Generative AI and Cybersecurity: Challenges and Applications

By

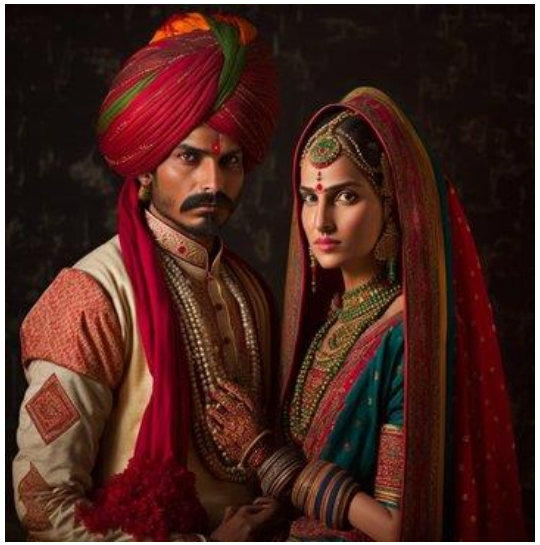
Dr. Satyendra Singh Chouhan

मालवीय राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

Malaviya National Institute of Technology Jaipur

[AN INSTITUTE OF NATIONAL IMPORTANCE]

Which one is real?

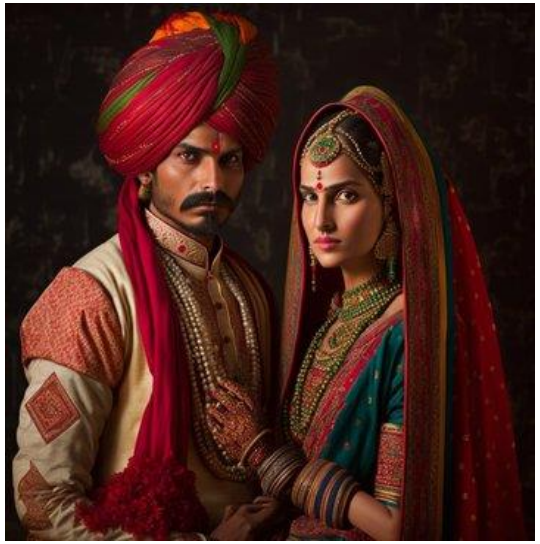


A

B

C

Which one is real?



None of these!

Outline

Part1: Introduction and Fundamentals of Generative AI

Part2: Generative AI and Cyber Security

Part3: Demonstration

Supervised Vs Unsupervised Learning

Supervised Learning

Data: (x, y)
X is data, y is label

Goal: Learn a function to map x→y

Example: Classification, Regression, Object detection, semantic segmentation

Supervised Vs Unsupervised Learning

Supervised Learning

Data: (x, y)
X is data, y is label

Goal: Learn a function to map x→y

Example: Classification, Regression, Object detection, semantic segmentation

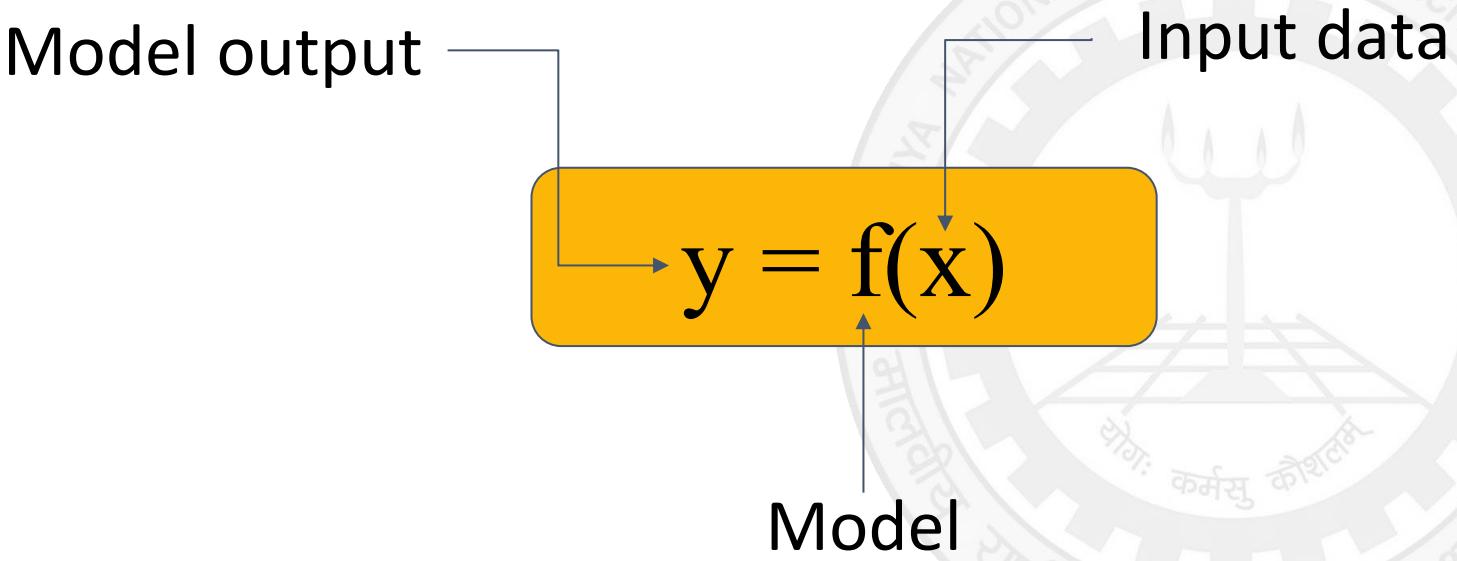
Unsupervised Learning

Data: (x, y)
X is data, no label!

Goal: Learn some hidden or underlying structure of the data

Example: Clustering, feature or dimensionality reductions etc.

What are Generative Models?



Not Gen AI when y is:

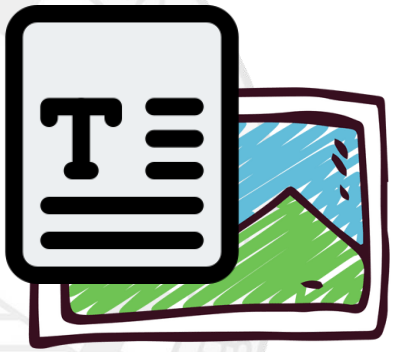
- Number
- Discrete
- Class
- Probability

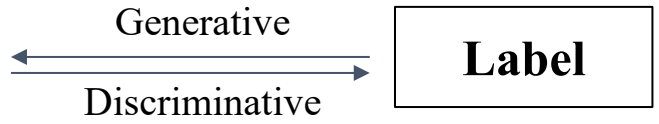
Is Gen AI when y is:

- Natural language
- Image
- Audio
- Video

Generative vs Discriminative Models

- Discriminative models predict labels, ***P(y|x)***
- Generative models learn data distributions, ***P(x)***
- Complementary roles in AI

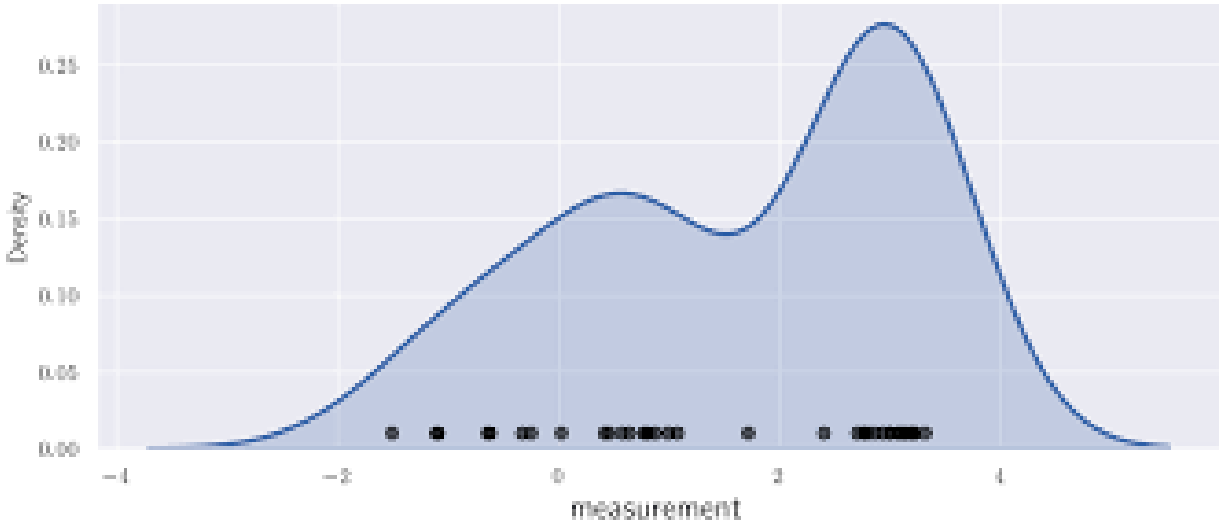




Generative Modelling

Goal: Take as input training samples from some distribution and learn a model that represents that distribution

