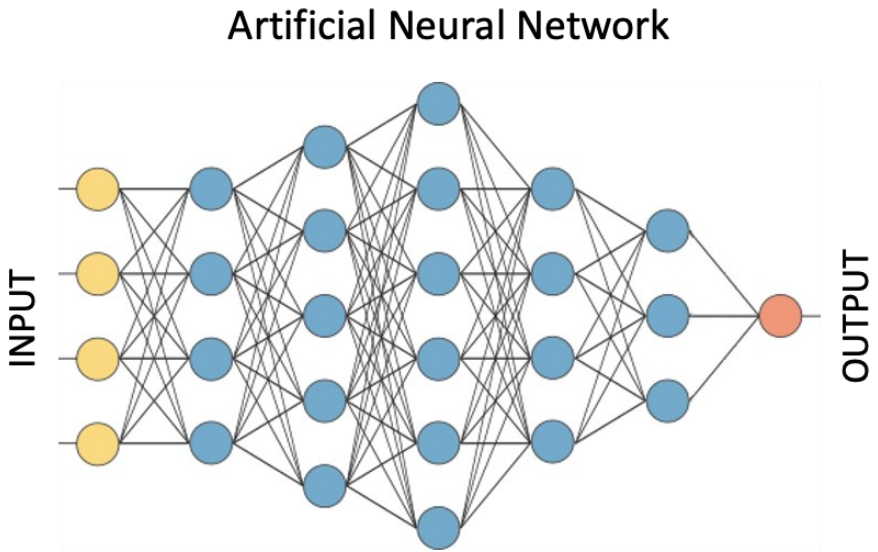
End-to-end learning



- Deep learning, via artificial neural network models, can learn to extract useful information from raw input directly
- The catch is a lot of data and supervision is needed

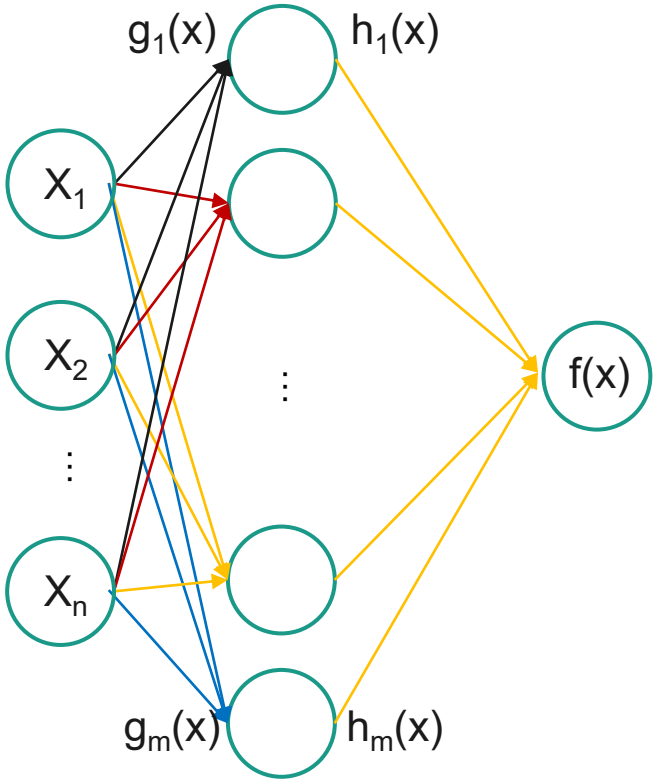
Artificial neural network





- Network of simple computation nodes: $out = f(w_1in_1 + w_2in_2 + \dots + w_nin_n)$

Calculations inside neural nentwork



Linear neuron input

- $g_1(x) = w_{1,1}x_1 + \cdots + w_{1,n}x_n$
- $g_m(x) = w_{m,1}x_1 + \cdots + w_{m,n}x_n$

Sigmoid activation

- $h_1(x) = \frac{1}{1+e^{-g_1(x)}}$
- $h_m(x) = \frac{1}{1+e^{-g_m(x)}}$

Linear aggregated output

- $f(x) = u_1 h_1(x) + \cdots + u_m h_m(x)$

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

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

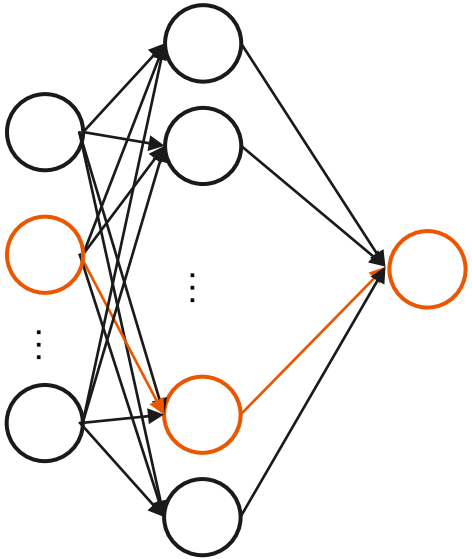
where W_2, W_1 are composable affine maps and $\circ$ denotes component-wise composition, such that the approximation bound

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

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

- Neural network with one hidden layer can mimic any mathematical function

Gradient of a neural network



Neuron input: $g_i(x) = w_{i,1}x_1 + \cdots + w_{i,n}x_n$

Sigmoid activation: $h_i(x) = \frac{1}{1+e^{-g_i(x)}}$

Linear output: $f(x) = u_1 h_1(x) + \cdots + u_m h_m(x)$

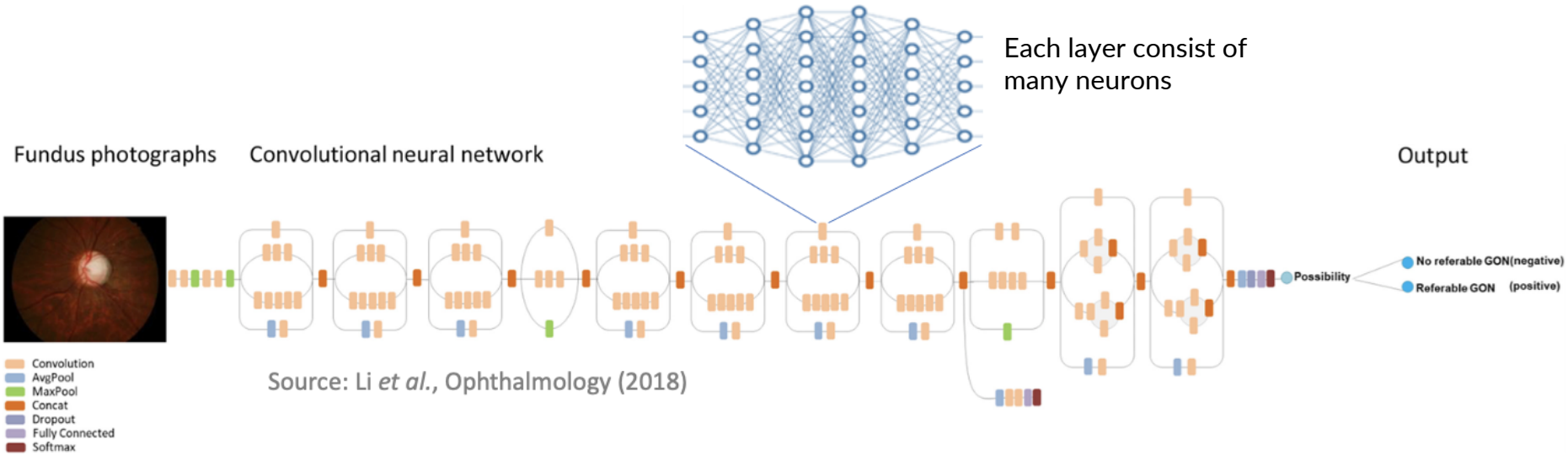
MSE loss: $L(f(x), y) = \frac{1}{2}\|f(x) - y\|^2$

Gradient: $\frac{\delta L}{\delta w_{i,j}} = ?$ ($w_{i,j}$ is the weight for *j*-th feature entering *i*-th neuron)

$$\frac{\delta L}{\delta w_{i,j}} = \frac{\delta L}{\delta f}\frac{\delta f}{\delta h_i}\frac{\delta h_i}{\delta g_i}\frac{\delta g_i}{\delta w_{i,j}} = (f(x) - y) \cdot u_i \cdot g_i(\mathrm{x})(1 - g_i(\mathrm{x}))\, x_j$$

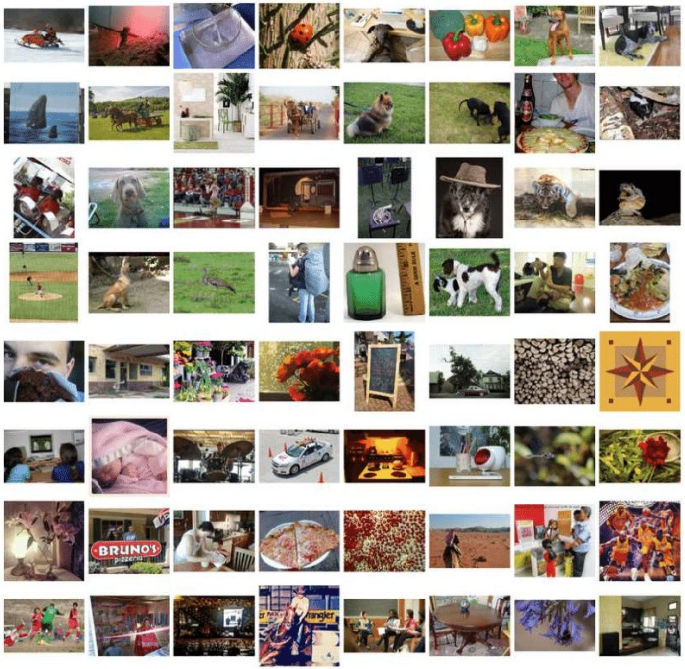
Toward deep(er) learning

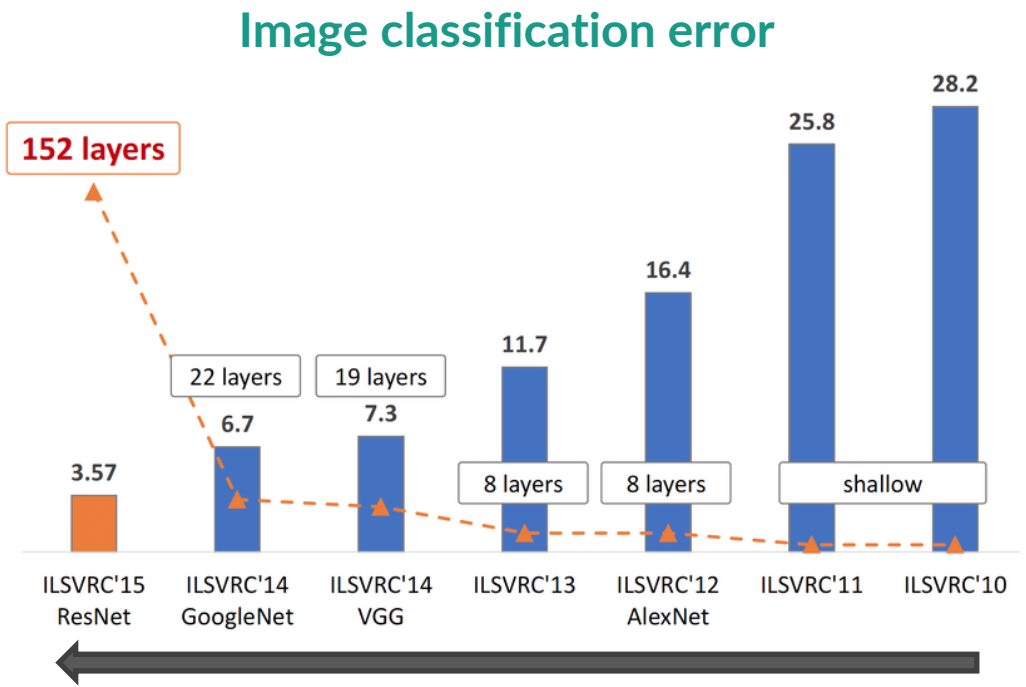
Deep artificial neural network



- Up to billions of parameters
- Deep learning is the technique for developing deep artificial neural network and theory on how such feat is possible

ImageNet: The rise of deep artificial neural network





Graphical processing unit (GPU)

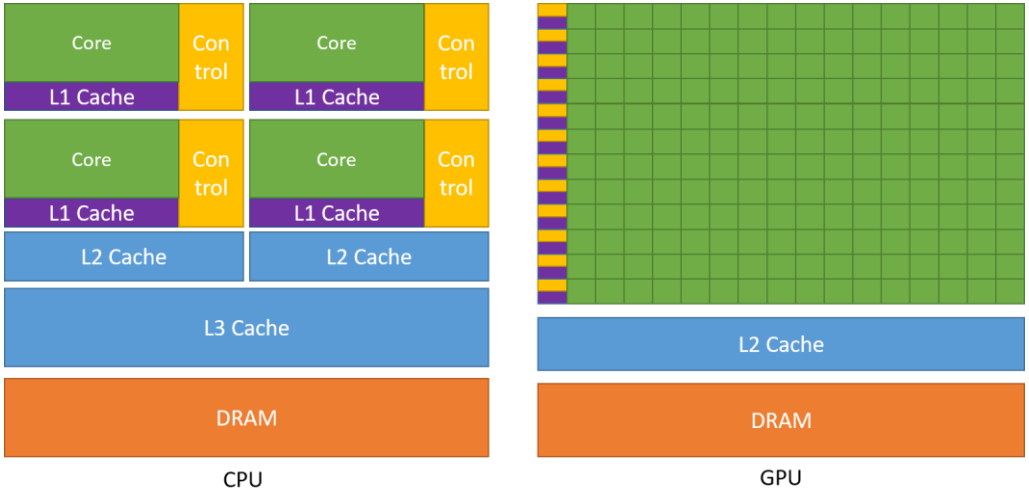


Image from analyticsvidhya.com

- Calculation of gradient for an ANN requires millions of simple operations that can be performed in parallel → Similar to the calculation of graphics

Representation learning

Naïve representations

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
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

