جامعة الحسين التقنية
Al Hussein Technical University

Center of Professional Development and Community Outreach

Generative AI Training Program - Professional Development (Reskilling) Training Program

Generative AI - The Pre-Transformer Architecture
(A quick overview of probability)

جامعة الحسين التقنية
Al Hussein Technical University

Module 2-A:
The Pre-Transformer Architecture
(A quick overview of probability)

Dr. Abeer Al-Hyari
Abeer.alhyariups@htu.edu.jo

Module-2-A

جامعة الحسين التقنية
Al Hussein Technical University

Module	Module Name and Focus	Key Topics to Cover	PLOs Addressed
Module 2-A	**The Pre-Transformer Architecture (A quick overview of probability)**	• Probability and Likelihood • Bayes Theorem • Architectures for Sequence Modeling (Pre-Transformer) • Architectures for Generation (The Generative Models)	Probability Foundations (Primary), Model Development

Objectives

جامعة الحسين التقنية
Al Hussein Technical University

1. Understand Probability and Likelihood
2. Explain the principles of Bayesian inference and conditional probability.
3. Examine classical sequence modeling architectures such as RNNs, LSTMs, and GRUs.
4. Identify key generative model families: Autoregressive, Variational Autoencoders (VAEs), GANs, flow-based and diffusion models.

Topics Covered

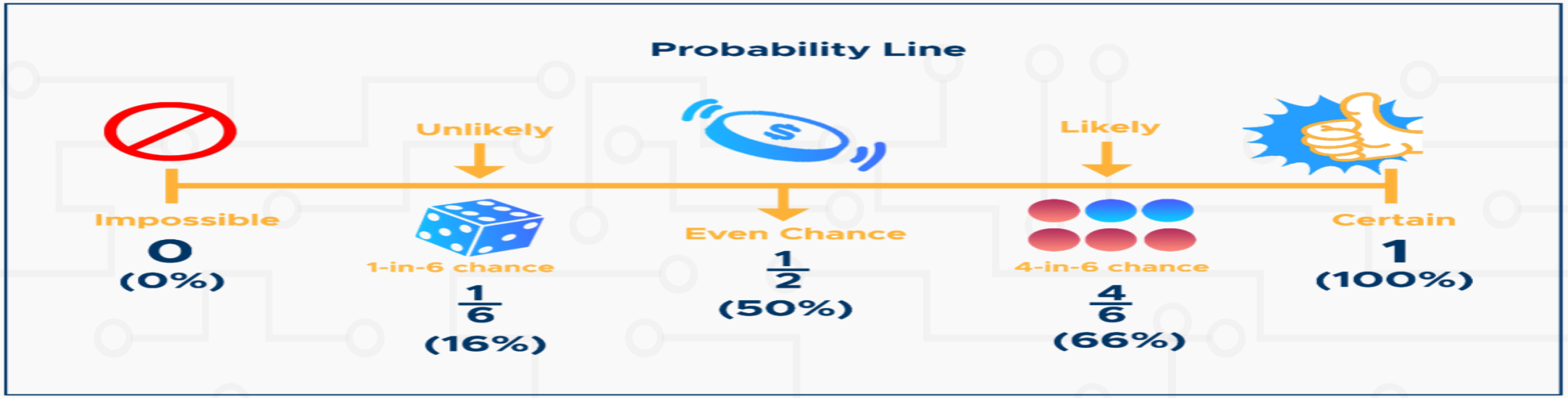
- Probability and Likelihood
- Bayes Theorem
- Architectures for Sequence Modeling (Pre-Transformer)
- Architectures for Generation (The Generative Models)

Probability

جامعة الحسين التقنية
Al Hussein Technical University

Probability is a branch of mathematics and statistics concerning events and numerical descriptions of how likely they are to occur. The probability of an event is a number between 0 and 1; the larger the probability, the more likely an event is to occur.

Probability



https://curvebreakerstestprep.com/wp-content/uploads/2021/04/Probability-Line.png

Bayes Theorem

جامعة الحسين التقنية
Al Hussein Technical University

Bayes' theorem is used in Bayesian methods to update probabilities, which are degrees of belief, after obtaining new data. Given two events A and B, the conditional probability of given that is true is expressed as follows:[*]

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where P(B)≠0. Although Bayes' theorem is a fundamental result of probability theory, it has a specific interpretation in Bayesian statistics.

*: https://en.wikipedia.org/wiki/Bayesian_statistics

Bayes Theorem – Definitions

جامعة الحسين التقنية
Al Hussein Technical University

- **Hypotheses**: refer to possible events or outcomes in the sample space; they are denoted as $\mathbf{E_1, E_2, \ldots, E_n}$. Each hypothesis represents a distinct scenario that could explain an observed event.
- **Priori Probability:** $P(E_i)$ is the initial probability of an event occurring before any new data is taken into account. It reflects existing knowledge or assumptions about the event.

  Example: The probability of a person having a disease before taking a test.
- **Posterior Probability** (P(Ei|A) is the updated probability of an event after considering new information. It is derived using the Bayes Theorem.