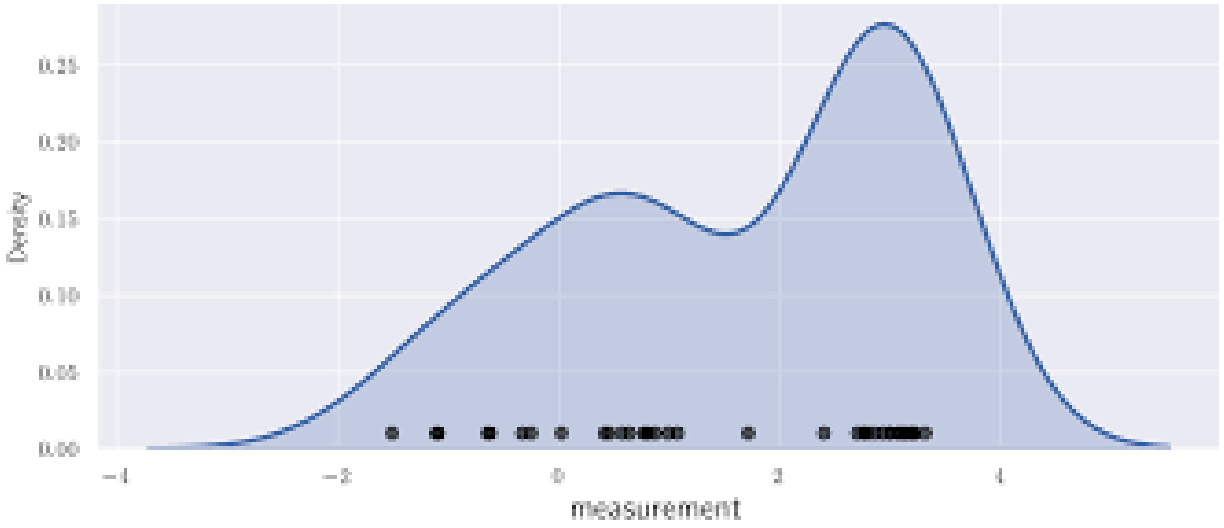
Density Estimation



Generative Modelling

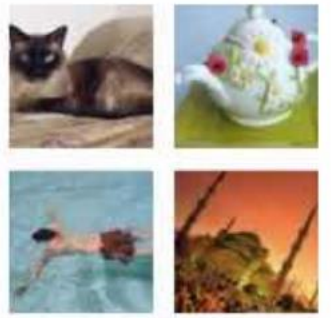
Goal: Take as input training samples from some distribution and learn a model that represents that distribution

Density Estimation



Sample Generation

Input samples



Generated samples

Training data ~ $P_{data}(x)$

Generated ~ $P_{model}(x)$

How can we learn $P_{model}(x)$ similar to $P_{data}(x)$?

How generative Models works?

Intuition of Generative AI

X=[3.2, 3.8, 4.1, 5.0, 5.3, 6.1, 6.8]

We can estimate the probability distribution of this data by calculating:

- **Mean (μ)**: The average of all points. $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Standard deviation (σ)**: The spread of the points around the mean.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

From the dataset, let's assume:

- μ=5.0(mean)
- σ=1.2(standard deviation)

To generate a new sample X_{new} we assume the data follows a normal distribution.

$$x_{\text{new}} = \mu + \sigma \cdot Z$$

Suppose $Z = 0.5$

$$x_{\text{new}} = \mu + \sigma \cdot Z$$

$$x_{\text{new}} = 4.9 + (1.19 \times 0.5)$$

$$x_{\text{new}} = 4.9 + 0.595 = 5.495$$

Similarly

[4.8, 5.5, 3.9, 6.3, 4.7]

In Real world?

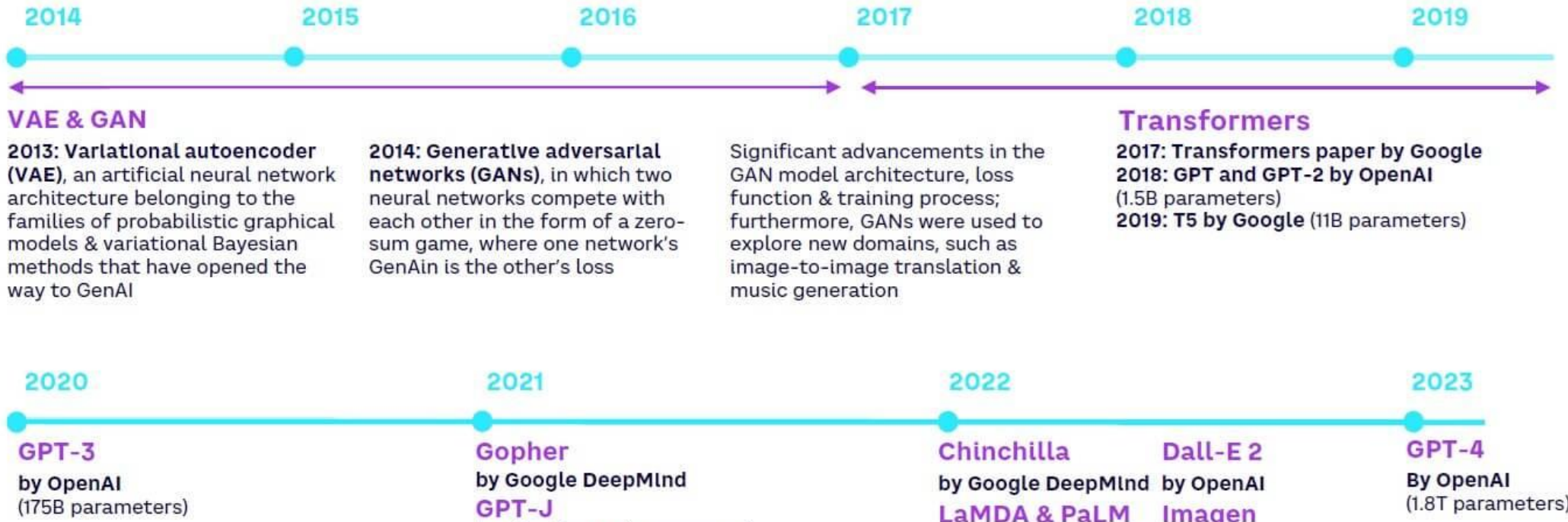
1. Complex Structures in Data
2. Multi-modal Distributions
3. High-Dimensionality
4. Implicit vs. Explicit Models

History

- Early approaches: Simple probabilistic models (Gaussian Mixture Models)

- Variational Autoencoders (VAE) [2013]
- Generative Adversarial Networks (GAN) [2014]
- Diffusion Models [2015]
- Transformers [2017]

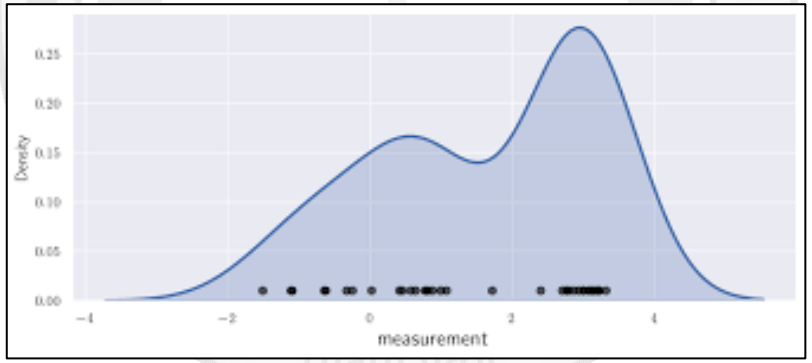
Timeline



Source: Arthur D. Little; Foster, David. *Generative Deep Learning*. O'Reilly Media, 2019; LinkedIn