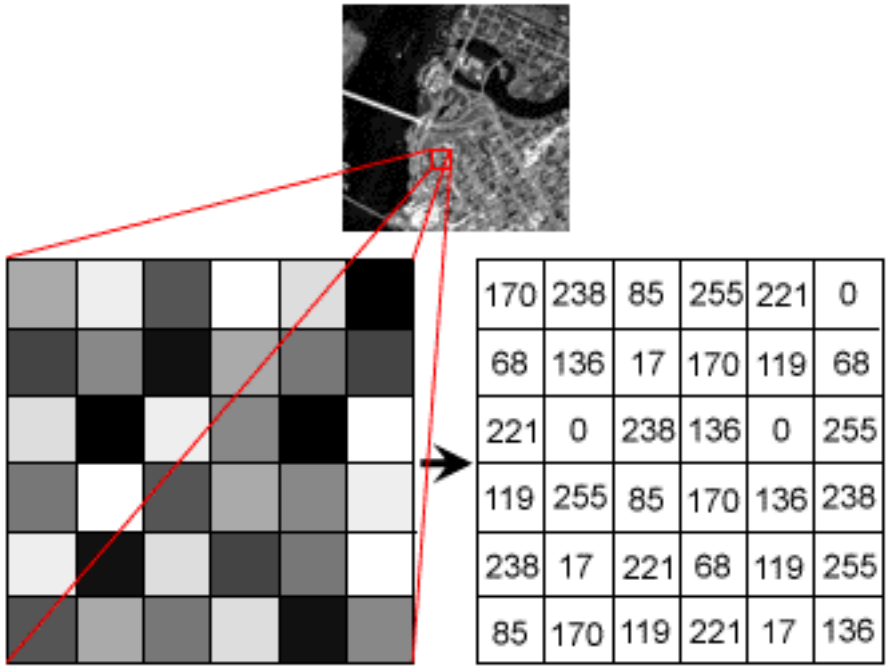
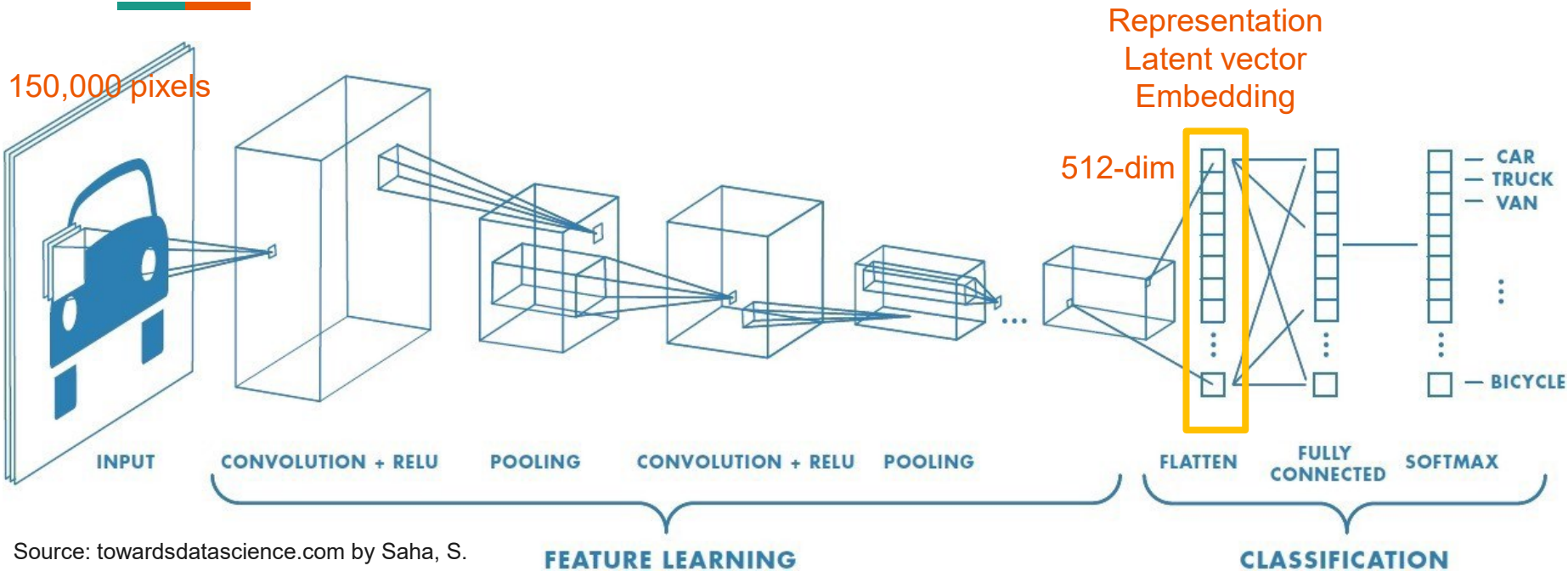


Image from naushardsblog.wordpress.com

Encoder-Decoder view of neural network

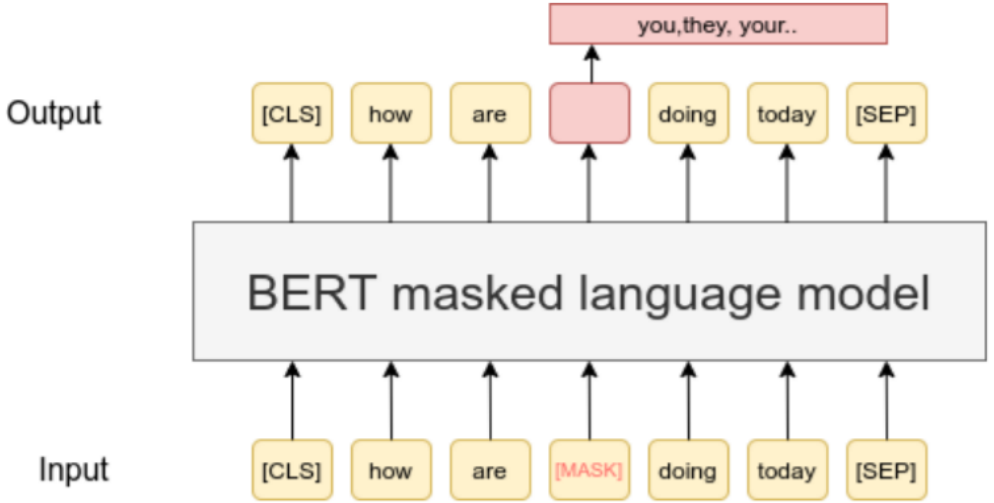


Source: towardsdatascience.com by Saha, S.

- Encode raw data into useful features → decode features for prediction

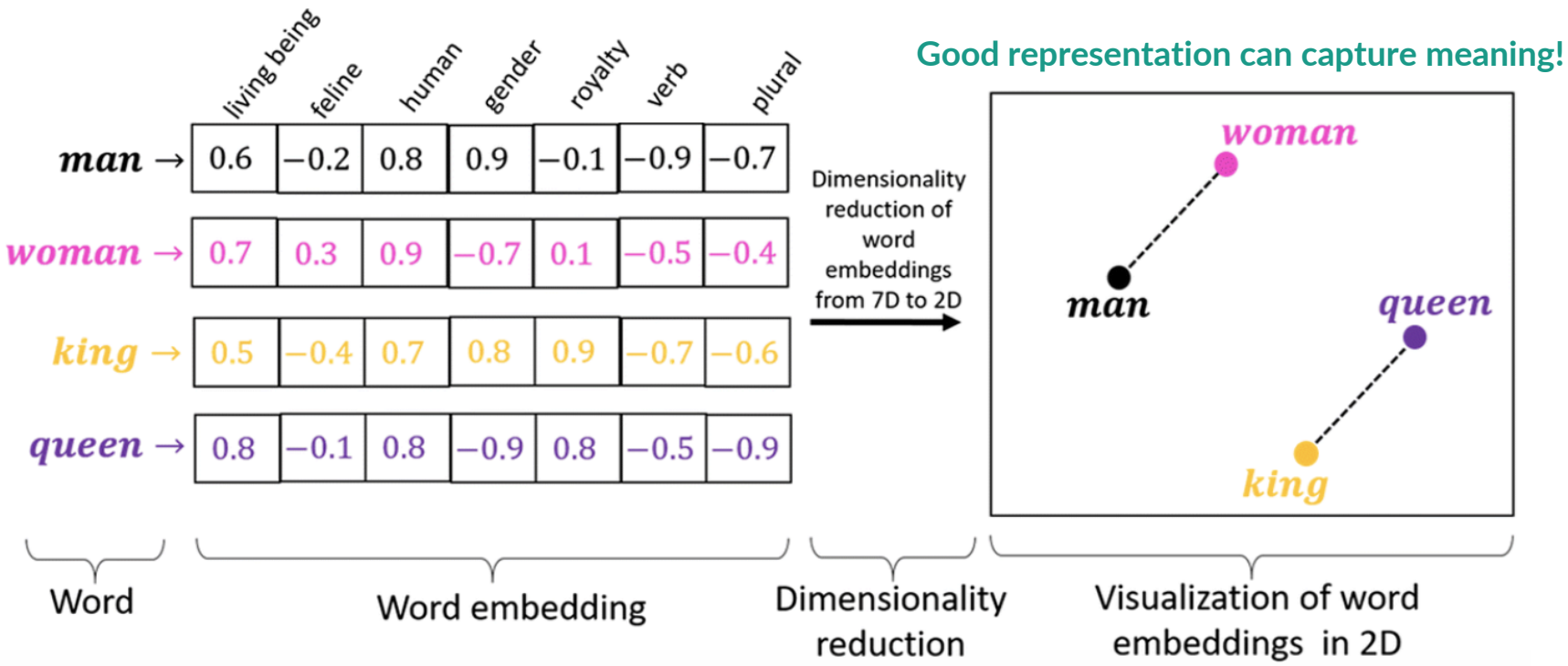
Word representation learning

- Transforming words to vectors (embeddings) that capture some meanings or characteristics
- The model obtains good embeddings by learning to predict the missing words



https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

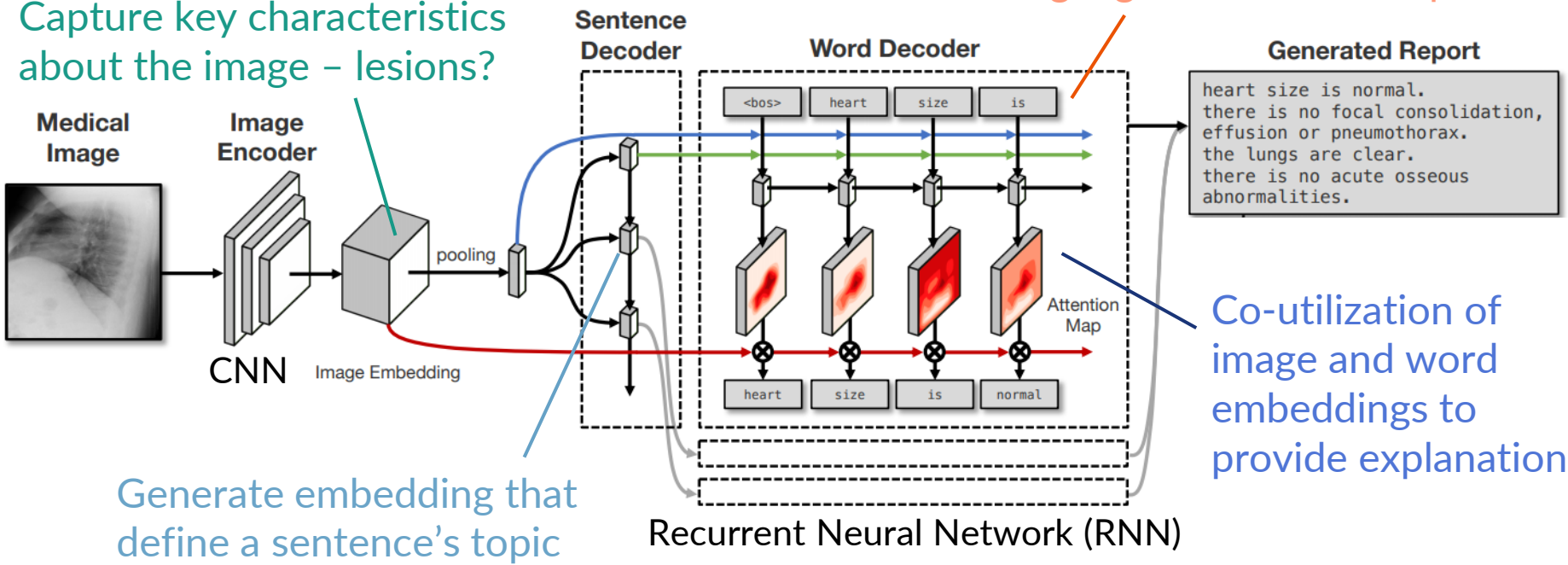
Meaningful word embeddings



Combining image and word embeddings

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

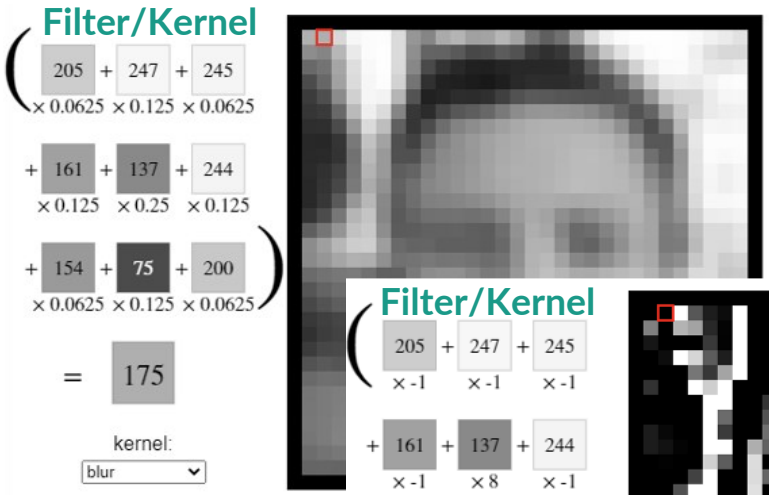
Generate embedding that define a sentence's topic

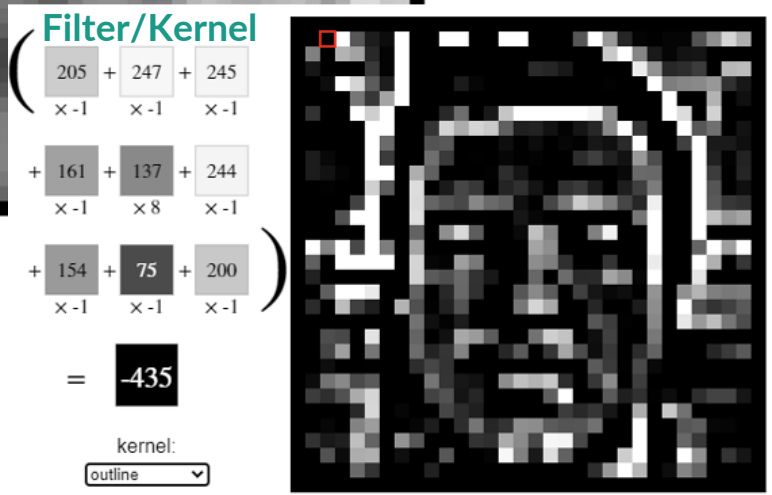
Liu et al. Clinically Accurate Chest X-Ray Report Generation (2019)