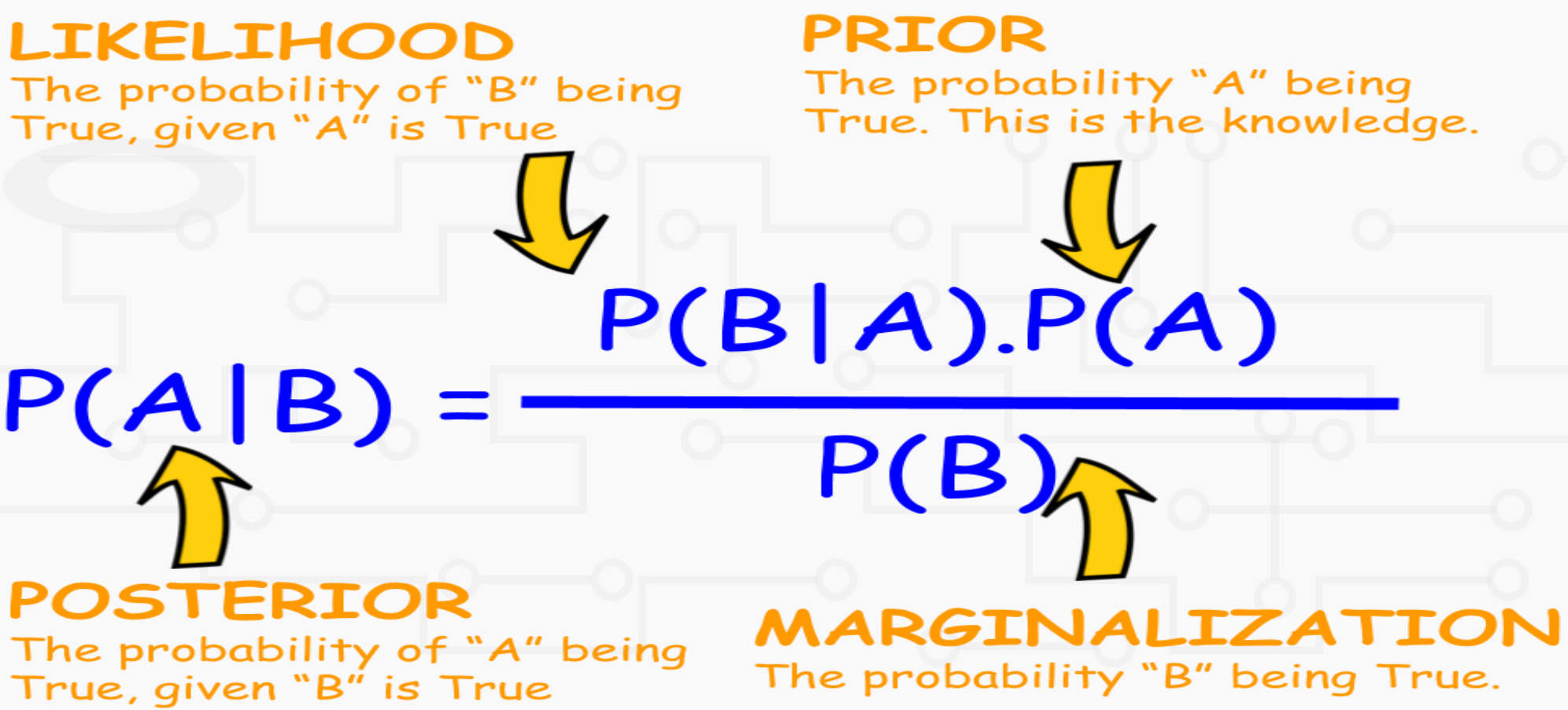
  Example: The probability of having a disease given a positive test result.

Bayes Theorem – Definitions

جامعة الحسين التقنية
Al Hussein Technical University

- **Conditional Probability**: The probability of an event A based on the occurrence of another event B is termed conditional Probability. It is denoted as **P(A|B)** and represents the probability of A when event B has already happened.
- **Joint Probability**: When the probability of two or more events occurring together and at the same time is measured, it is marked as Joint Probability. For two events A and B, it is denoted by joint probability is denoted as **P(A∩B).**
- **Random Variables:** Real-valued variables whose possible values are determined by random experiments are called random variables.
- The probability of finding such variables is the experimental probability

Bayes Theorem – Definitions



https://miro.medium.com/v2/resize:fit:1400/1*CnoTGGO7XeUpUMeXDrIfvA.png

Bayes Theorem - Example

جامعة الحسين التقنية
Al Hussein Technical University

- A person has undertaken a job. The probability of completing the job on time if it rains is 0.44, and the probability of completing the job on time if it does not rain is 0.95. If the probability that it will rain is 0.45, then determine the probability that the job will be completed on time.

Let:

- *R: event that it rains*
- *R^c: event that it does not rain*
- *C: event that the job is completed on time*

Bayes Theorem - Example

جامعة الحسين التقنية
Al Hussein Technical University

- A person has undertaken a job. The probability of completing the job on time if it rains is 0.44, and the probability of completing the job on time if it does not rain is 0.95. If the probability that it will rain is 0.45, then determine the probability that the job will be completed on time.

Let:

- *R: event that it rains*
- *R^c: event that it does not rain*
- *C: event that the job is completed on time*

Bayes Theorem -Task

جامعة الحسين التقنية
Al Hussein Technical University

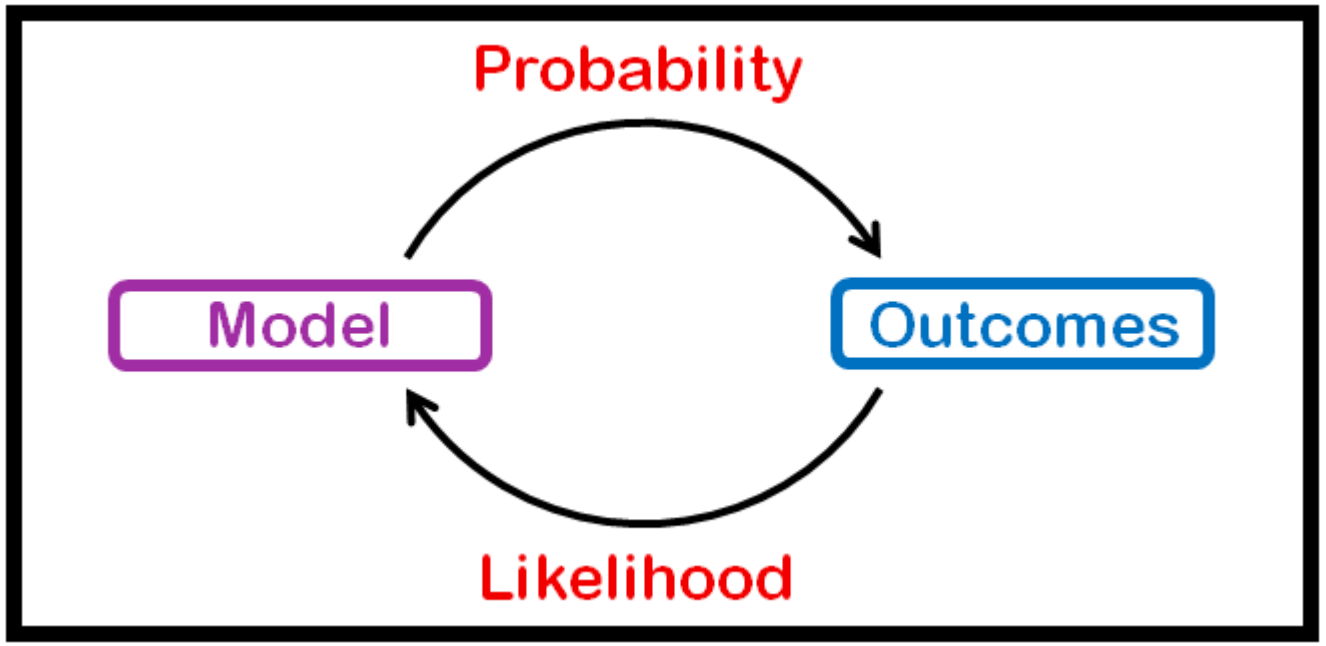
- A person has undertaken a job. The probability of completing the job on time if it rains is 0.44, and the probability of completing the job on time if it does not rain is 0.95. If the probability that it will rain is 0.45, then determine the probability that the job will be completed on time.
  - Work in groups
  - Discuss your results in front of class
  - Time Limit 10 mins

Likelihood

- likelihood, In mathematics, a subjective assessment of possibility that, when assigned a numerical value on a scale between impossibility (0) and absolute certainty (1), becomes a probability.
- Thus, the numerical assignment of a probability depends on the notion of likelihood.
- If, for example, an experiment (e.g., a die toss) can result in six equally likely possible outcomes, the probability of each is 1/6.