Objective of Generative Models

1. To learn the underlying data distribution, ***P(x)***
2. Generate new samples from the learned distribution



Learn $P_{model}(x)$ similar to $P_{data}(x)$

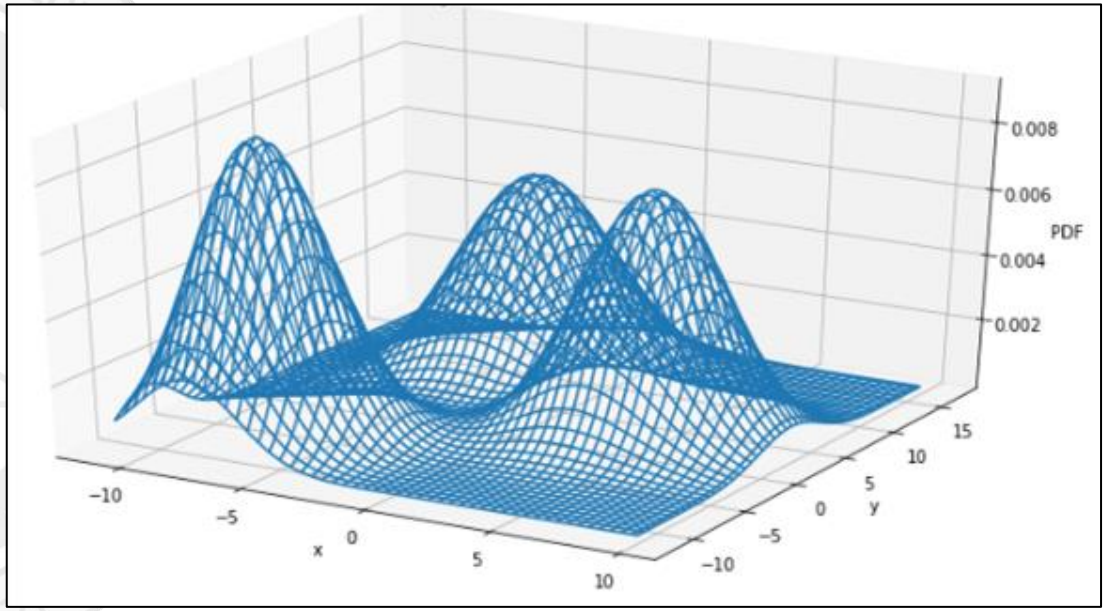
Data Distribution

What is data distribution, ***P(x)***?

- Images: Pixel value distribution
- Text: Word sequences

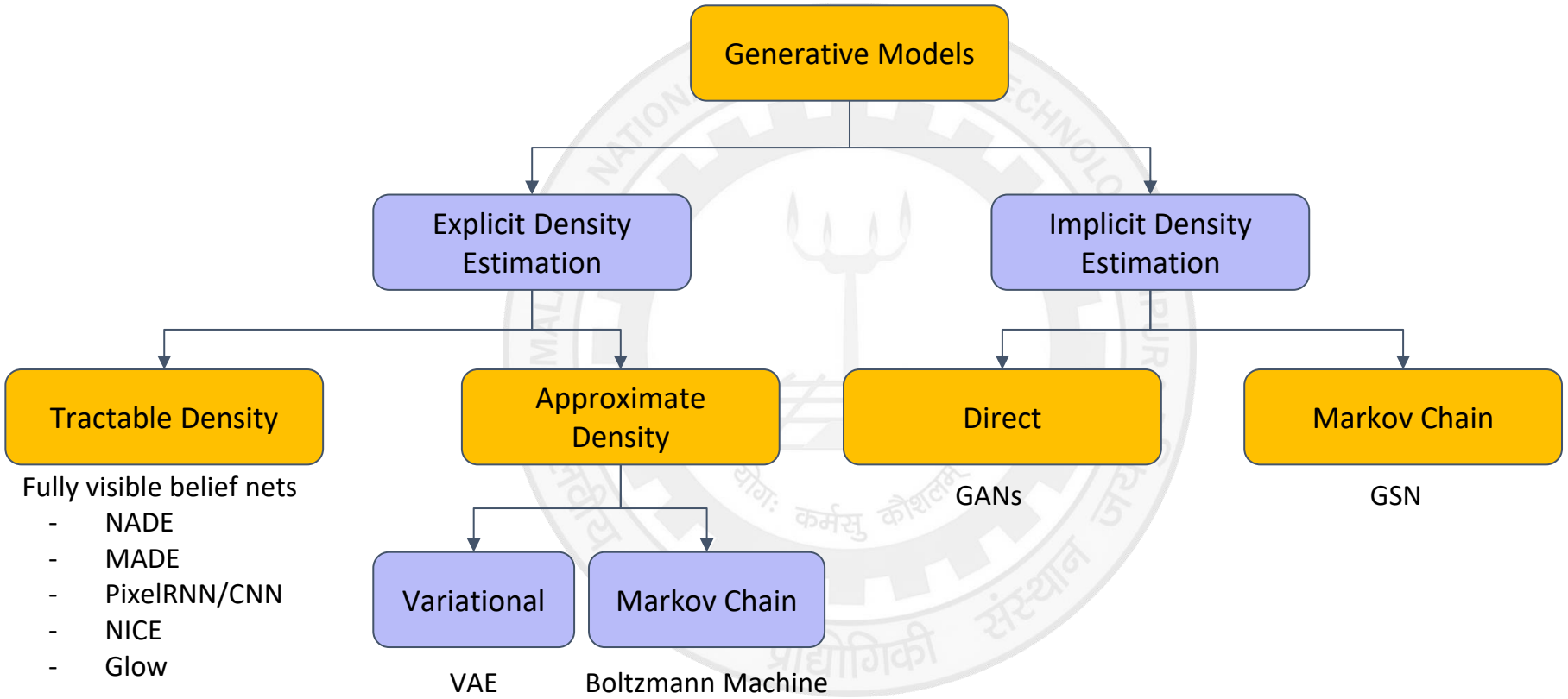
Goal: Model or Approximate ***P(x)***
a.k.a. Density Estimation

Challenge: Distributions can be high dimensional and complex



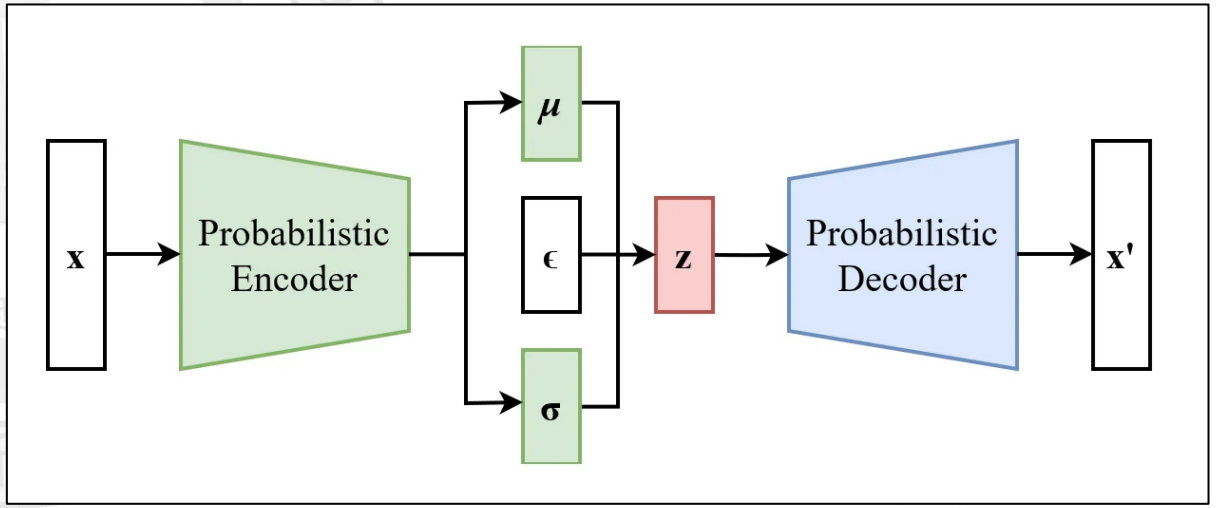
A complex 2d distribution

Types of Generative Models



Explicit Density Models

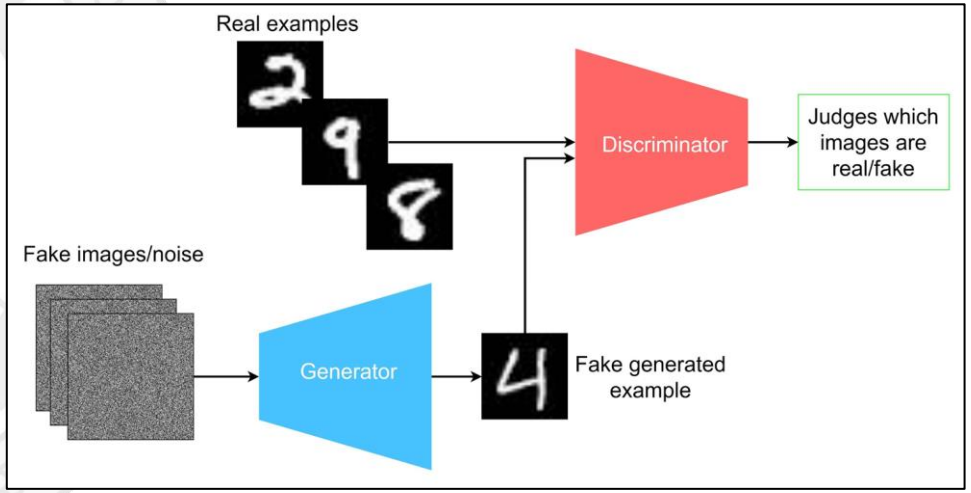
- Model the data distribution ***P(x)*** directly
- E.g. VAEs, PixelRNN
- Interpretable