Convolutional neural network

Extracting contextual pattern with filter

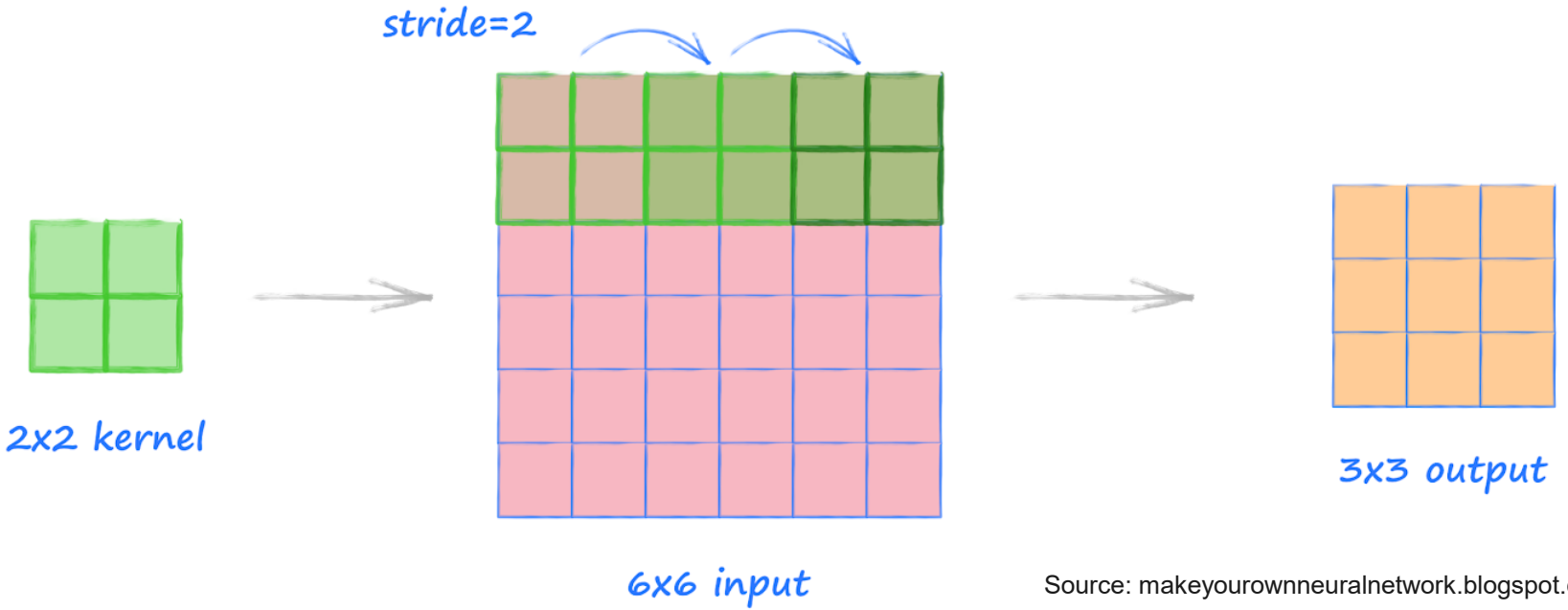
input image





output image

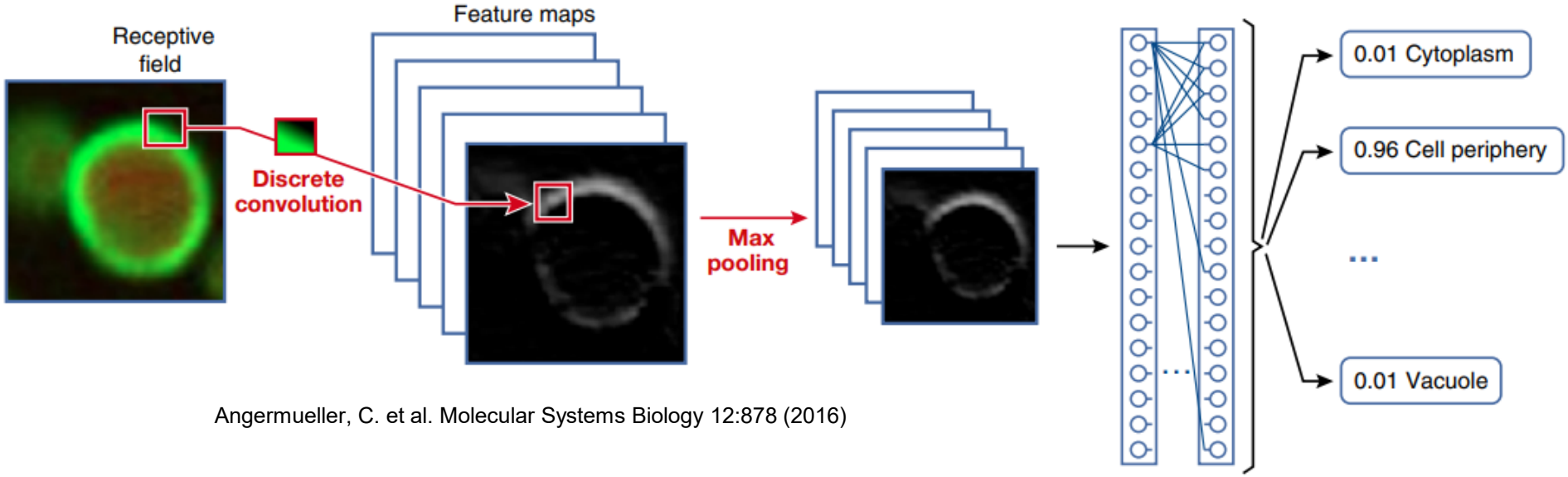
Convolutional operation



Source: makeyourownneuralnetwork.blogspot.com

- Linear combination of values in nearby pixels – applied throughout

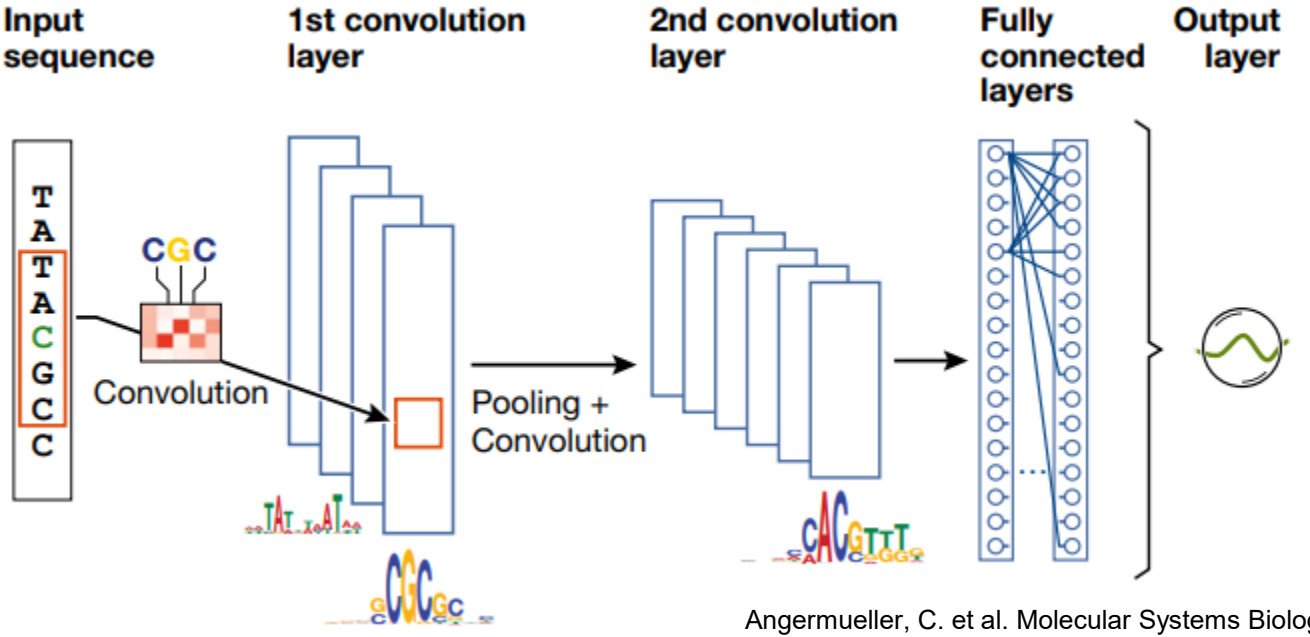
Convolutional neural network (CNN)



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

- Instead of using human-define filters to extract contextual pattern, CNN learns the best filters from the data

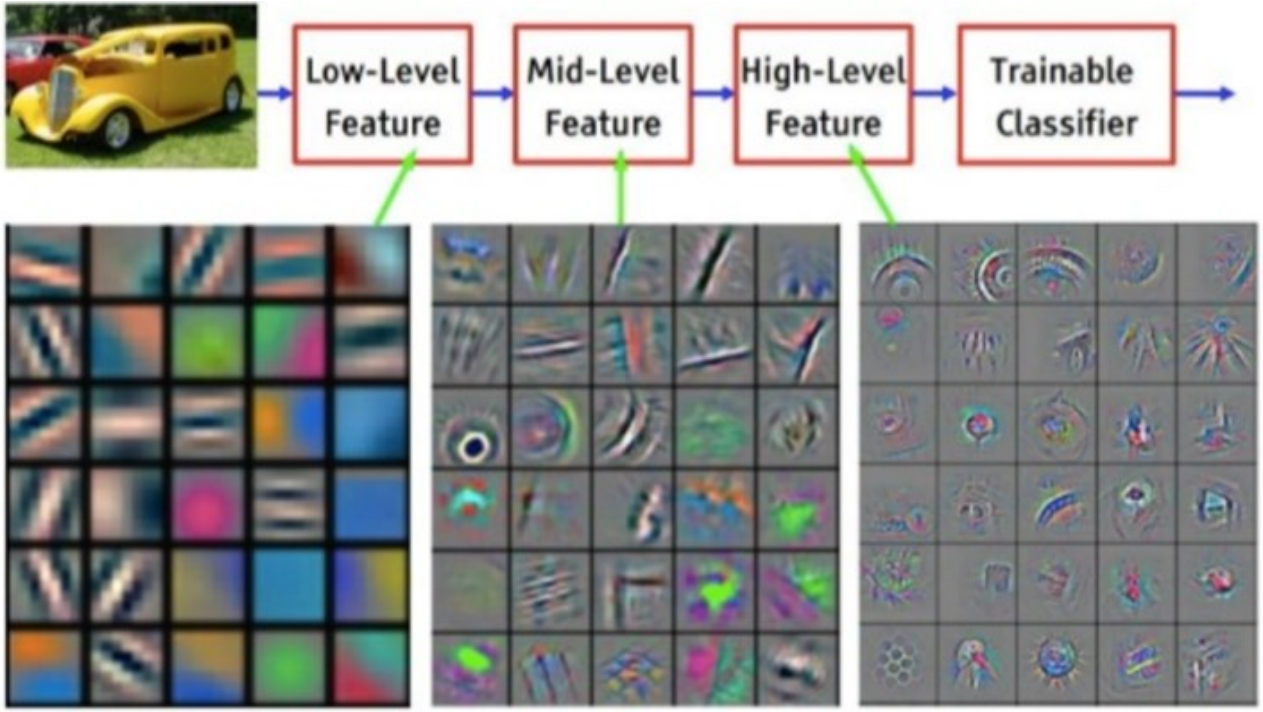
Convolution for DNA sequences



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

- Motif = contextual pattern on DNA sequence

Hierarchical feature assembly inside CNN



Source: Zeiler and Fergus (2013)

Some CNN designs

Vanishing and exploding gradient problems

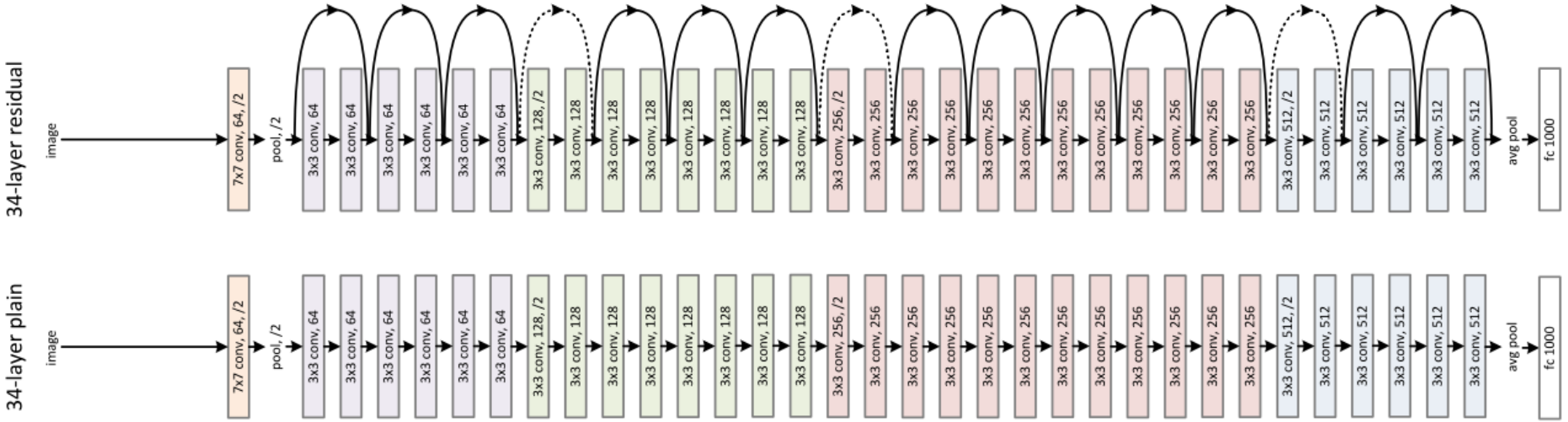
Gradient: $\frac{\delta L}{\delta w_{i,j}} = \frac{\delta L}{\delta f}\frac{\delta f}{\delta h_i}\frac{\delta h_i}{\delta g_i}\frac{\delta g_i}{\delta w_{i,j}} = (f(x) - y) \cdot u_i \cdot g_i(\mathrm{x})(1 - g_i(\mathrm{x}))\, x_j$

The number of multiplicative terms scales with the number of layers

What would happen if all values are << 1 or >> 1?