- https://www.britannica.com/science/likelihood

Probability vs. Likelihood



https://cdn.analyticsvidhya.com/wp-content/uploads/2023/06/1_I_ipLXJB6qMXOmEIk8bvdQ.png

Probability vs. Likelihood

- "**likelihood**" refers to the chance of observing data given a particular model or hypothesis
- while "**probability**" represents the chance of an event occurring beforehand.
- https://www.youtube.com/watch?v=pYxNSUDSFH4

Log likelihood

The likelihood function represents the probability of the observed data given specific model parameters. Taking the logarithm of the likelihood simplifies calculations, especially when dealing with very small probability values.

Equation:

Log Likelihood (LL) = log (P (Data | Parameters))

جامعة الحسين التقنية
Al Hussein Technical University

Pre-Transformer Architecture

Architectures for Sequence Modeling

جامعة الحسين التقنية
Al Hussein Technical University

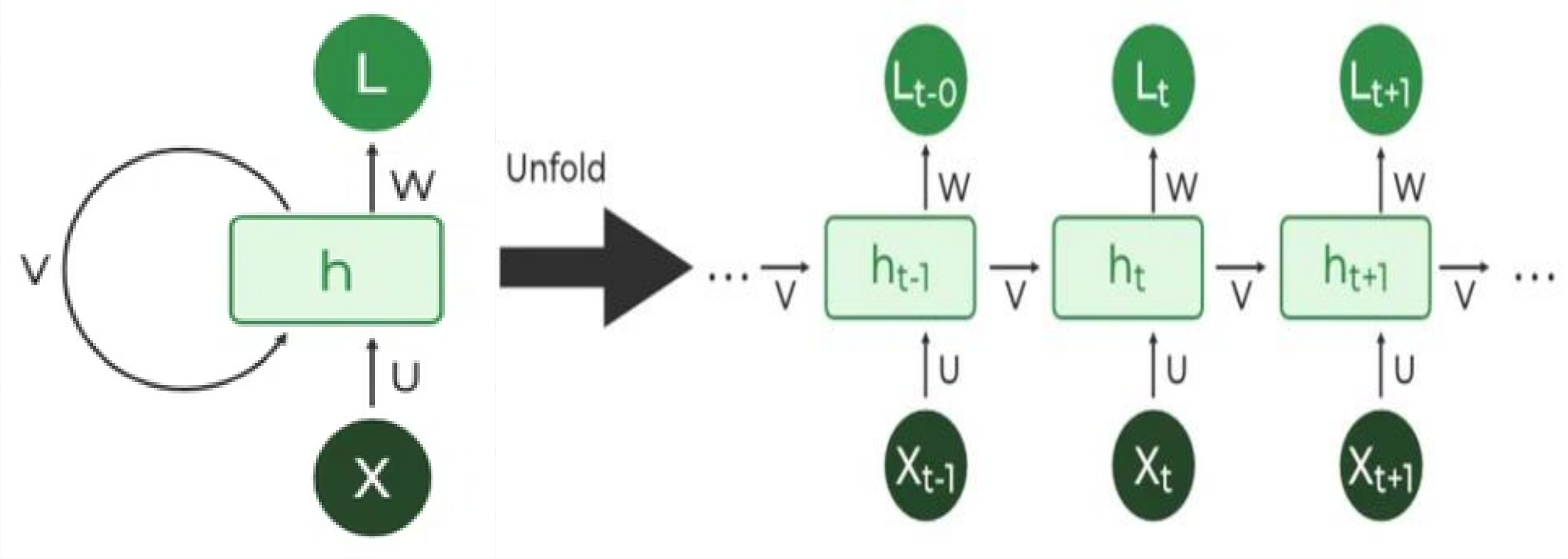
1. **RNN (Recurrent Neural Network):** The base sequential architecture where output depends on the current input and the previous hidden state (memory). Suffers from the vanishing gradient problem.
2. **GRU (Gated Recurrent Unit):** An evolution of the RNN that uses update and reset gates to control information flow, improving memory retention over long sequences. It is simpler and faster to compute than the LSTM.
3. **LSTM (Long Short-Term Memory):** An improvement over RNNs that uses complex **gates** (input, forget, output) and a dedicated **Cell State** to solve the vanishing gradient problem, allowing the network to "remember" crucial information over long time steps.

Recurrent Neural Networks (RNNs)

جامعة الحسين التقنية
Al Hussein Technical University

- **How It Works**:
  RNNs process **sequential data** by passing information through hidden states. They excel in tasks where context matters, such as **text and speech generation**.
- **Limitations**:
  - Struggles with **long-range dependencies** due to the vanishing gradient problem/exploding gradient problem.
  - Slower training compared to newer architectures.

Recurrent Neural Networks (RNNs)



https://www.geeksforgeeks.org/machine-learning/introduction-to-recurrent-neural-network/

Recurrent Neural Networks (RNNs)

جامعة الحسين التقنية
Al Hussein Technical University

Example Model:

- **Early Chatbots & Text Generators**
- **Example: OpenAI's early text-generation models (pre-GPT)** used RNNs and LSTMs to generate text.

Applications:

1. **Speech-to-Text AI** — Used in early **Google Voice & Siri**.
   Music Composition — AI composers like **MuseNet** generate melodies.
2. **Predictive Text** — Keyboard suggestions powered by simple RNN-based AI.

Long Short-Term Memory Networks (LSTMs)

جامعة الحسين التقنية
Al Hussein Technical University

- LSTM is an enhanced version of the RNN
- LSTMs can capture long-term dependencies in sequential data making them ideal for tasks like language translation, speech recognition and time series forecasting.
- LSTMs introduce a memory cell that holds information over extended periods addressing the challenge of learning long-term dependencies.
- This allows LSTM networks to selectively retain or discard information as it flows through the network which allows them to learn long-term dependencies.