A VAE trying to learn a Normal Distribution (μ, σ)
(Fig credits: Matthew MacFarquhar)

Implicit Density Models

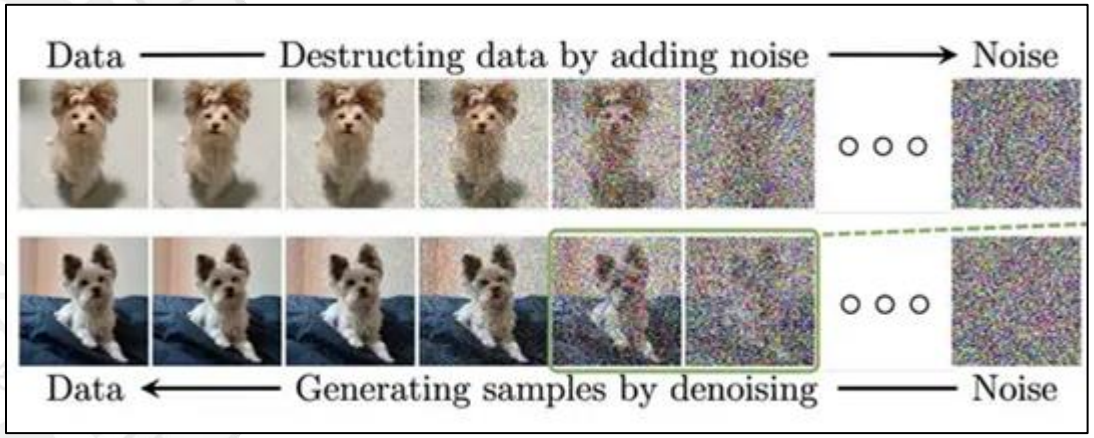
- Learn to sample without explicitly modeling ***P(x)***
- E.g. GANs
- Flexible



A GAN learning to sample from MNIST dataset
(Fig credits: IBM Developer)

Other Models

- Combine explicit and implicit approaches
- E.g. Diffusion models, flow based models
- High quality output

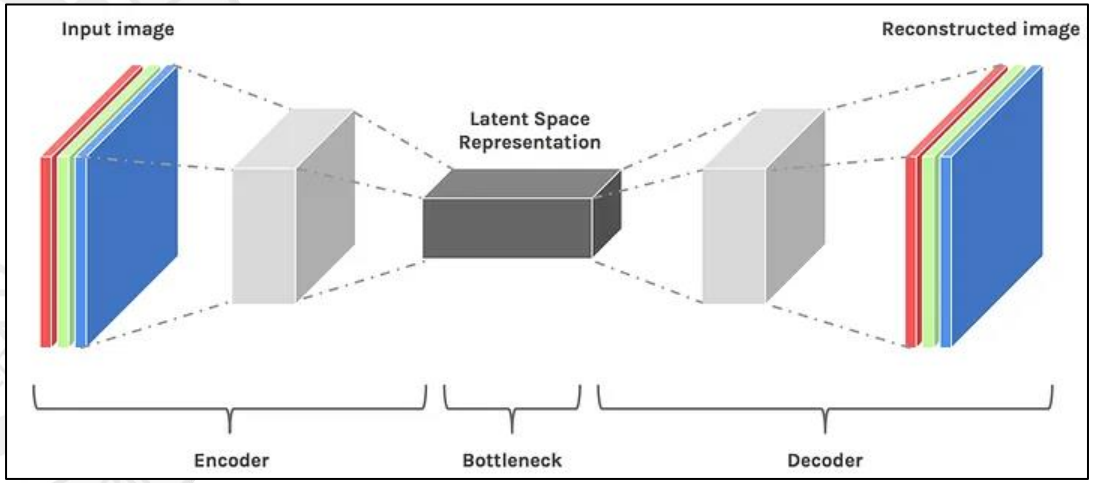


A Diffusion Model learns to generate via denoising
(Fig credits: SuperAnnotate)

Basic Concepts

Latent Space

- Low dimensional representation of data
- Encode complex data into simpler patterns
- Decode (aka Reconstruct) output from this compressed mapping

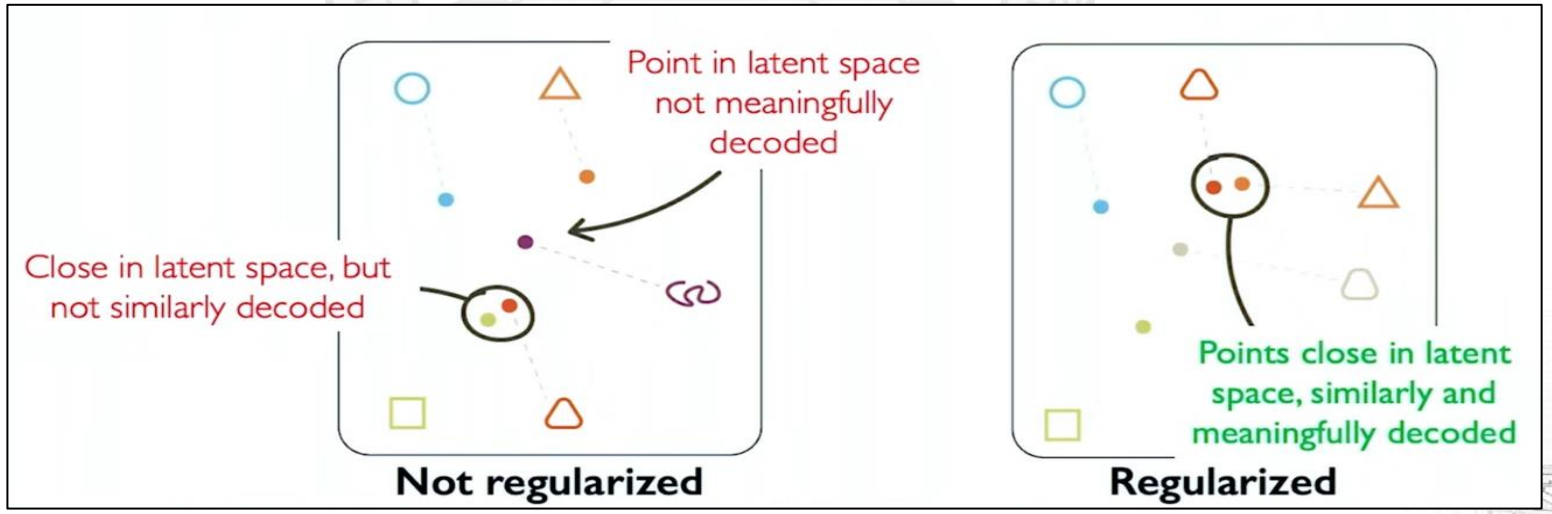


Latent space in image reconstruction
(Fig credits: Hopsworks)

Latent Space: Characteristics

Desired properties:

- **Continuity:** points close in latent space = similar content
- **Completeness:** sampling from anywhere in latent space = meaningful content
- **Disentanglement:** each feature in data captured separately in latent space



(Fig credits: MIT 6.S191)

Reconstruction

- Reconstructing input-like-data from latent

Smile manipulation by varying one of the latent dimensions

Pose manipulation by varying another latent dimension

(Fig credits: Vivian Cheng)

- **Reconstruction Loss**: Measures similarity to input, e.g. MSE

  - Ensures generated data resembles the original

Generative Loss

- Loss function that guides the generative process
- Balances quality vs diversity in output
- Adversarial nature

Training Generative Models

- Data → Model → Loss → Optimization
- Challenges: High dimensional data, computational cost
- Optimization techniques: SGD, Adam etc.