- Gradient became very small → No weight update
- Gradient became very large → Unstable

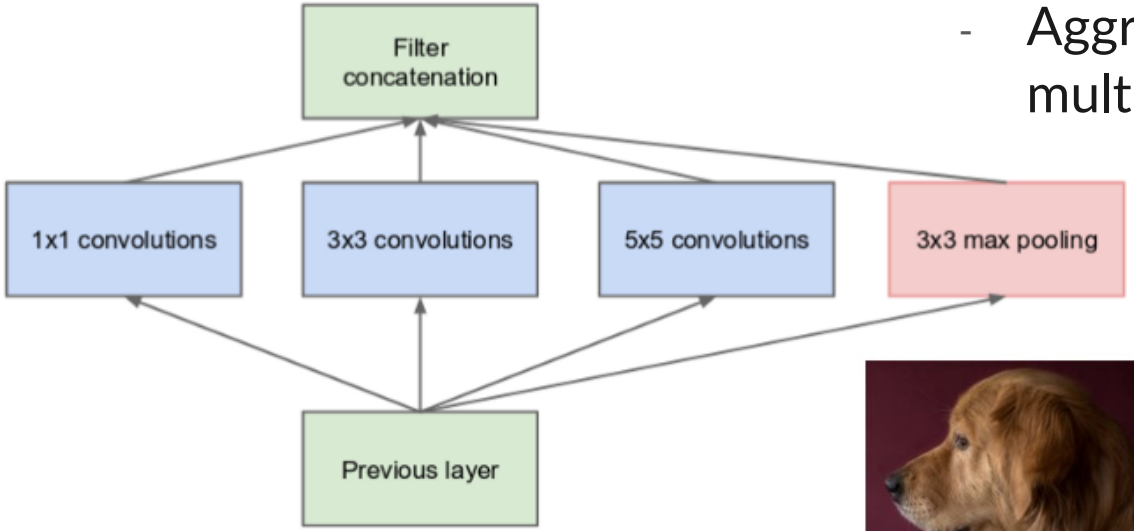
Residual network (ResNet)



Source: medium.com

- Adding skip connections jumping over blocks of convolutional layers
- Reduce the number of terms in gradient of early weights

Inception = multi-resolution layer

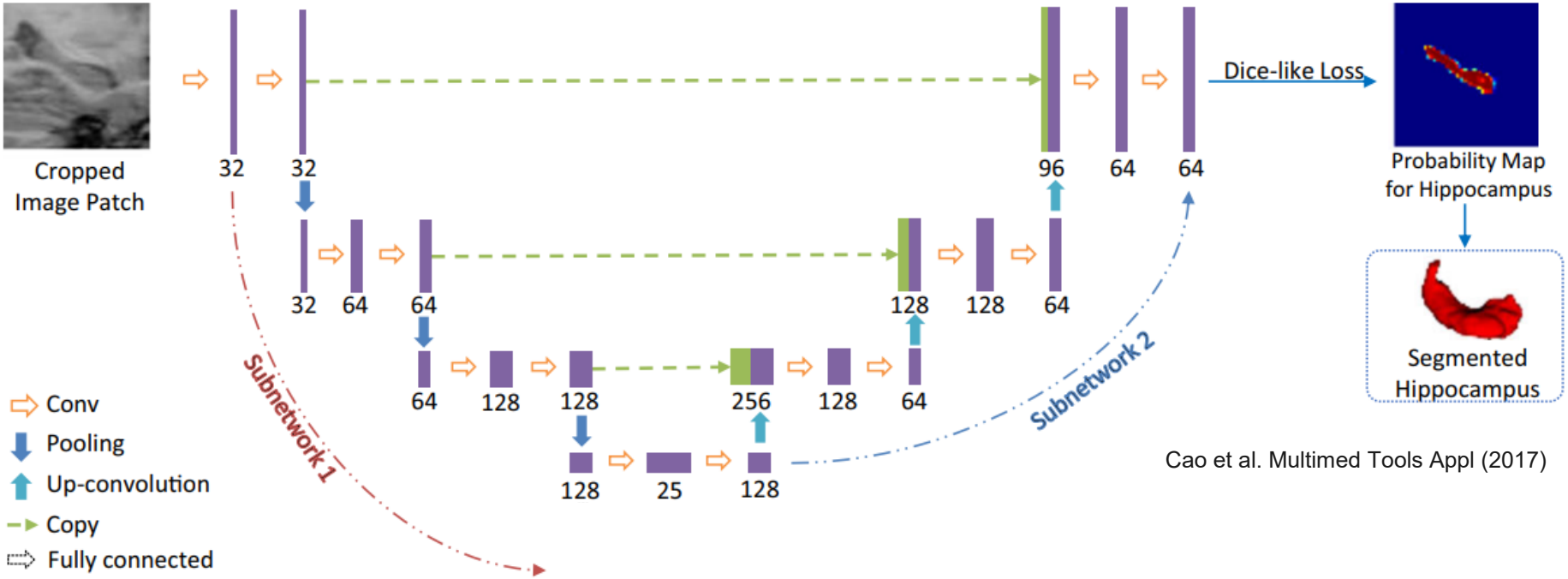


- Aggregate information from multiple resolutions together

- Convolution with multiple filter sizes per layer

From left: A dog occupying most of the image, a dog occupying a part of it, and a dog occupying very little space (Images obtained from Unsplash).

U Net = producing image from image

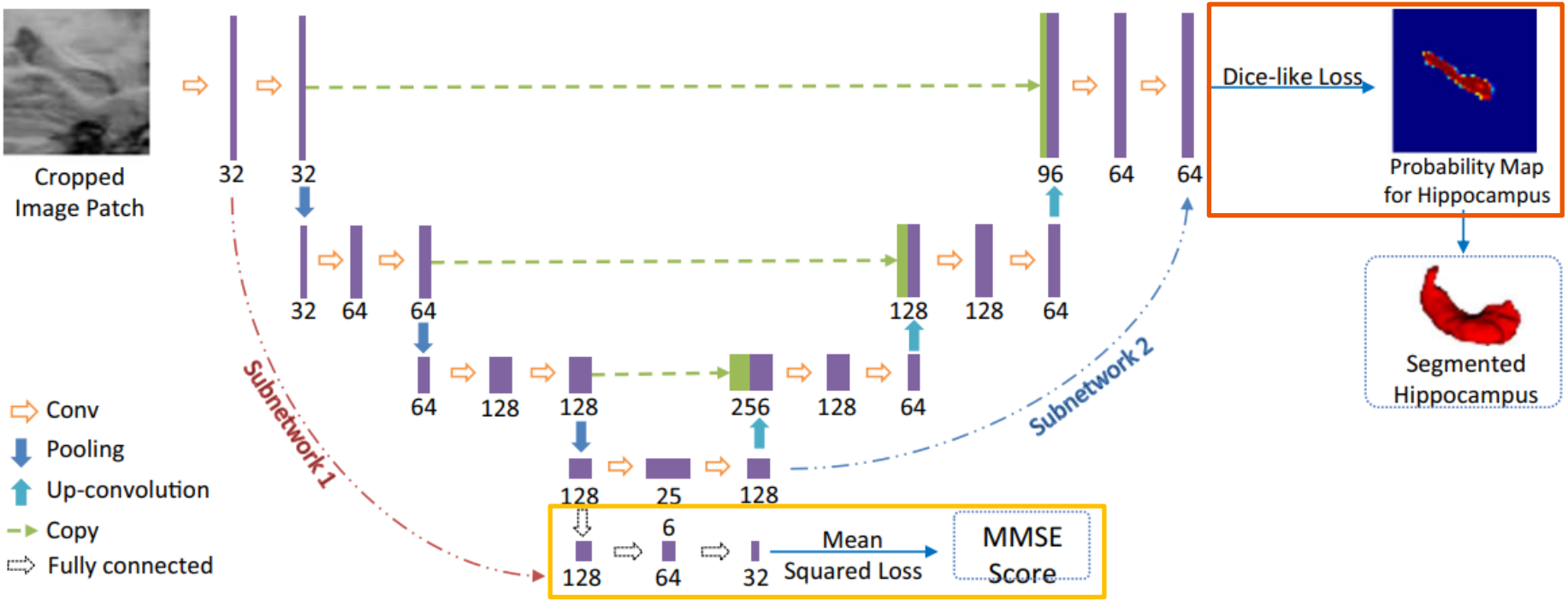


Cao et al. Multimed Tools Appl (2017)

- Make prediction for every pixel → output size = input size

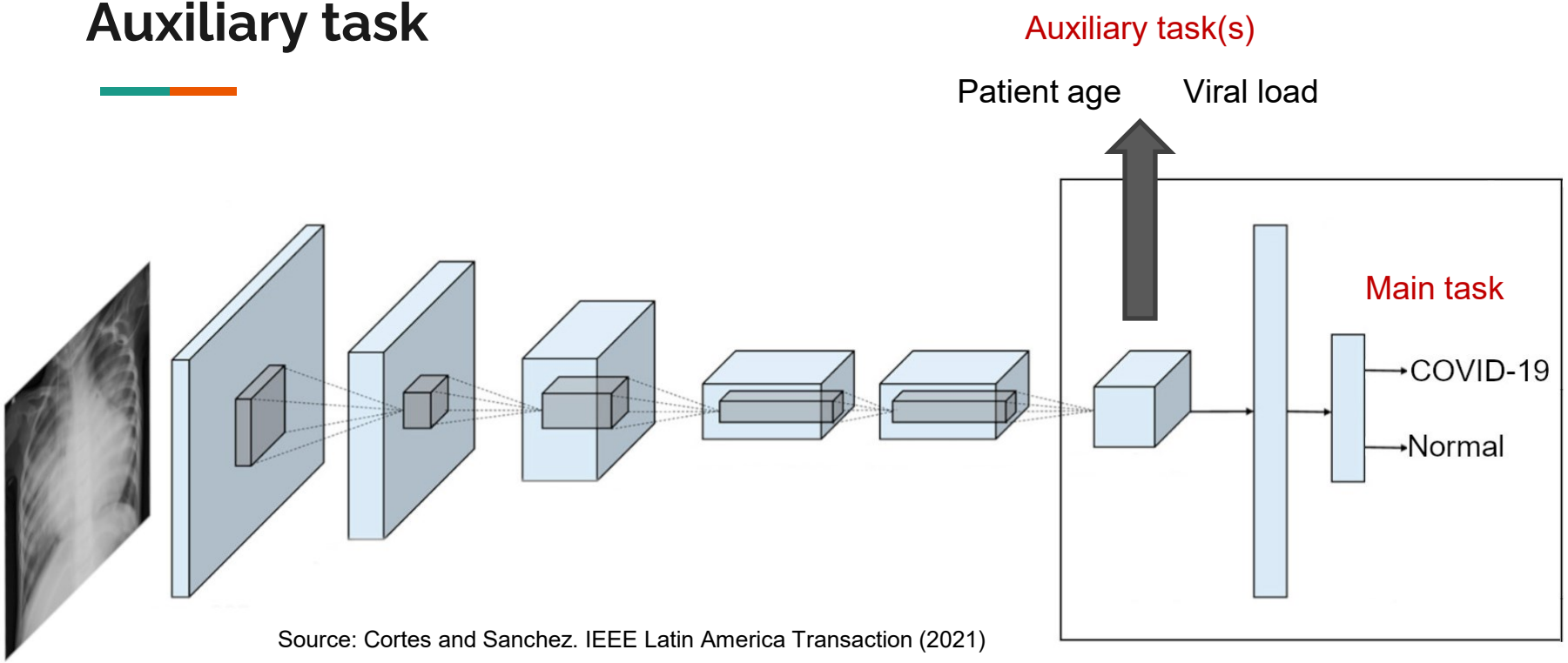
Multitasking

Simultaneous segmentation & classification



- Combine gradients from both tasks

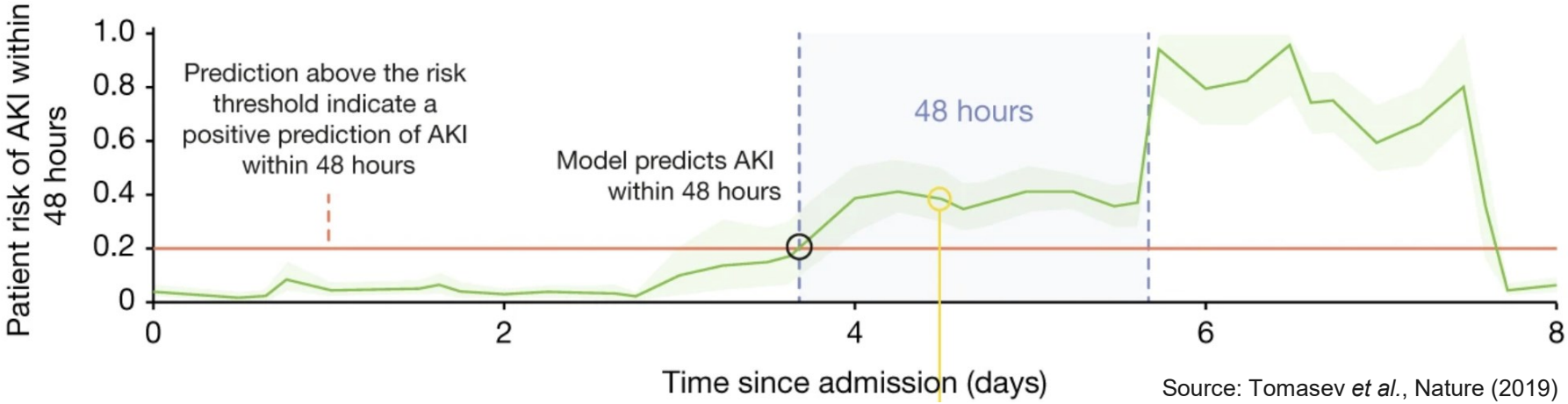
Auxiliary task



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

- Encourage the learned representation to include more information

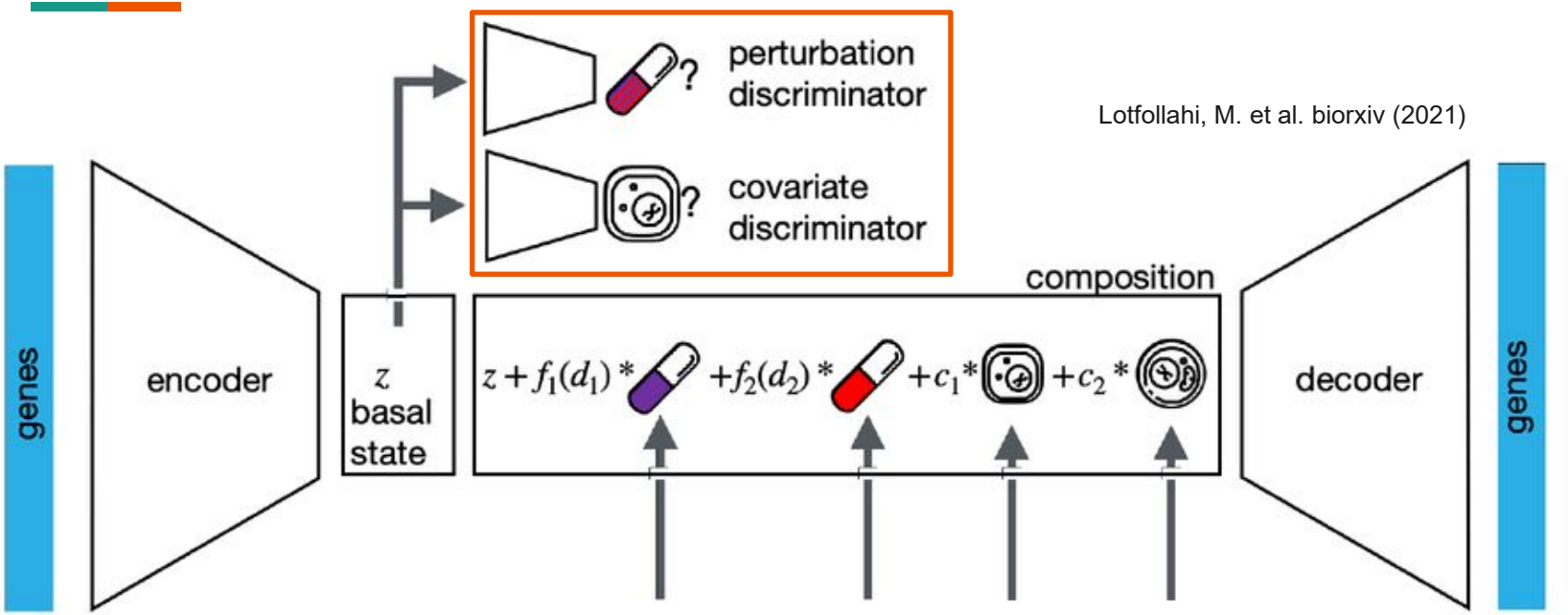
Acute kidney injury prediction



Source: Tomasev *et al.*, Nature (2019)