Long Short-Term Memory Networks (LSTMs)

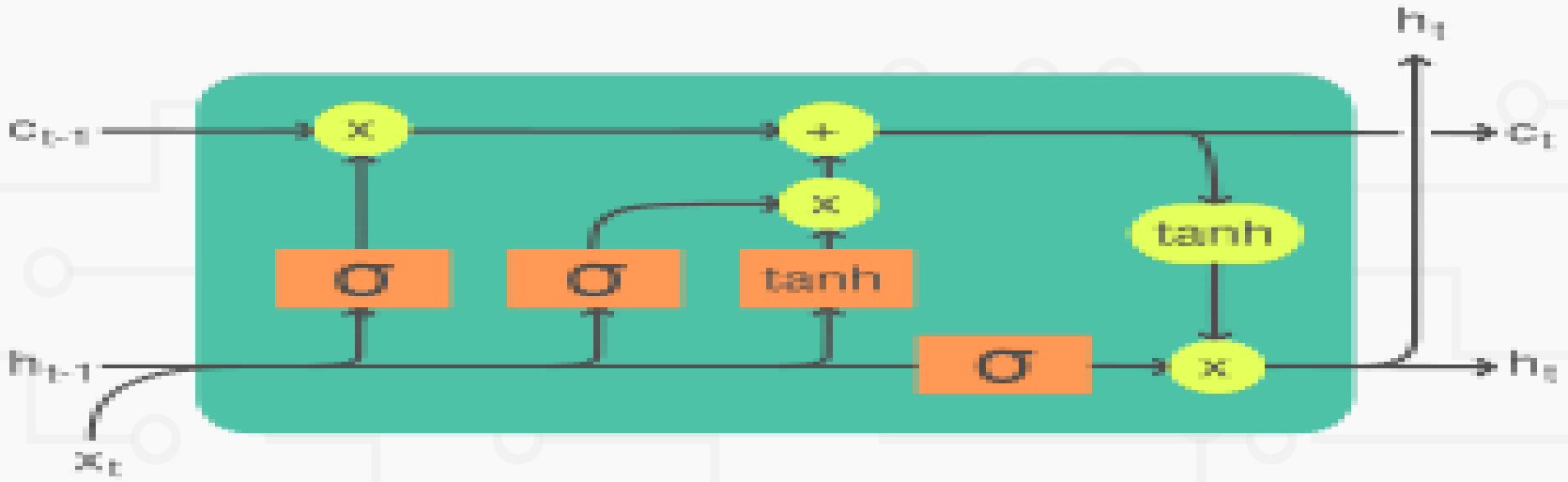
جامعة الحسين التقنية
Al Hussein Technical University

LSTM architectures involves the memory cell which is controlled by three gates:

- **Input gate**: Controls what information is added to the memory cell.
- **Forget gate**: Determines what information is removed from the memory cell.
- **Output gate**: Controls what information is output from the memory cell.

- https://towardsdatascience.com/wp-content/uploads/2022/02/1Zrht4QBK5_hAxif17ED4ew-1112x1536.png
-

Long Short-Term Memory Networks (LSTMs)



Long Short-Term Memory Networks (LSTMs)

جامعة الحسين التقنية
Al Hussein Technical University

Example models

Baidu's Deep Speech
Early versions of **SwiftKey** or **Gboard**
Early versions of **Google Translate**

Applications:

1. **Language Modeling**
2. **Speech Recognition**
3. **Forecasting**.
4. **Anomaly Detection**
5. **Recommender Systems**
6. **Video Analysis**

جامعة الحسين التقنية
Al Hussein Technical University

Architectures for Generation

The Generative Models

جامعة الحسين التقنية
Al Hussein Technical University

- Autoregressive models
- Variational Autoencoders (VAEs)
- Generative Adversarial Networks (GANs)
- Flow-Based model
- Diffusion models (modern approach)

Autoregressive Models

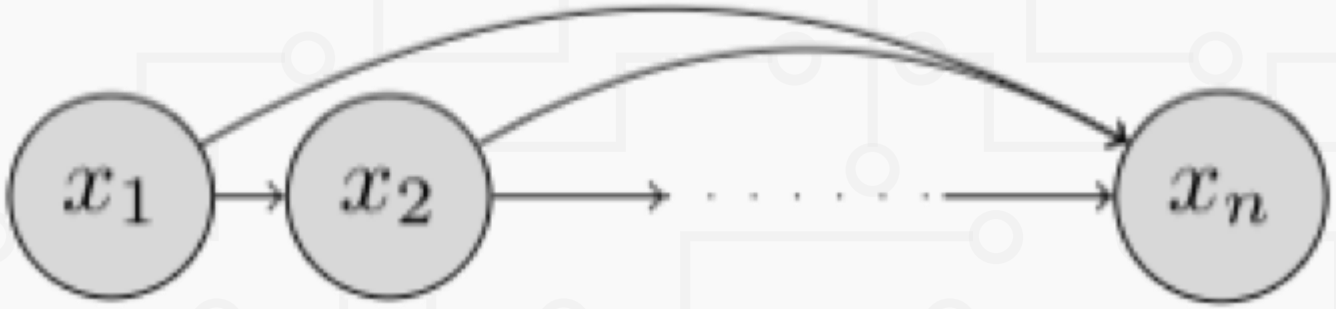
جامعة الحسين التقنية
Al Hussein Technical University

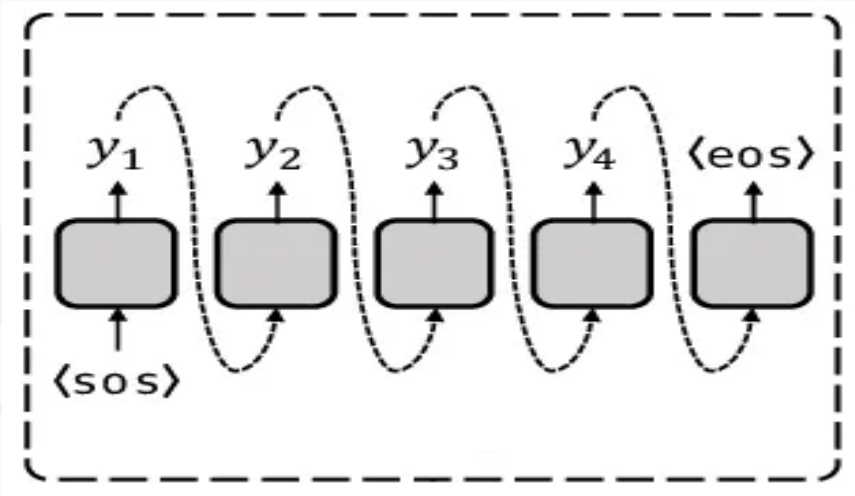
- **How It Works**:
  An **autoregressive (AR) model** is a type of model that **predicts the next value (or token)** in a sequence **based on previous ones**.

$$P(x) = P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_{<t})$$

- **Key Strength**:
  - Simple and **stable** training (maximum likelihood)
  - Great for sequence generation (text, speech)
  - Captures strong local dependencies