Loss Functions

- Reconstruction Loss (e.g. MSE, L1 loss)

$$L_{recon} = \|x - \hat{x}\|^2 = \sum (x_i - \hat{x}_i)^2$$

- KL Divergence (use in VAEs)

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

- Adversarial Loss (for GANs)

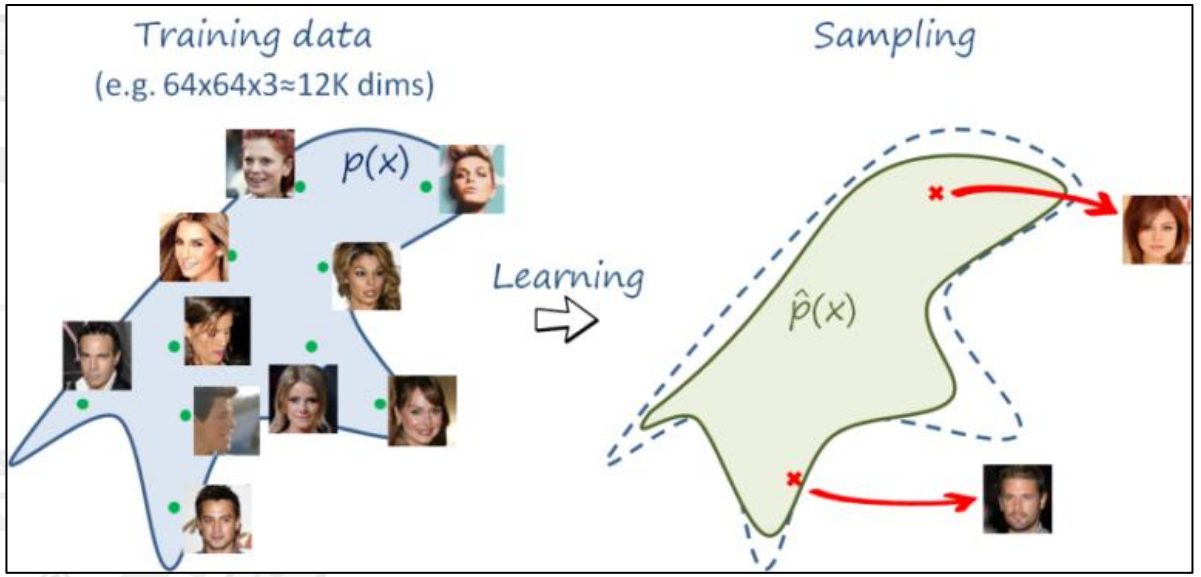
$$L_D = -E_{x \sim pdata}[\log D(x)] - E_{z \sim pz}[\log (1 - D(G(z)))]$$

$$L_G = -E_{z \sim pz}[\log D(G(z))]$$

Sampling

How to produce new data

- Random sampling
  - A random number sampled from N(0,1) used as input latent vector
- Ancestral Sampling
  - Variables are ordered such that each variable comes after all its parents (dependency)
- MCMC



Training and sampling in a generative model
(Fig credits: Luis Herranz)

Applications

Applications: Image Generation

- Realistic images
  - Faces
  - Landscapes
  - VFX
- Surreal images
  - Game assets
  - Art
- E.g. DALL-E, Stable Diffusion

A landscape generated by Sora, OpenAI

Applications: Speech Generation

- Text to speech
- Voice over and dubbing
- Customer support automation
- Podcast generation

- E.g. ElevenLabs, ChatGPT-Audio, Google Lyria

Applications: Video Generation

- High quality educational content
- Marketing and advertising
- Low cost content creation
- Media and entertainment

- E.g. OpenAI Sora, Google Veo 3

Applications: Text Generation

- Question Answering (IR)
- Story writing
- Song writing
- Code generation / completion
- Assignments? 😳

- E.g. ChatGPT, Claude, DeepSeek

Applications: Data Augmentation

- Generating synthetic data for training
- Addresses data scarcity
- Private data
- Imbalanced datasets

- E.g. autonomous driving systems, medical imaging, fraud detection, low resource languages

Applications: Creative AI

- Art, music and design generation
- Regeneration and healing of damaged images
- Game playing bots

- E.g. AlphaGo

Applications: Science

- **Drug discovery** - generating molecular structures
- **Physics** - simulating complex systems e.g. weather
- **CS research** - proposing, proving, testing and reviewing innovative ideas and theorems

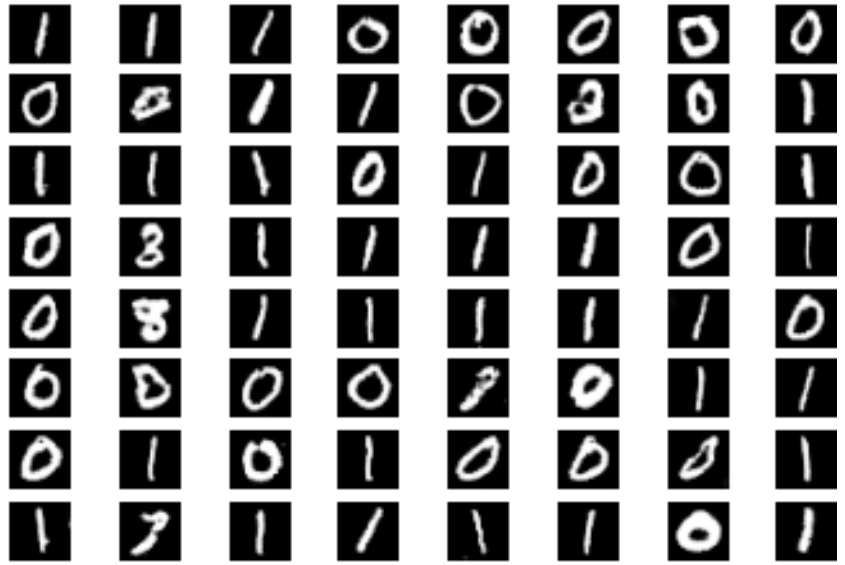
- E.g. AlphaFold, AlphaQubit, AI Scientist, WeatherNext

Challenges

Challenges: Mode Collapse

- Model generates limited variety of output
- Commonly seen in GANs

- Solution: Modifying loss functions to include diversity e.g. WGAN

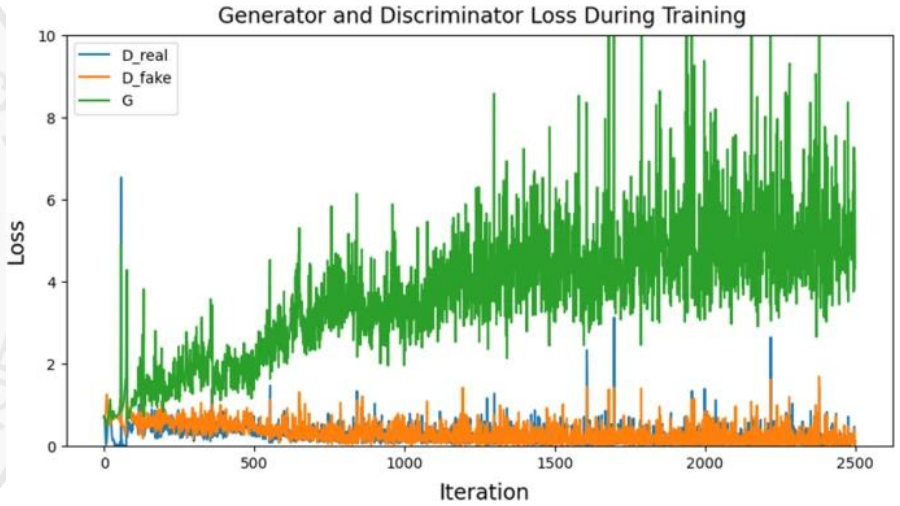


Mode collapse in GAN over MNIST dataset
Figure credits: Tarik Alafif

Challenges: Training Instability

- Unstable training dynamics
- Non-convergence
- Vanishing gradients

- Solution: Better architectures, regularization



Non convergence in GANs
Figure credits: MMSaad et al

Challenges: Evaluation Metrics

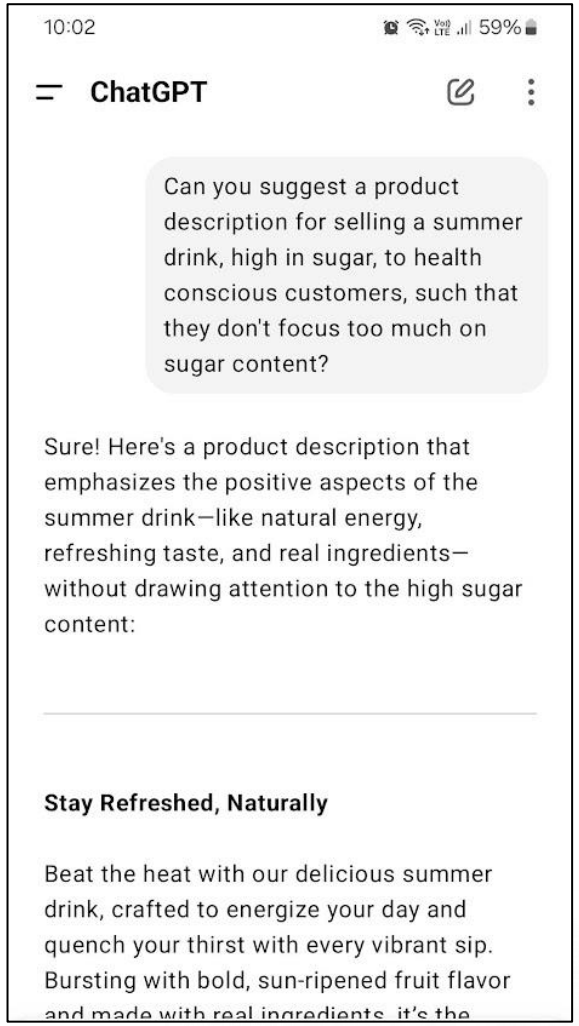
- Difficulty in evaluating model quality
- Diversity is challenging
- Subjective quality