- **Main task**: Occurrence of acute kidney injury within 48 hours
- **Auxiliary tasks**: Maximal values of 7 key lab tests within 48 hours
  - Provide more feedback on what the model gets wrong

Decoupling / debiasing

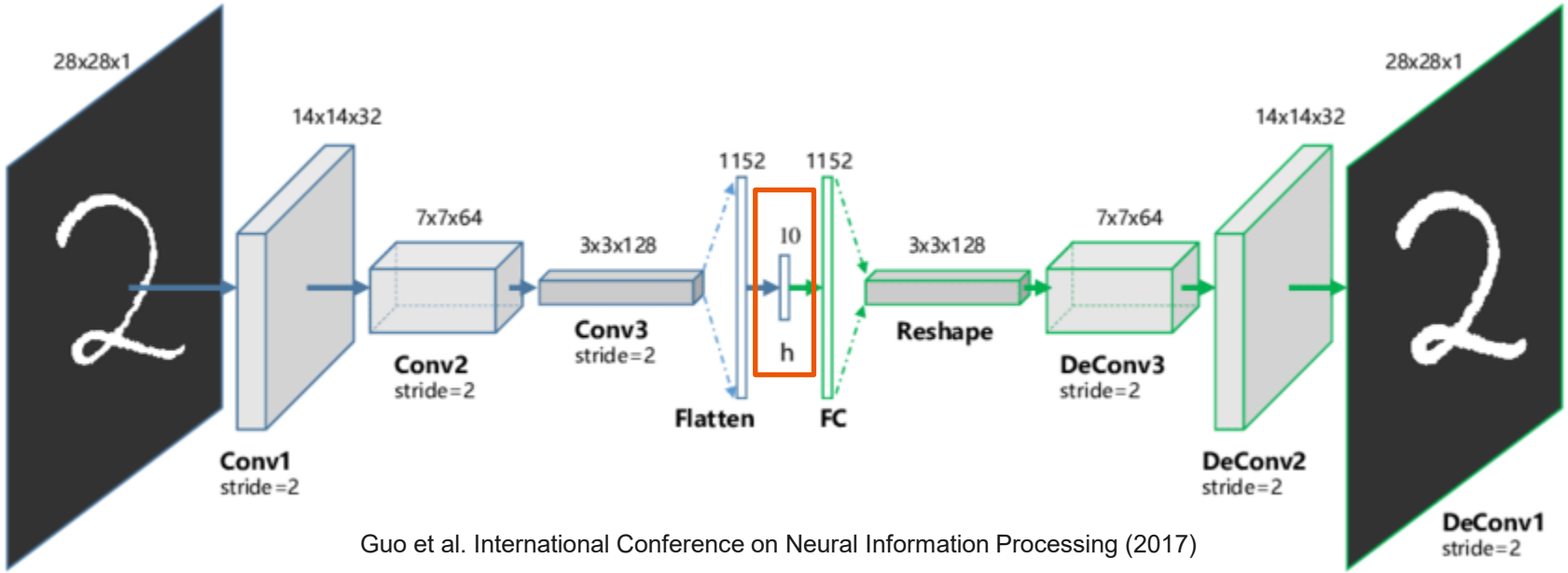


Lotfollahi, M. et al. biorxiv (2021)

- Deconvolute cell basal state from perturbation and covariate
- Update weights in the opposite direction of gradient

Autoencoder

Representation learning via self-reconstruction



Guo et al. International Conference on Neural Information Processing (2017)

- Similar to dimensionality reduction

Denoising autoencoder

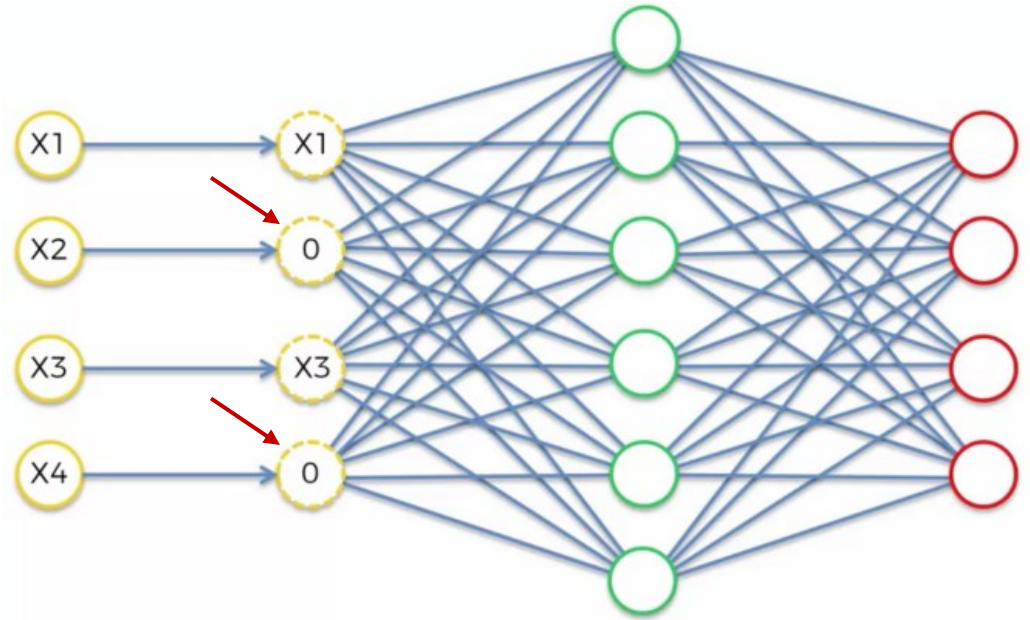


Image from towardsdatascience.com/denoising-autoencoders-explained-dbb82467fc2

- Randomly set some inputs to zero → robust representation

Variational autoencoder (VAE)

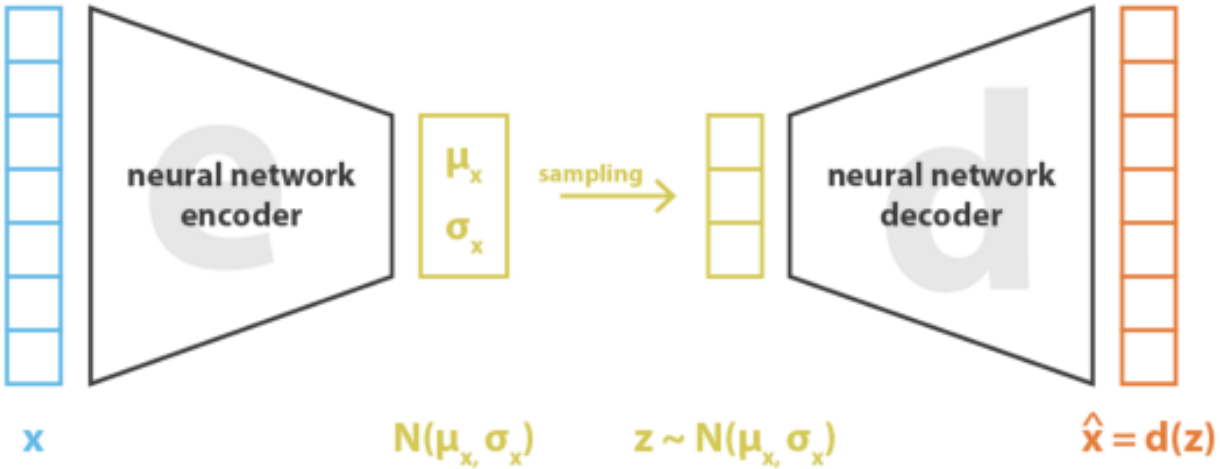
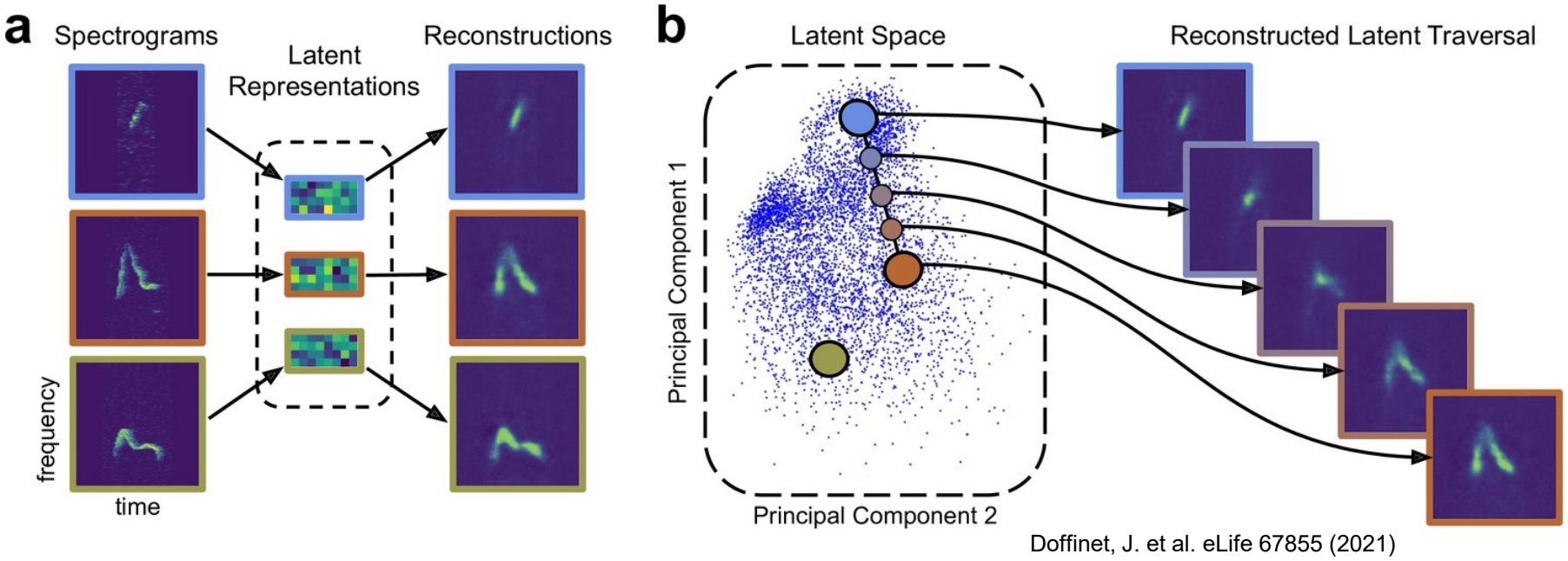


Image from www.jeremyjordan.me/variational-autoencoders/

- Learned representation = parameters for distribution
- Decoder is robust to small changes in the representation
  - Smooth representation space

VAE generates smoother representation space



Doffinet, J. et al. eLife 67855 (2021)

- VAE learn representation distribution, not just individual vectors

Generative model

Why generative model?

FAKE

REAL

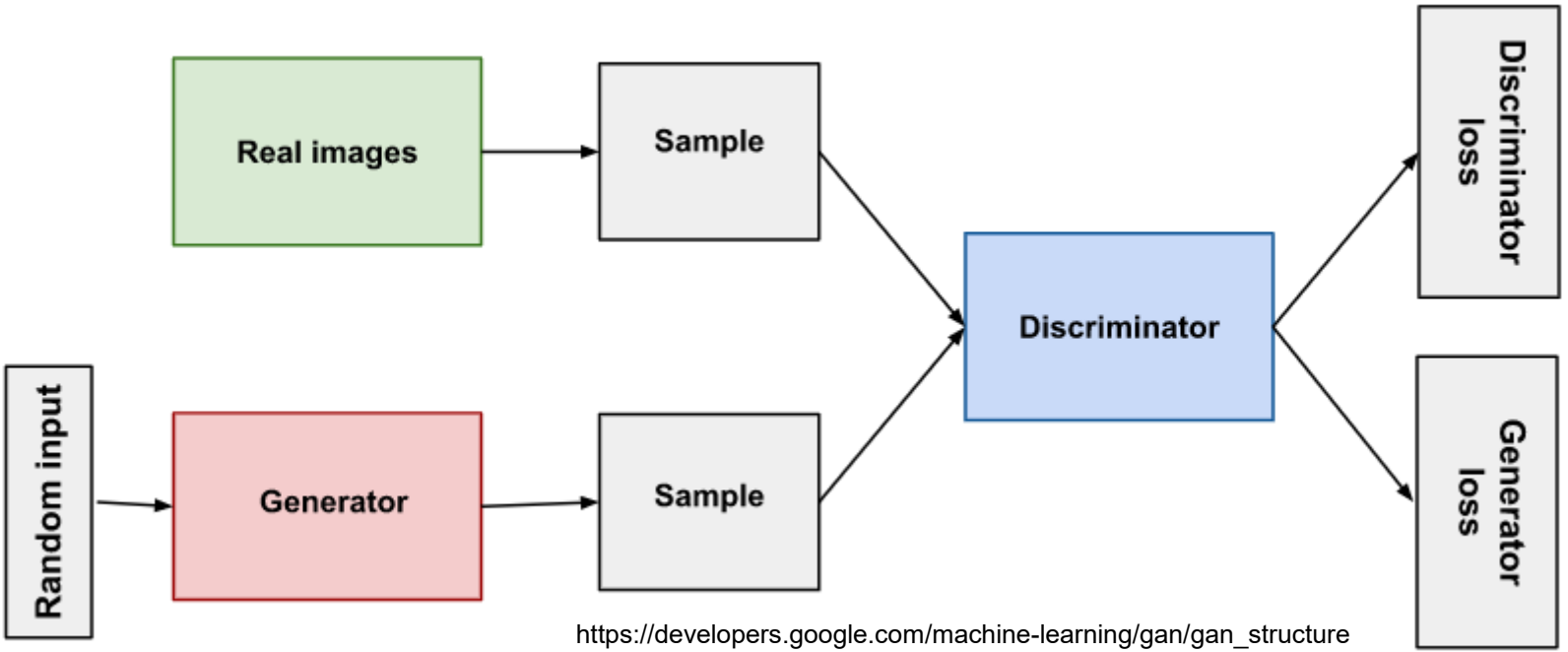
https://developers.google.com/machine-learning/gan/gan_structure

FAKE

REAL

- Models that generate realistic data can tell us about the underlying mechanisms of the system