Autoregressive Models





https://deepgenerativemodels.github.io/notes/autoregressive/autoregressive.png

Autoregressive Models

جامعة الحسين التقنية
Al Hussein Technical University

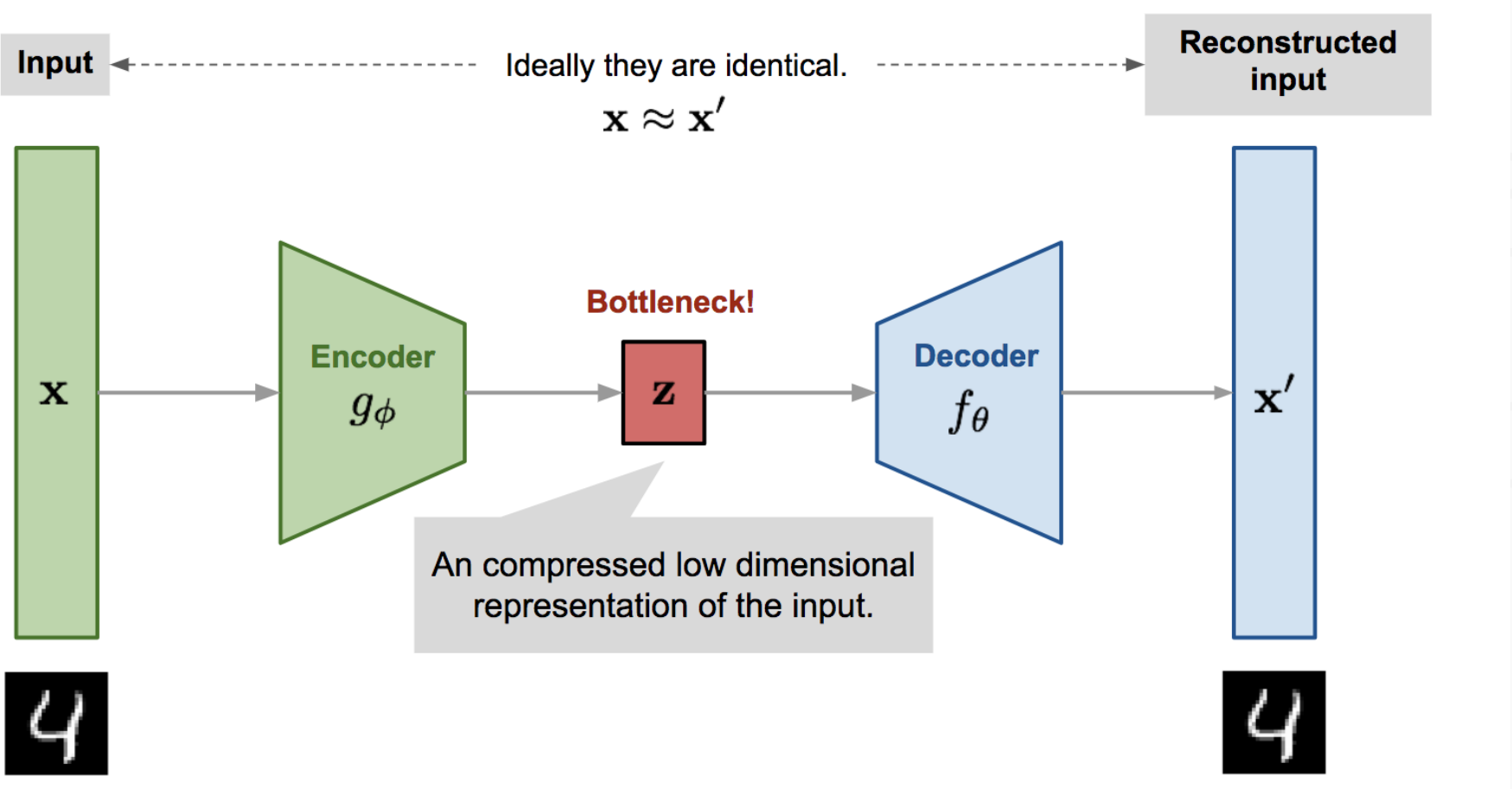
- **How It Works**:
  - Slow generation — must predict tokens one at a time.
  - No global view — can struggle with long-term dependencies.
  - Exposure bias — model only sees gold-standard history during training, not its own errors.

Variational Autoencoders (VAEs)

جامعة الحسين التقنية
Al Hussein Technical University

- **How It Works**:
  VAEs encode input data into a **latent space representation**, then reconstruct it. Unlike GANs, VAEs learn **probabilistic distributions**, making them more interpretable.
- **Key Strength**:
  - Better at **structured data generation** than GANs.
  - Can control specific **features in generated data**.

Variational Autoencoders (VAEs)



https://lilianweng.github.io/posts/2018-08-12-vae/autoencoder-architecture.png

Variational Autoencoders (VAEs)

جامعة الحسين التقنية
Al Hussein Technical University

Example Model:

- **VAE-Based Face Filters** — Snapchat & Instagram's AR face filters use VAEs.
- **Handwriting Synthesis** — AI handwriting generators use VAEs.

Applications:

1. **AI Face Filters** — Used in **Snapchat, TikTok, Instagram AR effects**.
2. **Anomaly Detection** — Used in **fraud detection & cybersecurity**.
3. **Synthetic Handwriting** — AI-generated fonts and styles.

Generative Adversarial Networks (GANs)

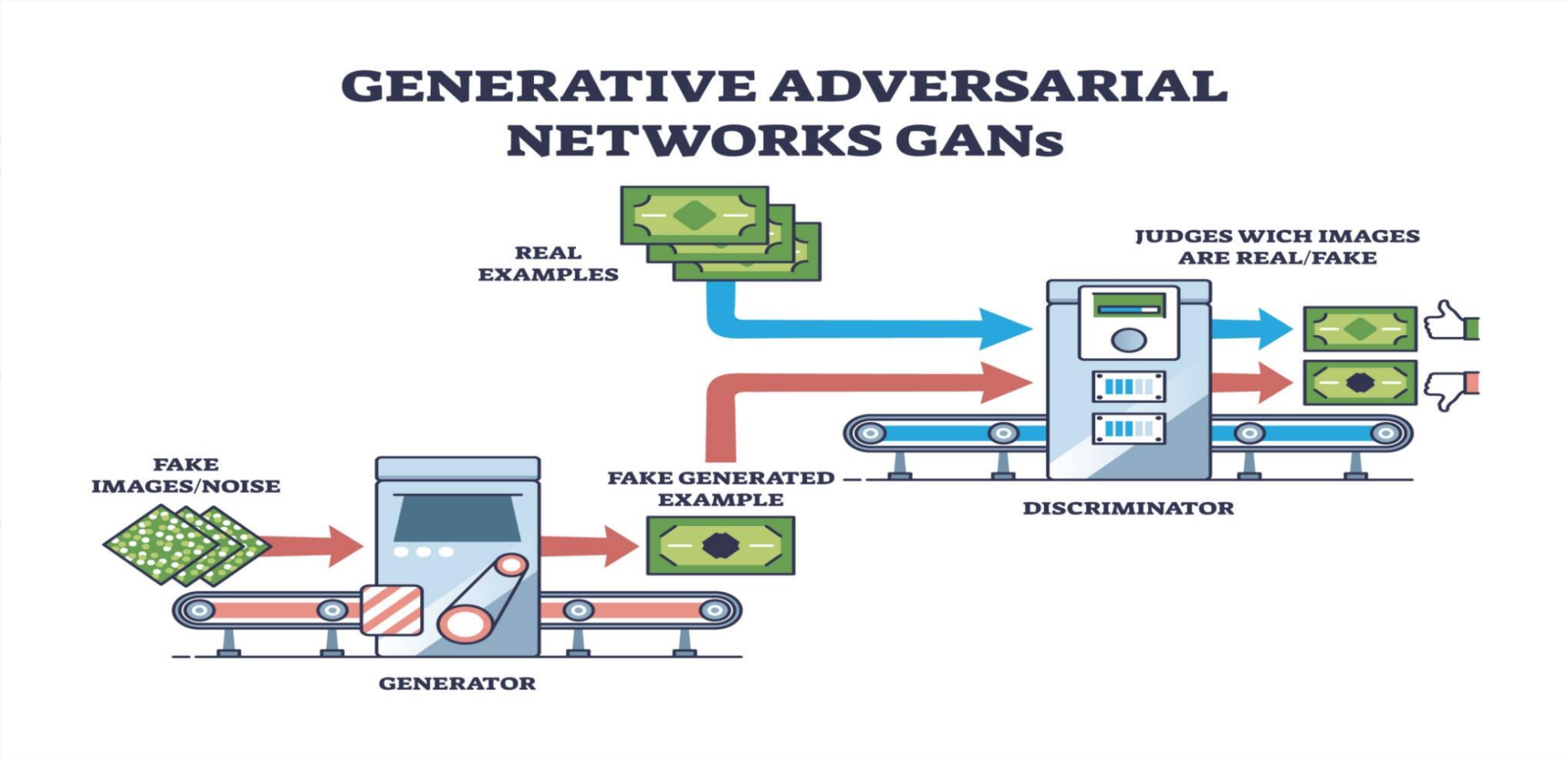
جامعة الحسين التقنية
Al Hussein Technical University