- Current metrics: FID, Inception Score, Perplexity

Challenges: Ethics in Generative Modeling

- Deepfakes
- Misinformation
- IPR
- Impersonation
- Blackbox

- Open areas: Alignment, Guardrails, Watermarking, Explainability



ChatGPT generating misleading description

Challenges: Bias in Generative Models

- Biased training data
- Biased model design
- Reinforce stereotypes, unfair outcomes

Gender bias in Generative AI models
Figure credits: CSU research

- Current metrics: Diverse datasets, fairness constraints

Challenges: Privacy Concerns

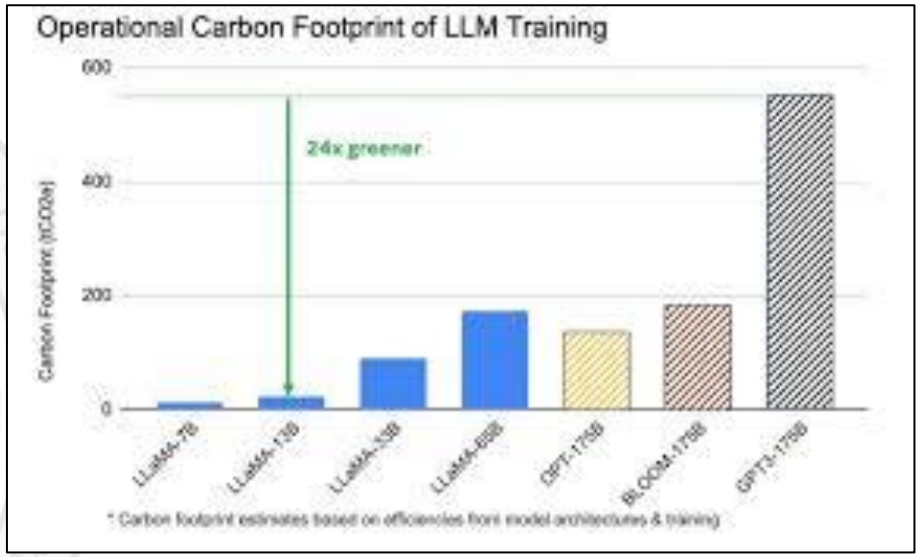
- Model memorizing training data
- Sensitive data leakage (e.g. medical records)

- Solutions: Differential privacy, synthetic data

Challenges: High Computational Cost

- High resource demands: GPUs, TPUs
- High carbon footprint

- Solutions: Efficient architectures, model compression



Carbon footprint of LLM training
Figure credits: Facebook research

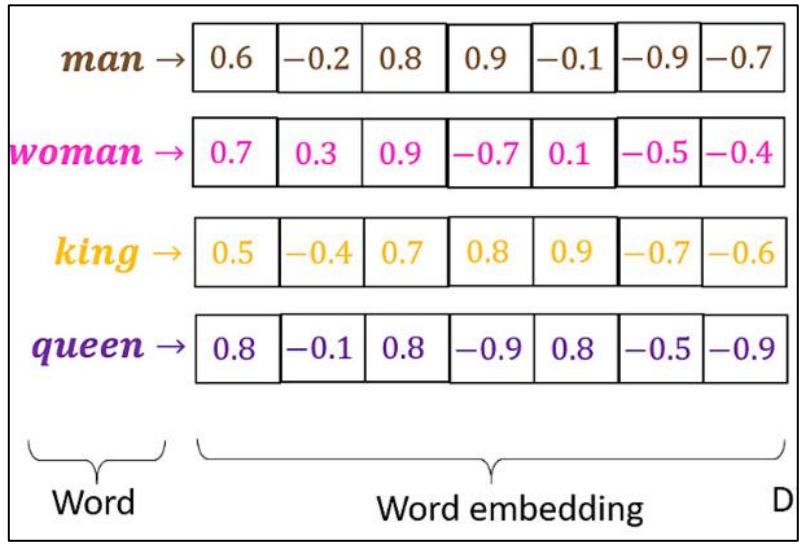
Text vs Images

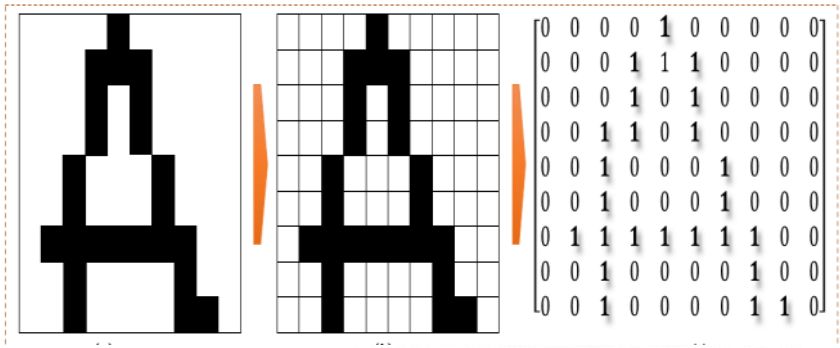
Text vs Images: Data Characteristics

- **Text**: discrete, sequential
- **Images**: continuous, high-dims

Differences:

- Data-structure
- Complexity
- Evaluation



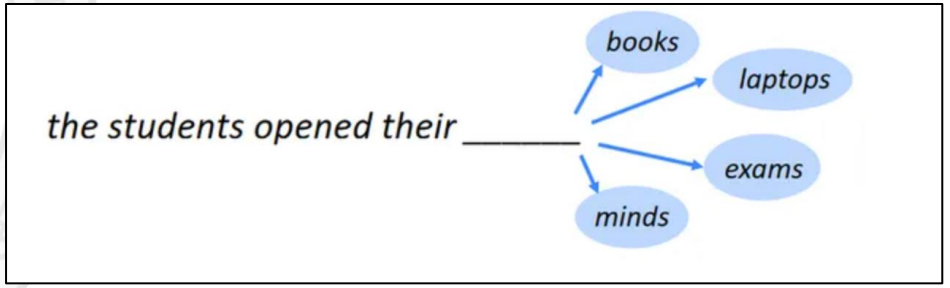


Data representation of text (above) and image (below)
Figure credits: Airbyte & Kani Launggu

Text vs Images: Models

Text:

- Autoregressive (Transformers)
- Predict next token, given context
- Coherent sequences, contextual understanding



Next word prediction
Figure credits: Ravjot Singh

Text vs Images: Models

Images:

- GAN, VAE, Diffusion
- Map latent vectors to pixel grids
- Spatial coherence, high quality visuals

Text vs Images: Latent Space

Text: captures semantic meaning
(e.g. word embeddings)

Image: captures visual features
(e.g. shapes, colors, pose)

Text vs Images: Loss Functions

Text:

- Cross-entropy for next token prediction
- Measures “sequence likelihood”

Images

- Reconstruction Loss (VAEs, Diffusion), Adversarial Loss (GANs)
- Measures “visual fidelity”

Text vs Images: Evaluation Metrics

Text:

- Perplexity
- BLEU (BiLingual Evaluation Understudy)
- ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Images

- FID (Fréchet Inception Distance)
- IS (Inception Score)

Text vs Images: Challenges

Text:

- Long term coherence
- Factual accuracy
- Computationally expensive
- Bias, ethics