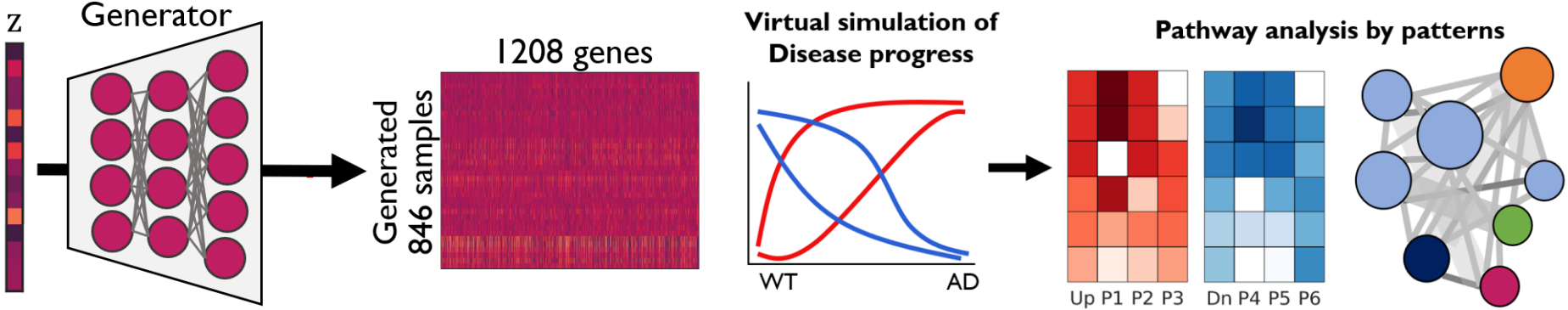
Generative adversarial network (GAN)



https://developers.google.com/machine-learning/gan/gan_structure

- Simultaneous training of generator and discriminator

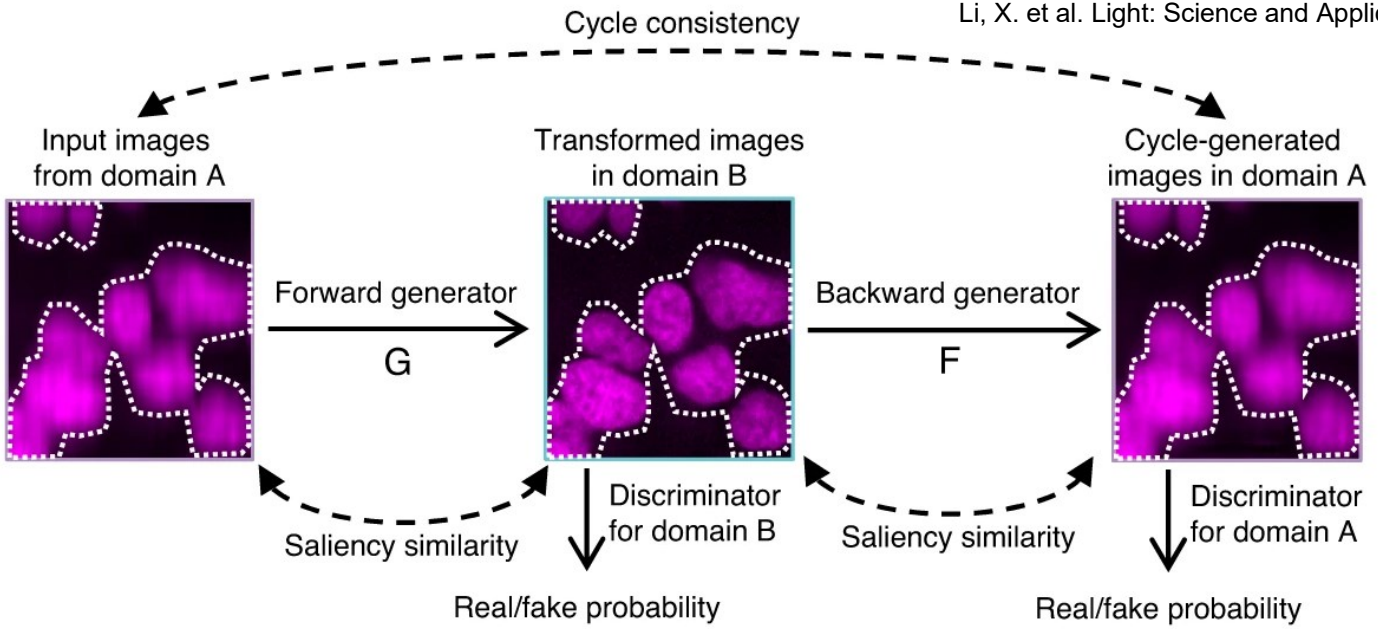
Knowledge from simulated data



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

- Train a generator with data from small-scale experiment
- Simulate time-course gene expression profiles
- Perform usual bioinformatics analyses to infer biological knowledge

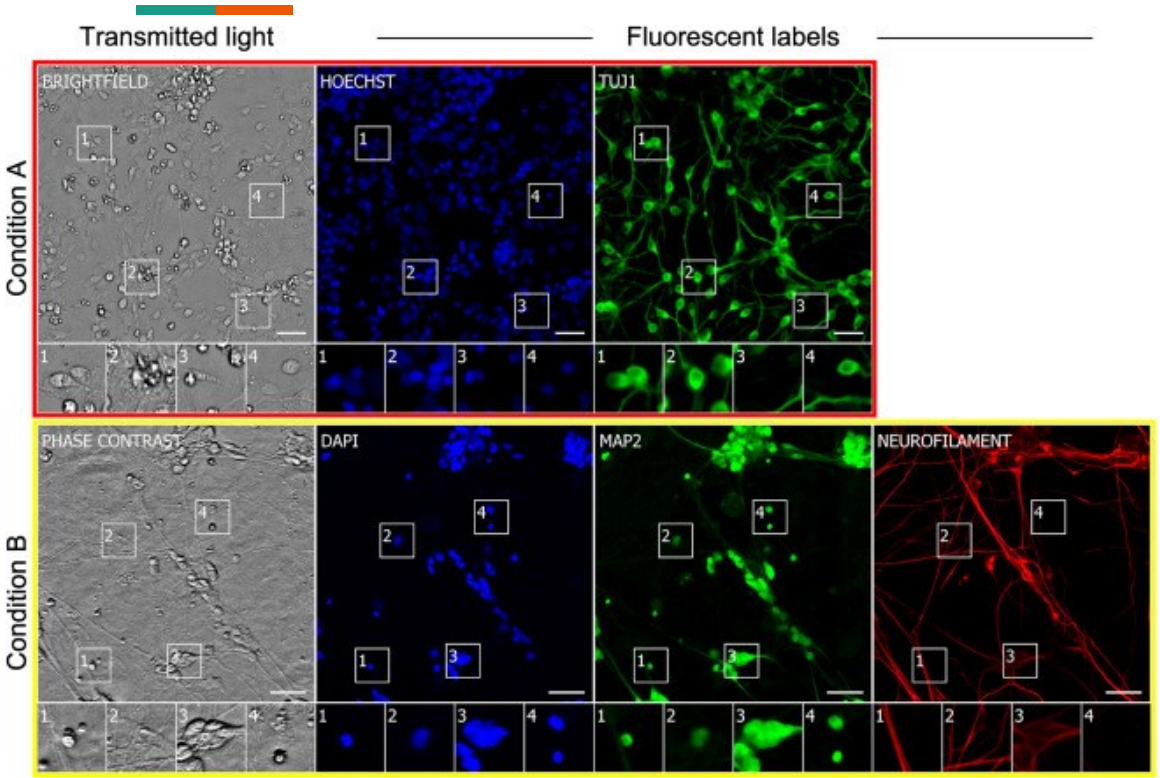
Cycle GAN for transforming image



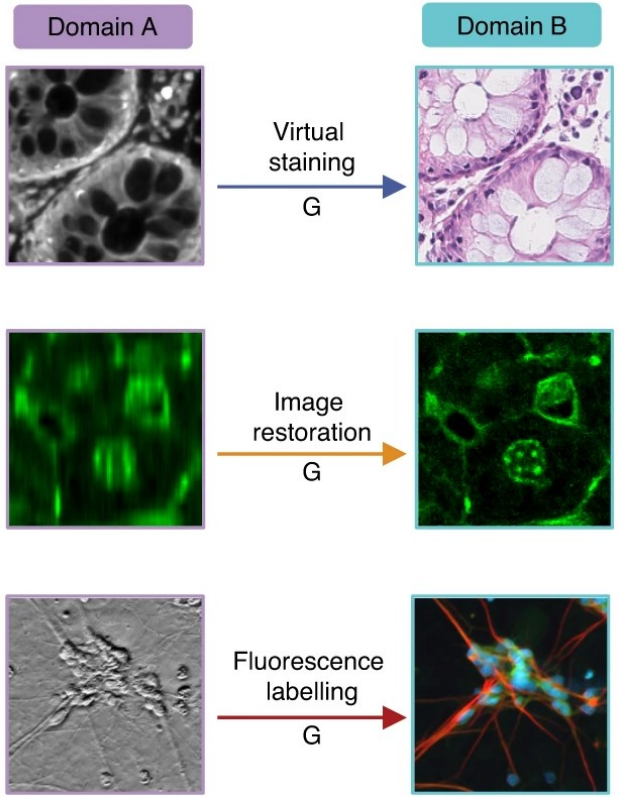
Li, X. et al. Light: Science and Applications10:44 (2021)

- Generate sharpened image from blurry image and back

Virtual staining



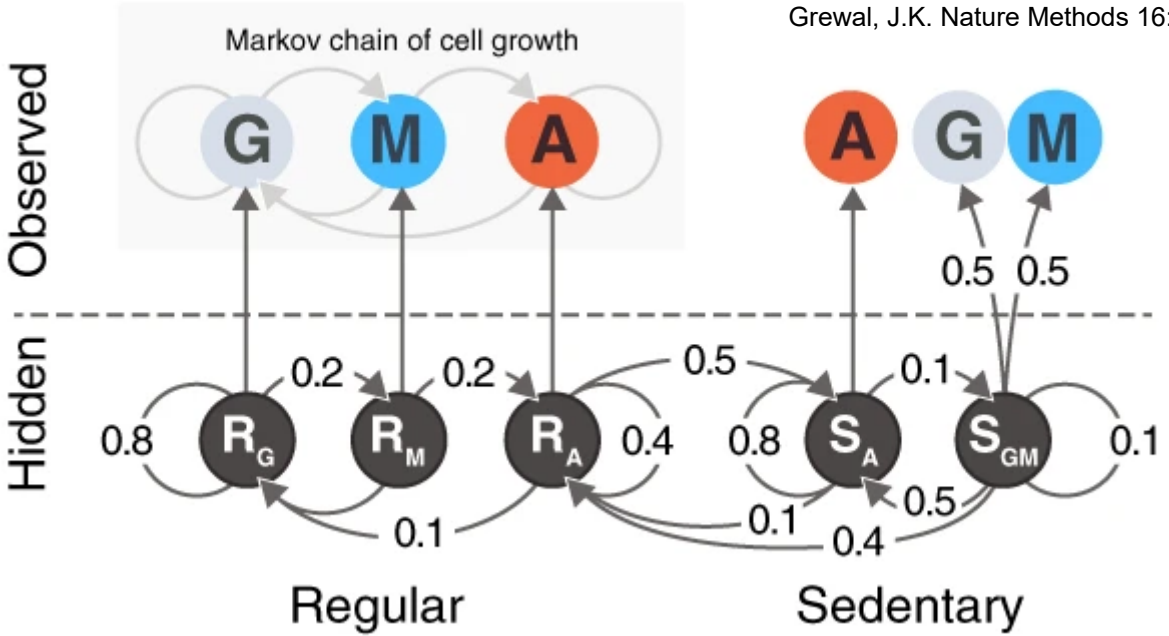
Christiansen, E.M. et al. Cell 173:792-803.e19 (2018)



Li, X. et al. Light: Science and Applications10:44 (2021)

Recurrent neural network

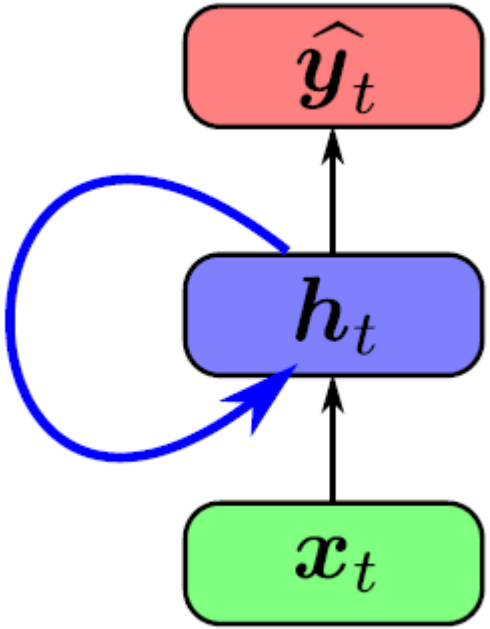
Hidden Markov Model



Grewal, J.K. Nature Methods 16:795-796 (2019)

- Sequence of observations, each generated from a model

Recurrent neural network



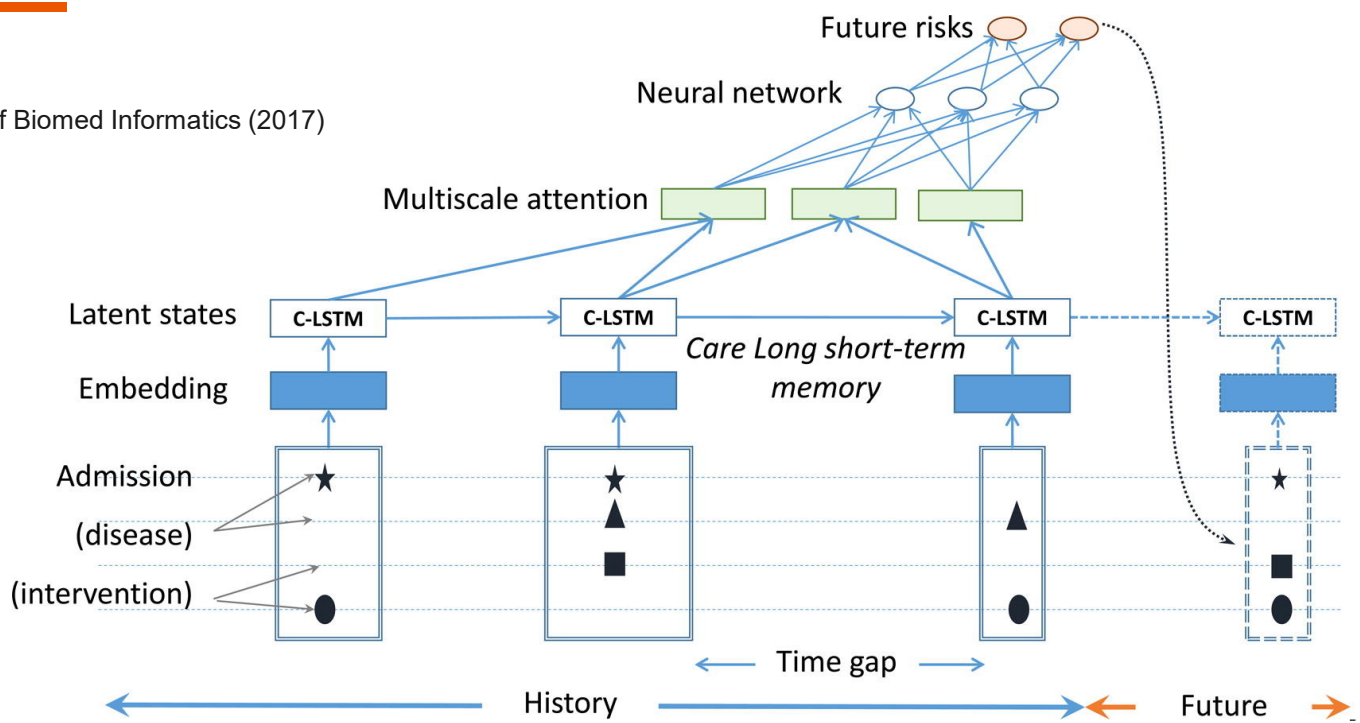
Shared weights!