- **How It Works**:
  GANs consist of two neural networks:
  - **Generator**: Creates synthetic data.
  - **Discriminator**: Evaluates if the generated data is real or fake.

They continuously compete, improving the realism of generated outputs.

- **Key Strength**:
  - Excellent at generating **high-quality synthetic media**.
  - Often used in AI **art, deepfakes, and upscaling images**.

Generative Adversarial Networks (GANs)



Generative Adversarial Networks (GANs)

جامعة الحسين التقنية
Al Hussein Technical University

Example Model:

- **StyleGAN (NVIDIA)** — Generates **hyper-realistic human faces** (ThisPersonDoesNotExist.com).
- **Deepfake AI** — GANs power face-swapping technology in videos.

Applications:

1. **AI-Generated Art** — Used in **AI painting and digital avatars**.
2. **Deepfake Videos** — **Face-swapping and synthetic actors** in Hollywood.
3. **Image Enhancement** — **NVIDIA DLSS** for upscaling low-res images.

Flow-Based Models

جامعة الحسين التقنية
Al Hussein Technical University

Flow-based generative models are a type of **likelihood-based model** that explicitly models a complex probability distribution by transforming a simple distribution (like a standard Gaussian/normal distribution) through a sequence of invertible transformations.

- **Key Innovation:**

  The transformations used are designed to be **invertible** and have a **tractable Jacobian determinant**. This mathematical design allows the model to calculate the **exact log-likelihood** of any new data sample.

Flow-Based Models

جامعة الحسين التقنية
Al Hussein Technical University

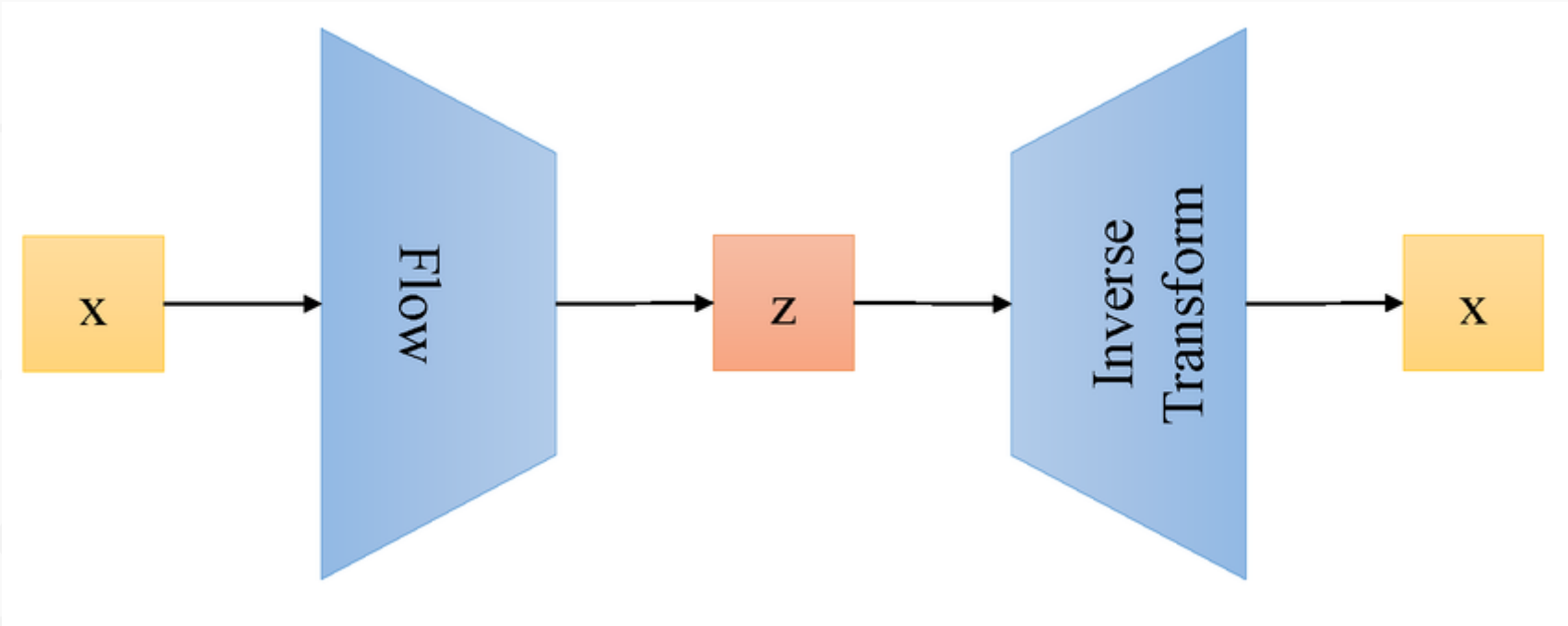
- **Training Objective:**

  They are trained directly by **maximizing the log-likelihood of the data**.