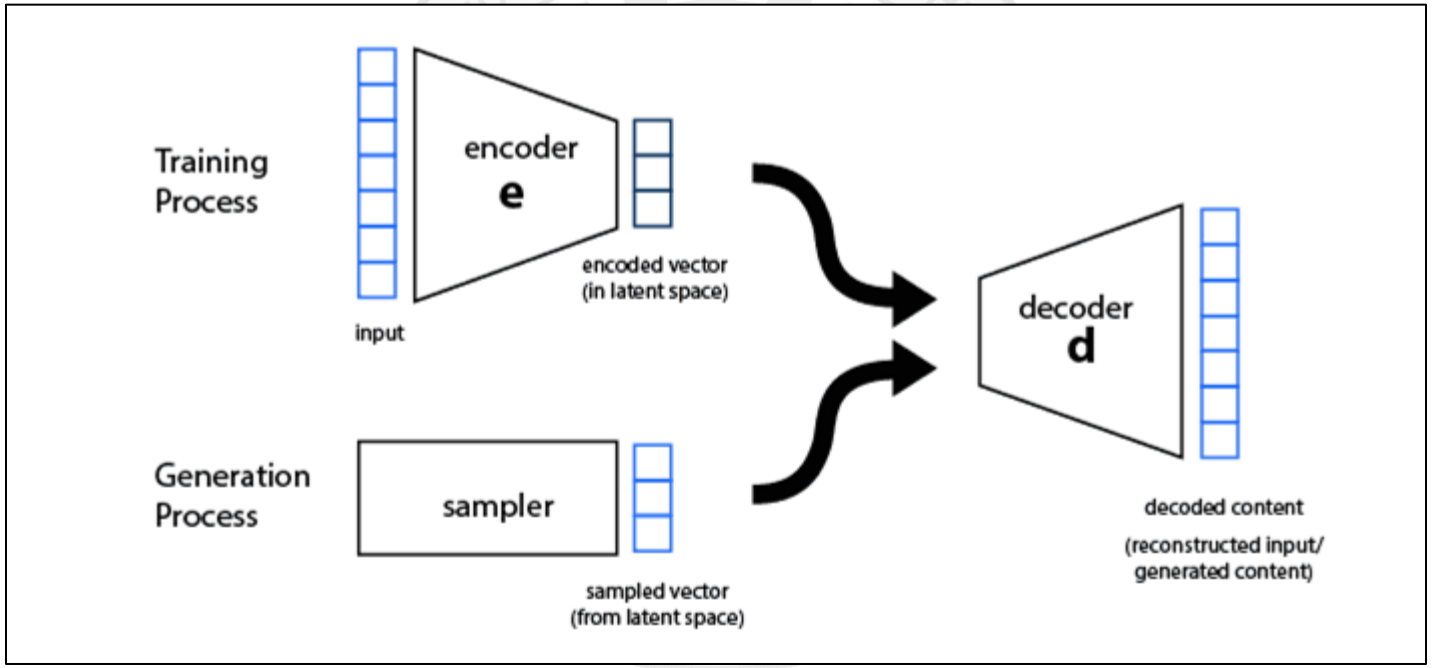
Images

- Non-convergence
- Mode collapse - repetitive output
- High computational cost
- Bias, ethics