$$h_1 = f(\boldsymbol{u} \cdot x_1 + \boldsymbol{v} \cdot h_0 + c)$$
$$h_2 = f(\boldsymbol{u} \cdot x_2 + \boldsymbol{v} \cdot h_1 + c)$$
$$\dots$$
$$h_t = f(\boldsymbol{u} \cdot x_t + \boldsymbol{v} \cdot h_{t-1} + c)$$
$$\hat{y}_t = \boldsymbol{w} \cdot h_t + b$$

- Reuse a single layer (weights) over time with different input

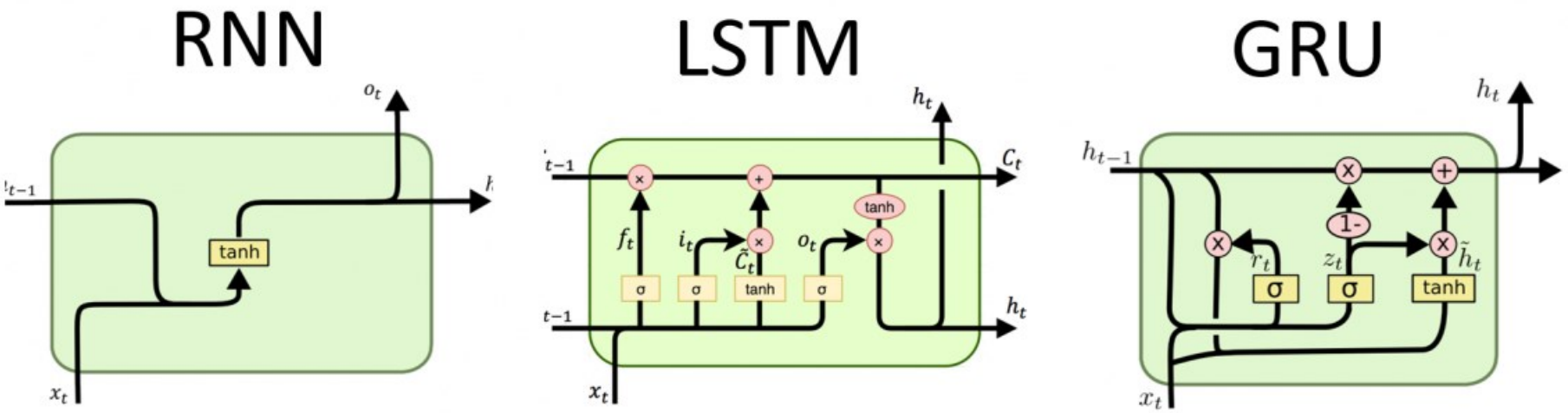
RNN on medical history

Pham et al. J of Biomed Informatics (2017)



- Aggregate information across time to make prediction

RNN architecture



Source: www.linkedin.com/pulse/recurrent-neural-networks-rnn-gated-units-gru-long-short-robin-kalia

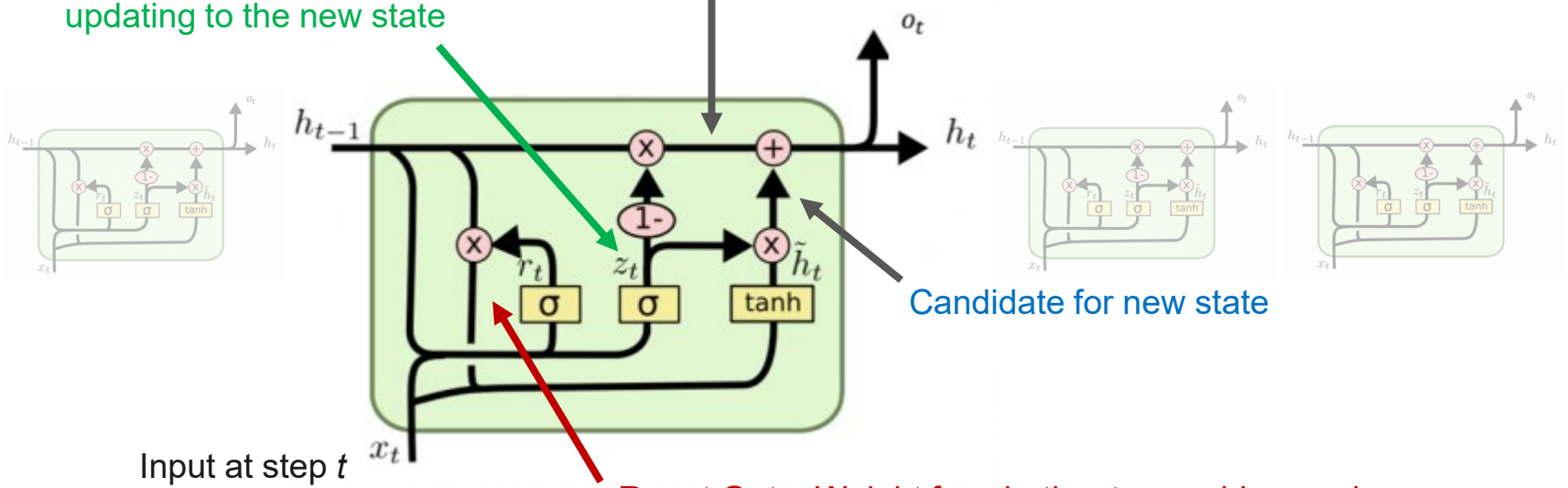
- Allow the model to retain / forget information from earlier time points
- Include shortcuts for gradient calculation – similar to ResNet

Gated recurrent unit (GRU)

Update Gate: Weight for keeping previous state or updating to the new state

Previous state

Output at step *t*

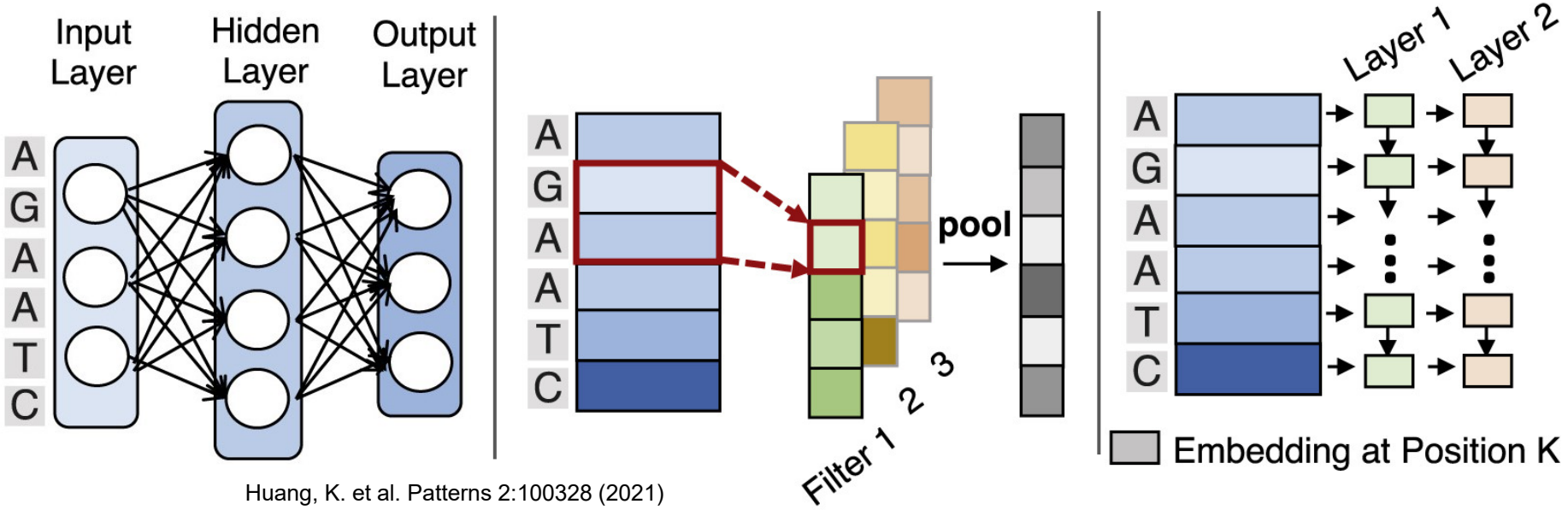


Candidate for new state

Input at step *t*

Reset Gate: Weight for whether to consider previous state when proposing the new state

Picking the right model

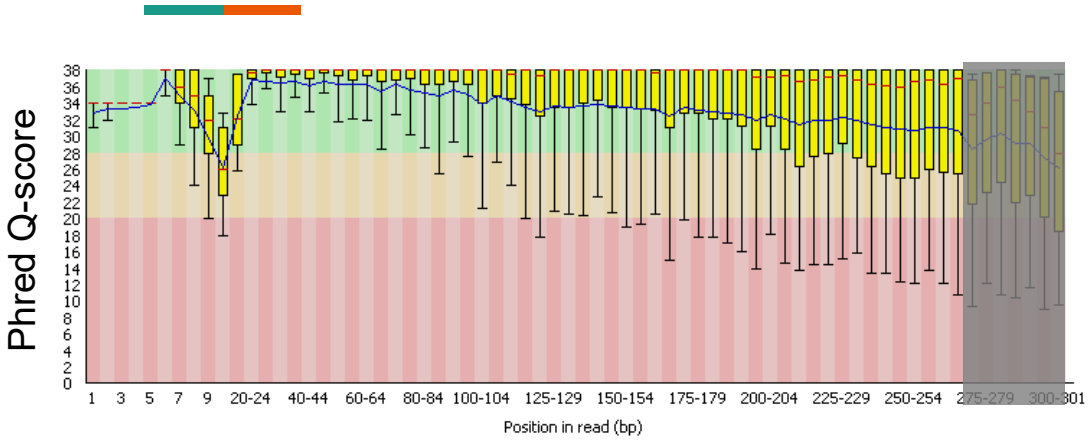


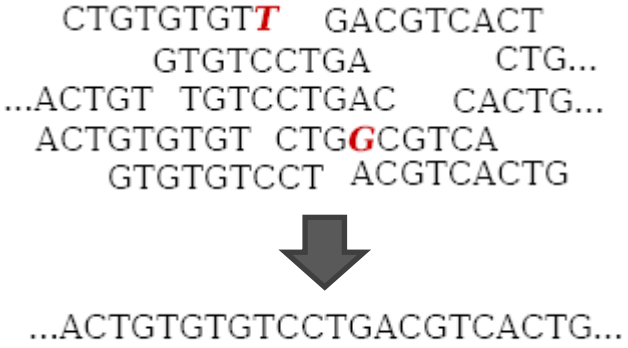
Huang, K. et al. Patterns 2:100328 (2021)

- Choosing the “right” model depends on the interpretation of the task and the underlying mechanisms – require domain knowledge

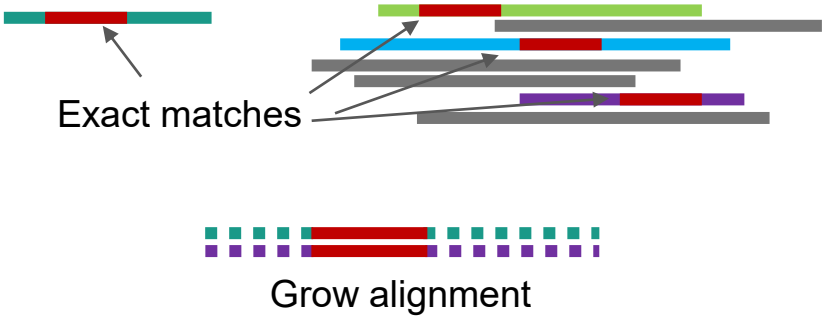
ML-enhanced bioinformatics

Bioinformatics relies on statistics and scoring





Chandra Varma Bogaraju, S. Int J Embed Syst 9:74 (2017)



- Instead of applying hand-made scoring + cutoffs, ANN model can be trained to predict the outcome directly

From bioinformatics to deep learning

Published: 24 September 2018

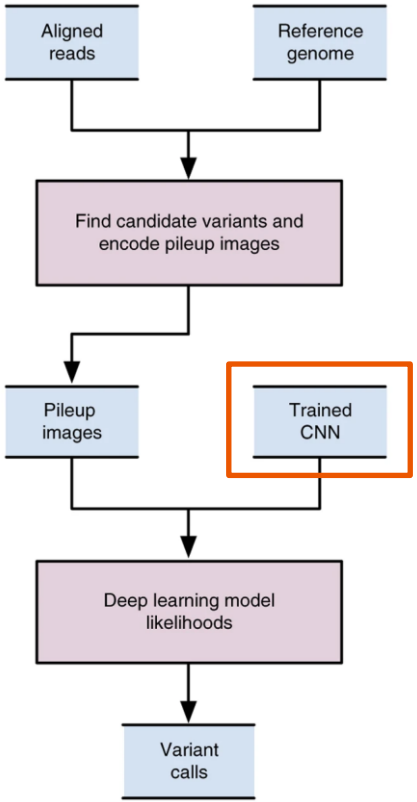
A universal SNP and small-indel variant caller using deep neural networks

Published: 27 July 2015

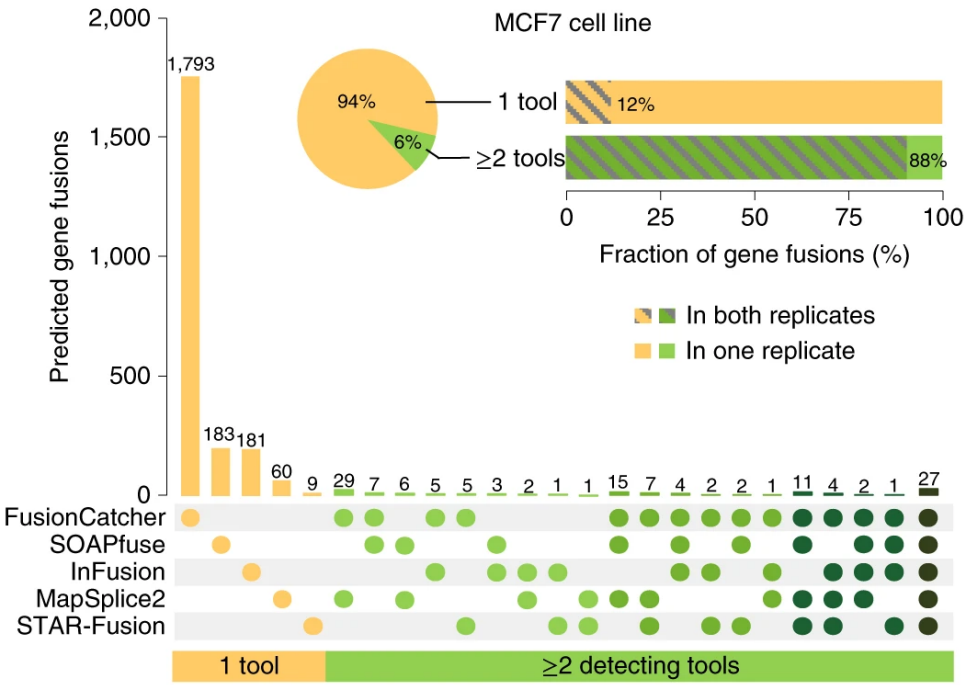
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