- **Trade-offs:**

  They provide **exact density estimation** and guarantee efficient, non-iterative sampling. However, the strict requirement that all transformations must be invertible often limits the model's architectural flexibility (expressiveness) compared to models like GANs.

Flow-Based Models



https://www.researchgate.net/publication/373998479/figure/fig3/AS:11431281359697729@1744083619823/The-flow-based-generative-models.tif

Flow-Based Models

جامعة الحسين التقنية
Al Hussein Technical University

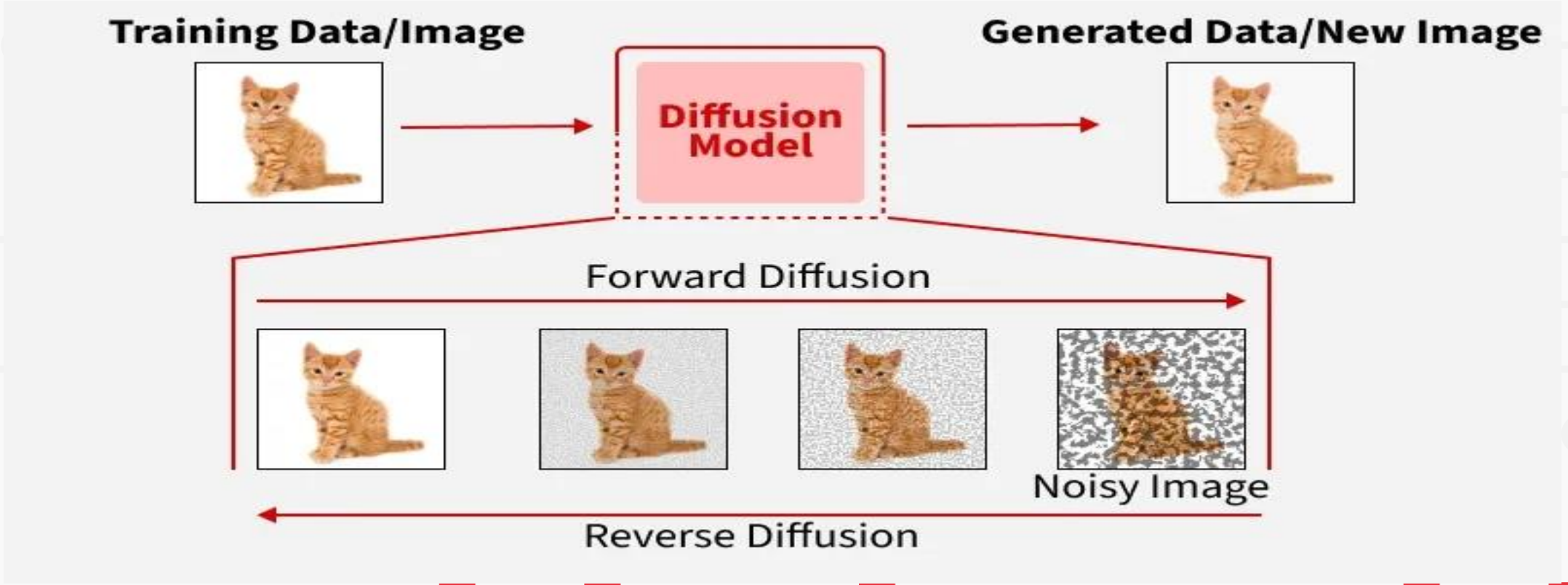
Aspect	Flow Model
Type	Generative model
Core principle	Learn invertible mapping between data and latent space
Training objective	Maximize exact data likelihood
Common architecture	RealNVP, Glow, Flow++, FFJORD
Pros	Exact likelihoods, stable training, fast sampling
Cons	Limited expressiveness, memory-heavy invertibility constraint
Used for	Image, speech, video synthesis, density estimation

Diffusion Models

Diffusion models represent a modern class of generative models that have gained prominence for their ability to produce highly realistic images, surpassing the quality of GANs.
They operate via an iterative, two-stage process:

- **Forward Diffusion:** Gradually adds random noise to the training data until the data is completely randomized.
- **Reverse Diffusion (Generation):** The model learns how to reverse this process, starting from pure noise and gradually removing the noise at every step until a coherent, high-quality image or video emerges.

Diffusion Model Architecture



https://media.geeksforgeeks.org/wp-content/uploads/20250804190118579985/diffusion_model.webp

Diffusion Models

جامعة الحسين التقنية
Al Hussein Technical University

- **How It Works**:
  Diffusion models start with **random noise** and gradually refine it into meaningful data. This process, inspired by physics, leads to **high-quality image and video generation**.
- **Key Strength**:
  - Produces **photo-realistic images and animations**.
  - Can generate **high-resolution media better than GANs**.

Diffusion Model Architecture

جامعة الحسين التقنية
Al Hussein Technical University

Denoising diffusion models

- **Forward / noising process**
  - Sample data $p(\mathbf{x}_0)$ → turn to noise

$p_0(\mathbf{x}_0)$ — $p_T(\mathbf{x}_T) \sim \mathcal{N}(0, I$

Clean sample — $\mathbf{x}_0$ $\mathbf{x}_1$ $\mathbf{x}_{T-1}$ $\mathbf{x}_T$ — *Pure noise*

- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

Diffusion Models

جامعة الحسين التقنية
Al Hussein Technical University

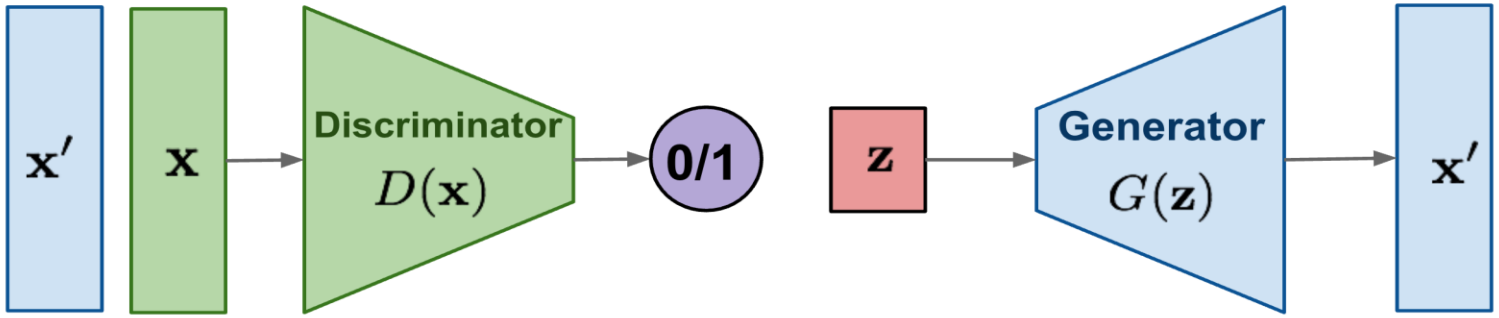
Example Model:

- **Stable Diffusion (Stability AI)** — Open-source AI that generates **high-quality images from text**.
- **Imagen (Google)** — Creates **ultra-realistic AI-generated photography**.

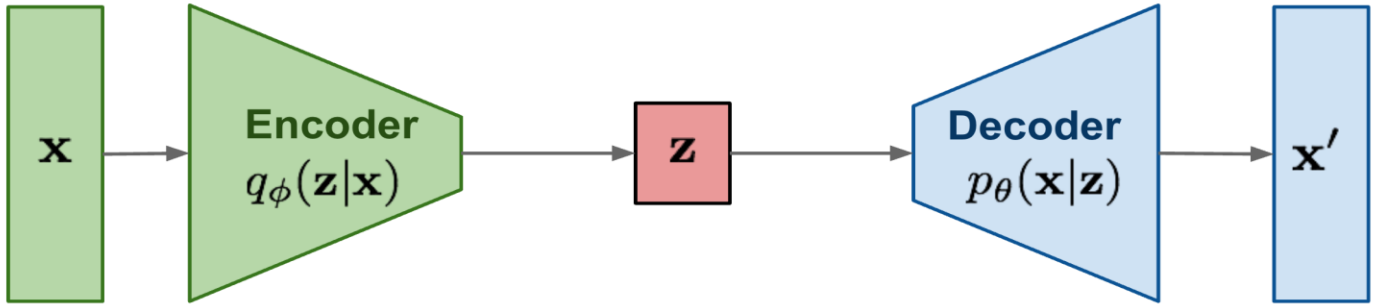
Applications:

1. **AI Art & Photography** — Used in **DALL·E 2, Midjourney, Stable Diffusion**.
2. **Video Generation** — AI-powered animations and deepfake videos.
3. **Image Restoration** — Enhances old or blurred photos.

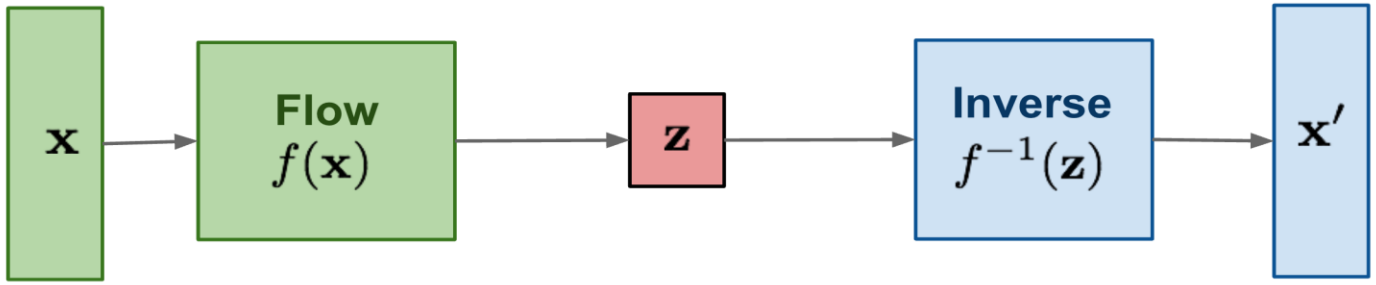
GAN: Adversarial training



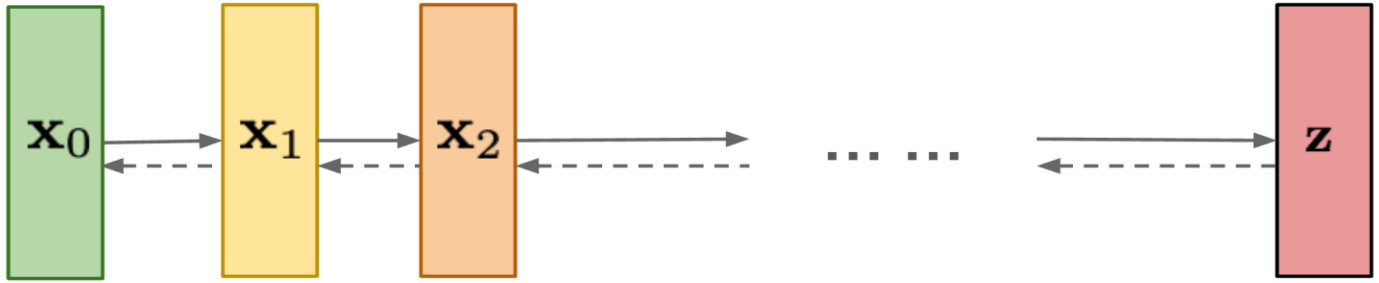
VAE: maximize variational lower bound



Flow-based models: Invertible transform of distributions



Diffusion models: Gradually add Gaussian noise and then reverse



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/generative-overview.png

Task

Given all the architectures we discussed so far in this course for sequence and language modelling, can you provide a timeline for the order of their appearance throughout the history:

- Work in groups
- Post your work on white board
- Time limit 15 minutes
- Assign a focal person for each group

Architecture Timeline

- 1980s
- 1986 RNN
- 1997 LSTM
- 2014 GRU
- 2014 GAN
- 2014 VAE
- 2017 Transformer
- 2018 Flow-based
- 2020 Diffusion
- 2020 GPT-3 / LLM Era begins
- 2021+ Foundation Models (FM) → multimodal

Classical neural & generative foundations

Module 2-A Summary & Transition

جامعة الحسين التقنية
Al Hussein Technical University

- **Probability vs. Likelihood:**
  The two are inverse perspectives of the same core formula. **Probability** predicts outcomes given a fixed model, while **Likelihood** estimates the plausibility of a model given fixed data.
- **Generative Goal:**
  The fundamental goal of generative models is **synthesis and creation** by modeling the underlying **joint probability distribution** of the data.
- **Pre-Transformer Models:**
  Classical sequence models like **RNNs, LSTMs, and GRUs** were limited by processing data **sequentially** and struggling with long-range dependencies due to the vanishing gradient problem.
- **Generative Architectures:**
  - **VAEs** (Likelihood-based) are stable and create smooth latent spaces.
  - **GANs** (Implicit) use competition to achieve high realism.
  - **Diffusion Models** (Modern) use noise removal for high-fidelity media generation.

Module 2-A Summary & Transition

جامعة الحسين التقنية
Al Hussein Technical University

- The previous modules have armed you with the technical understanding of how Foundation Models (LLMs) are built from the Transformer architecture and the probability they rely on.
- The next step is to master how to **control** and **direct** these massive models. We now move to the most immediate and hands-on skill required in the field: **Prompt Engineering**.

Task

Based on your findings in the previous task, can you provide a three examples of interesting applications/models for each milestone on your timeline

- Work in groups
- Discuss your finding in front of class at the beginning of the next class