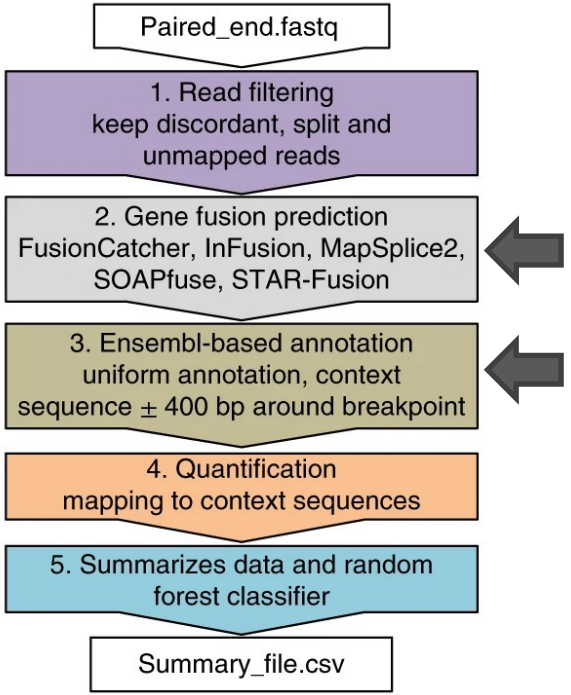
Article | Open Access | Published: 19 May 2022

Prediction of protein–protein interaction using graph neural networks



Aggregate scores from multiple tools





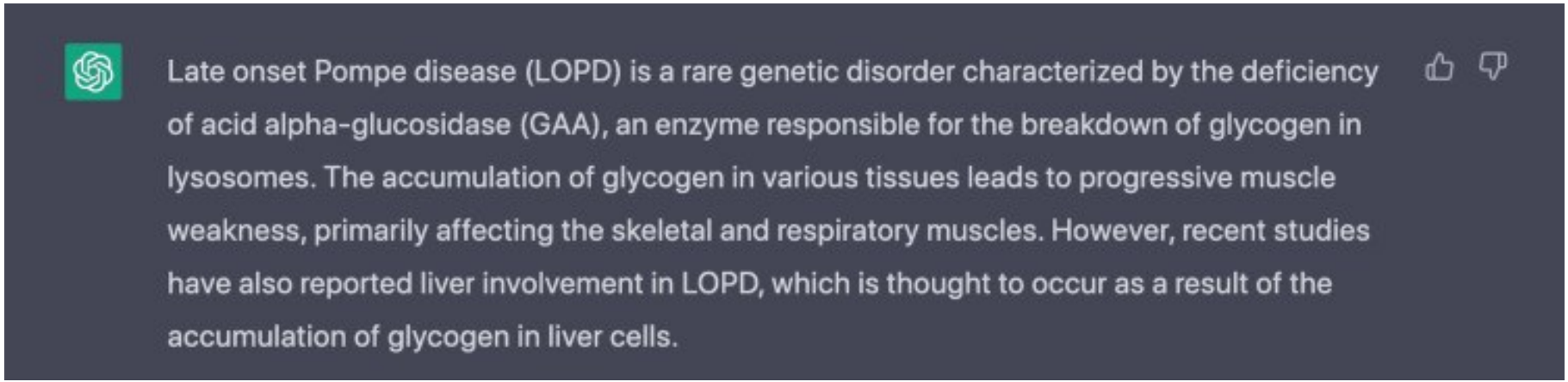
Weber, D. et al. Nature Biotechnology 40:1276-1284 (2022)

Caution when using AI

AI (silently) makes mistakes and biases

But can you spot them?

Alkaissi, H. et al. Cureus 15:e35179 (2023)



- There was no prior publication about liver involvement with LOPD
- However, the authors of this paper have an unpublished manuscript showing a link between liver disease and LOPD
  - *Did ChatGPT just synthesized new knowledge? Or simply hallucinated?*

Huge gap between development and actual use

Healthcare, Law, Regulation, and Policy, Machine Learning

“Flying in the Dark”: Hospital AI Tools Aren’t Well Documented

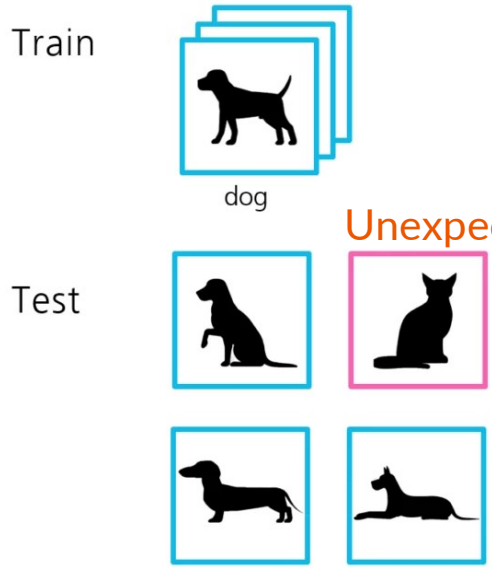
	EPIC MODEL BRIEFS											
MODEL REPORTING GUIDELINES	Deterioration Index	Early Detection of Sepsis	Risk of Unplanned Readmission	Risk of Patient No-Show	Pediatric Risk of Hospital Admission or ED Visit	Risk of Hospital Admission or ED Visit	Inpatient Risk of Falls	Projected Block Utilization	Remaining Length of Stay	Risk of Admission of Heart Failure	Risk of Hospital Admission or ED Visit for Asthma	Risk of Hypertension
TRIPOD	63%	63%	61%	48%	42%	61%	47%	36%	55%	48%	44%	51%
CONSORT-AI	63%	43%	63%	60%	33%	67%	53%	47%	47%	49%	42%	51%
SPIRIT-AI	61%	55%	54%	54%	38%	61%	44%	49%	51%	41%	39%	46%
Trust and Value	46%	33%	39%	50%	29%	42%	38%	46%	46%	25%	33%	46%
ML Test Score	27%	15%	33%	24%	9%	33%	15%	6%	18%	12%	9%	15%

Evaluation of sepsis diagnosis AI

Results We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

- AUC of 0.63 in practice
- Missed 67% of sepsis

Unexpected behaviors



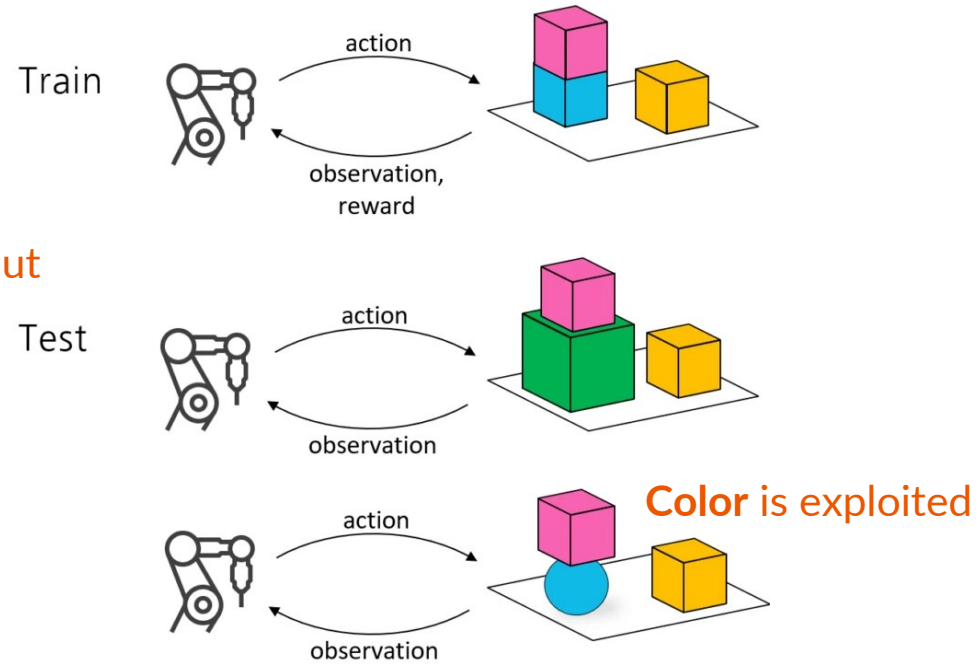
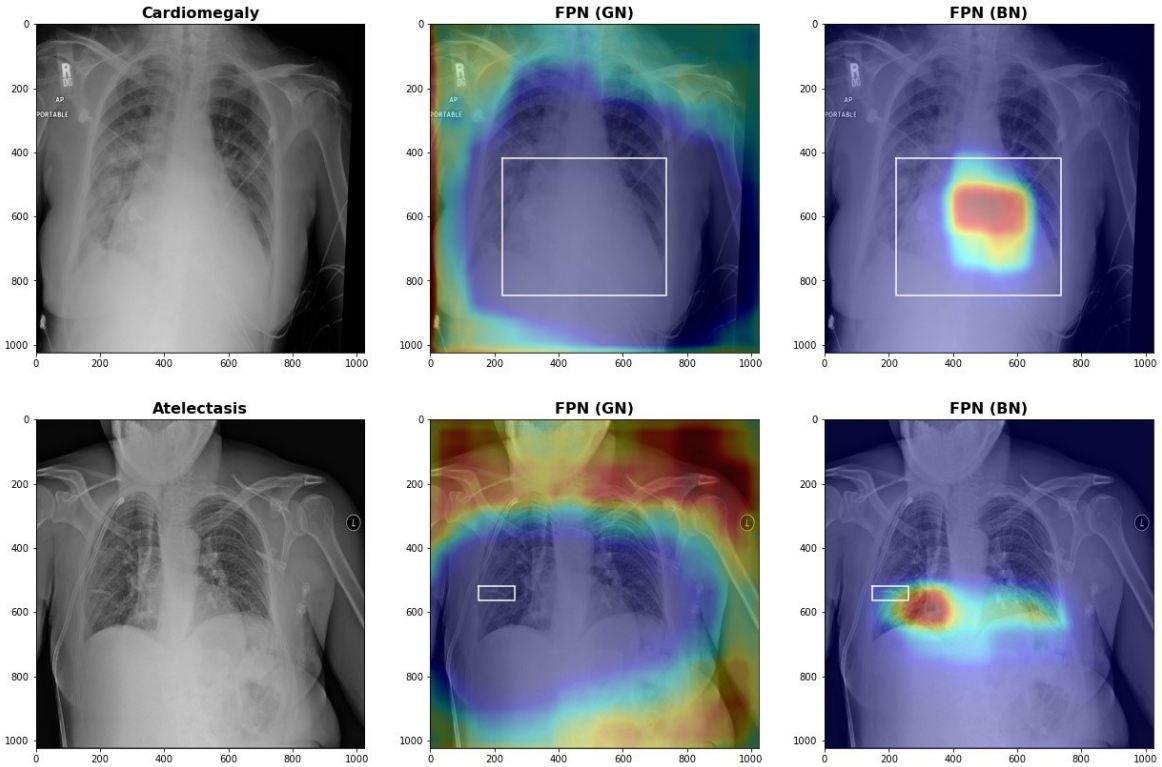


Image from https://safe-intelligence.fraunhofer.de/en/articles/out-of-distribution-detection-for-reinforcement-learning

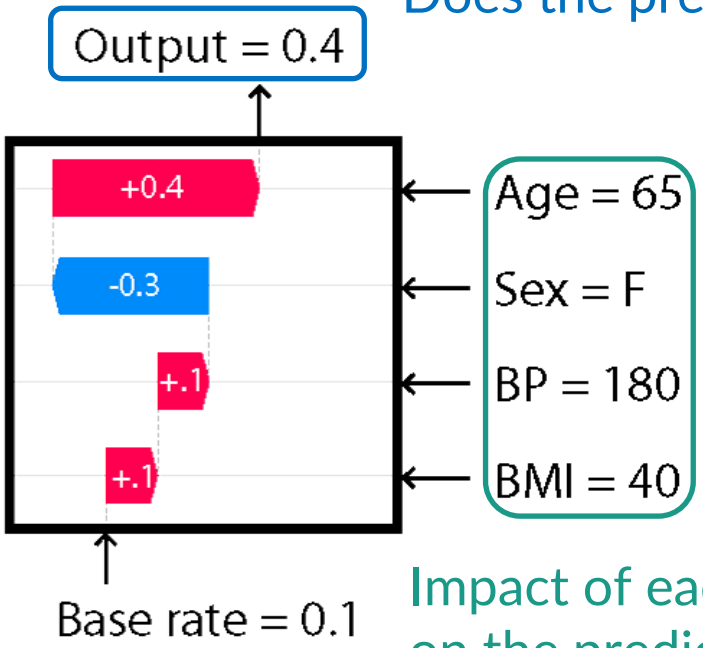
Correct prediction is not enough



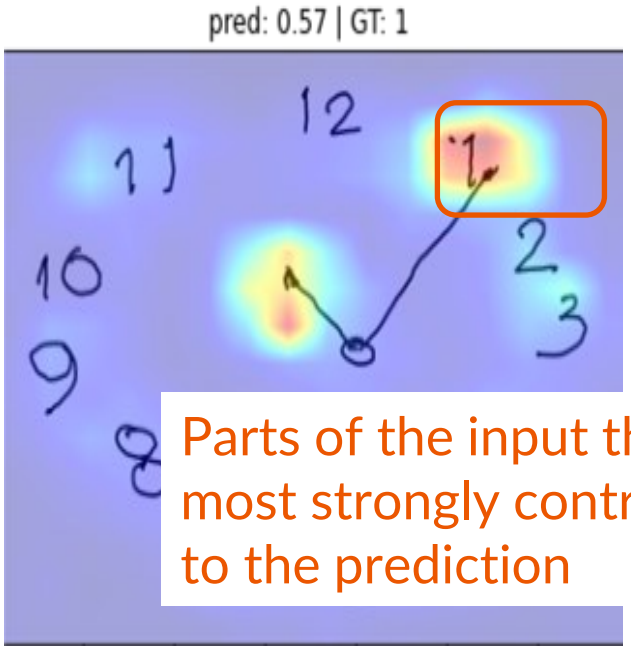
- Two models with the **same classification performance**
- Both images were **correctly classified**
- But the **explanations** complete differ

Explainability

Does the predicted confidence match your expectation?



Impact of each input on the prediction



Parts of the input that most strongly contribute to the prediction

And that's the end of the course!

- Bookmark my lab GitHub (https://github.com/cmb-chula) and website (https://cmb.md.chula.ac.th/) for more courses in the future

- Explore MIT OpenCourseWare
  - 6.0001 / 6.0002 for more Python and computational thinking
  - 7.91 for more rigorous bioinformatics
  - 8.591 for systems biology (dynamics modeling)
  - https://mit6874.github.io/ for deep learning in life sciences

- Stanford AI (https://ai.stanford.edu/courses/)
  - Find videos on YouTube based on course numbers