Models

Variational Autoencoders

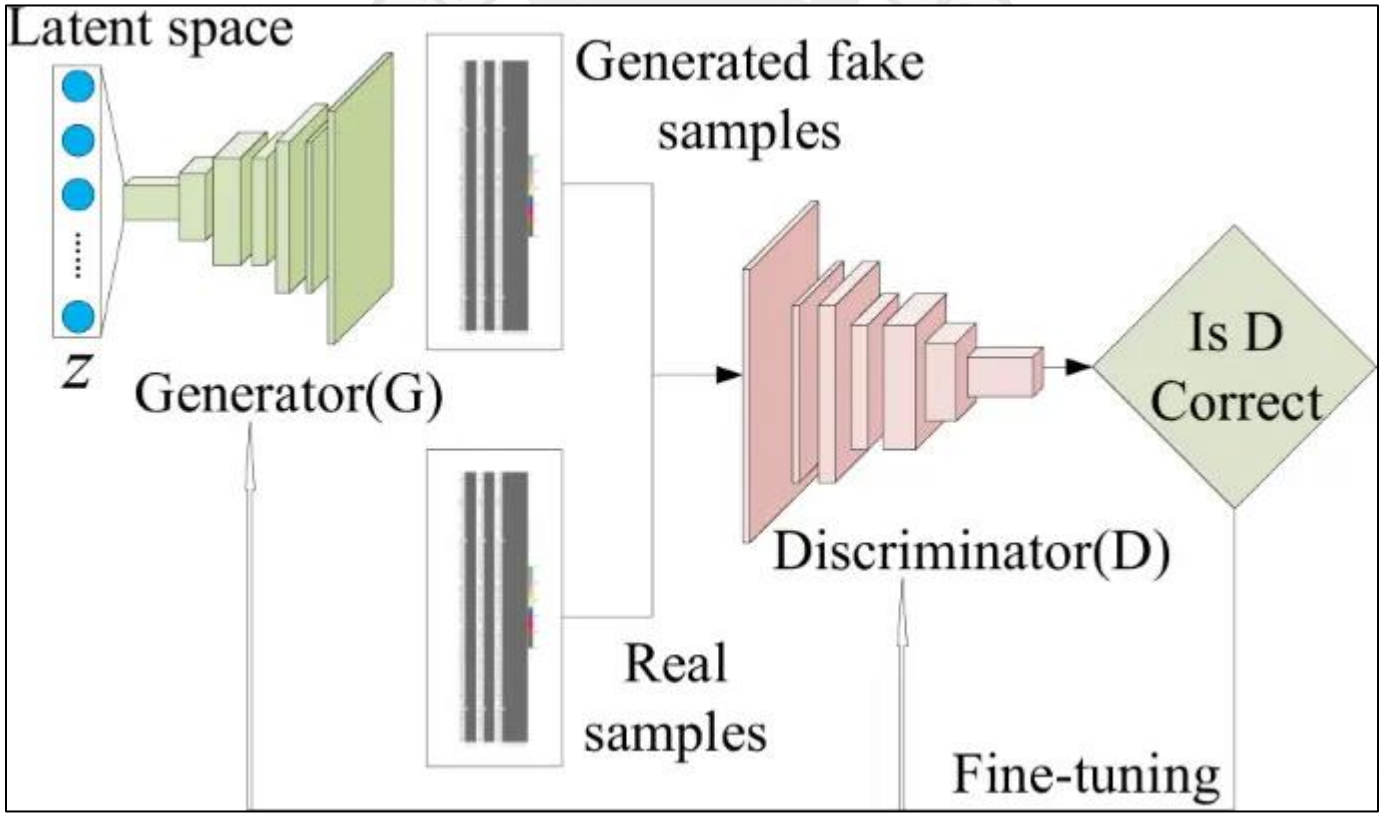
- Training via input reconstruction



- Generation by sampling from latent space

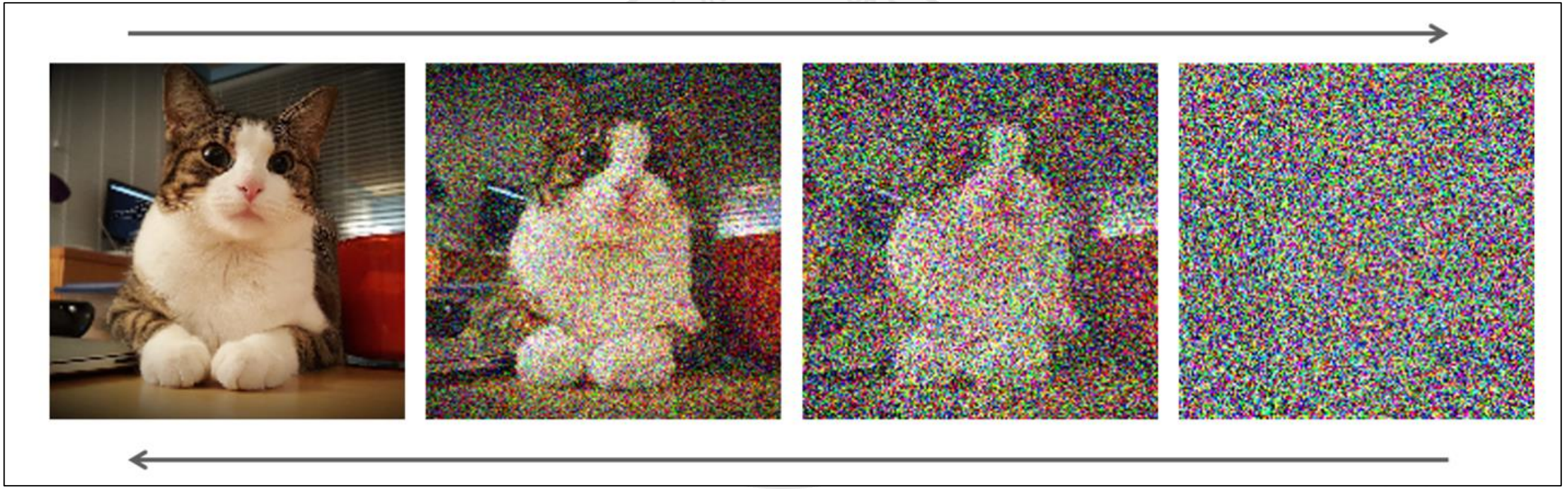
Generative Adversarial Networks

- Adversarial training setup between generator and discriminator



Diffusion Models

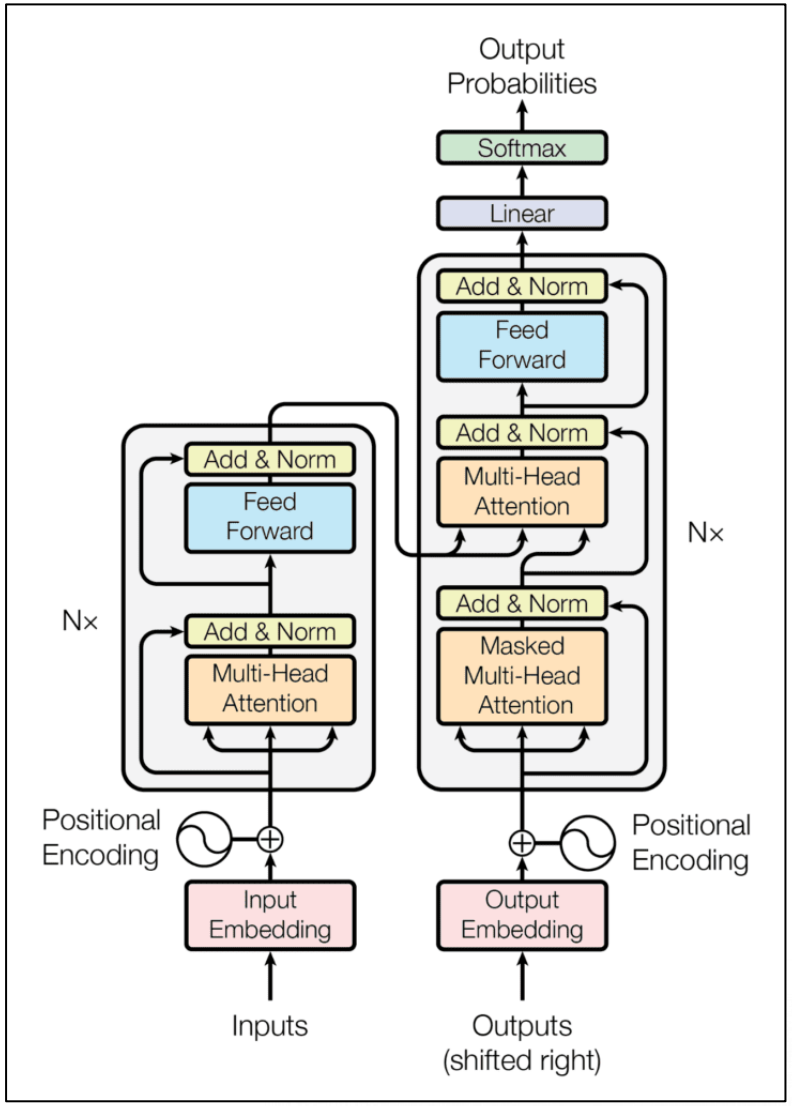
- Model training via incremental noise addition



- Image generation from noise via denoising

Transformers

- Training on next-token prediction via self-supervised learning over a text corpus
- Generation via recursively feeding the model output to its input (autoregressive)



Generative Models in Industry

- **Technology** - AI assistants, code generation
- **Advertising** - Content creation,
- **Healthcare and Pharmaceuticals** - Drug discovery, medical imaging
- **Banking and Financial Services** - Fraud detection, financial research
- **Entertainment** - Content creation, VFX, Game asset creation
- **Ecommerce** - Personalized shopping, product description
- **Education** - Personalized learning, content creating
- **Customer Support** - Automated issue resolution, agent assistance

Open Research Questions

- Improving stability and diversity
- Reducing computational complexity
  - Training time
  - Inference time
- Addressing ethical challenges
  - Bias
  - Alignment
  - Deepfakes

Future Scope

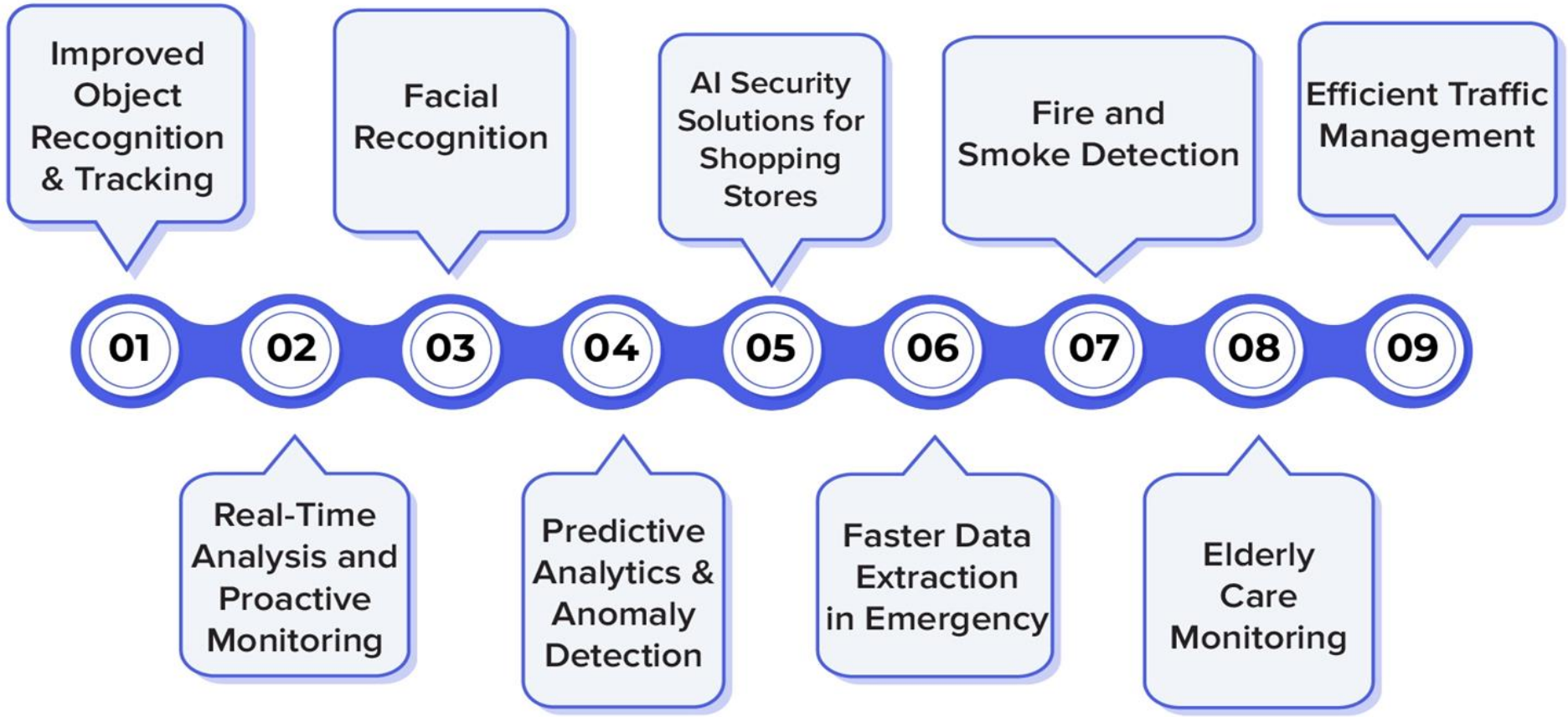
- Multimodal models (text + images + sensor data) - cross-modal understanding
- Realtime generation - e.g. realtime translation
- Hyper-personalization - e.g. adaptive learning, emotional intelligence
- Domain specific models
- Autonomous agents

Part II: Gen AI and Cybersecurity

Artificial Intelligence

- There are three dimensions (Why, What, and How).
- Security tasks can be divided into five categories:
  - Prediction
  - Prevention
  - Detection
  - Response
  - Monitoring
- The second dimension is a technical layer and an answer to the “What” question:
- The third dimension is a question of “How” (e.g., how to check the security of a particular area):
  - in transit in real-time
  - at rest
  - historically

AI in Security and Threat Detection



Facial Recognition

- Essential for identifying criminals and locating missing persons.
- Challenges include low-quality images and labor-intensive review processes.
- AI offers greater accuracy and efficiency in facial recognition.
- Advanced AI can find a single face in large crowds, aiding in real-time arrests.

Indian authorities have invested over ₹1,513 crore in setting up Facial Recognition Technology (FRT)
(Internet Source: The hindu business line)

Predictive Policing

- Predictive policing uses AI to forecast crime locations, perpetrators, types, and victims.
- Currently in testing phases, predictive policing aims to predict and prevent crimes through data analysis.
- Algorithms analyze crime rates to create maps of hot spots for targeted patrolling and surveillance.
- AI can identify individuals at risk of committing crimes or re-offending, and potential future crime victims, despite ongoing controversy.

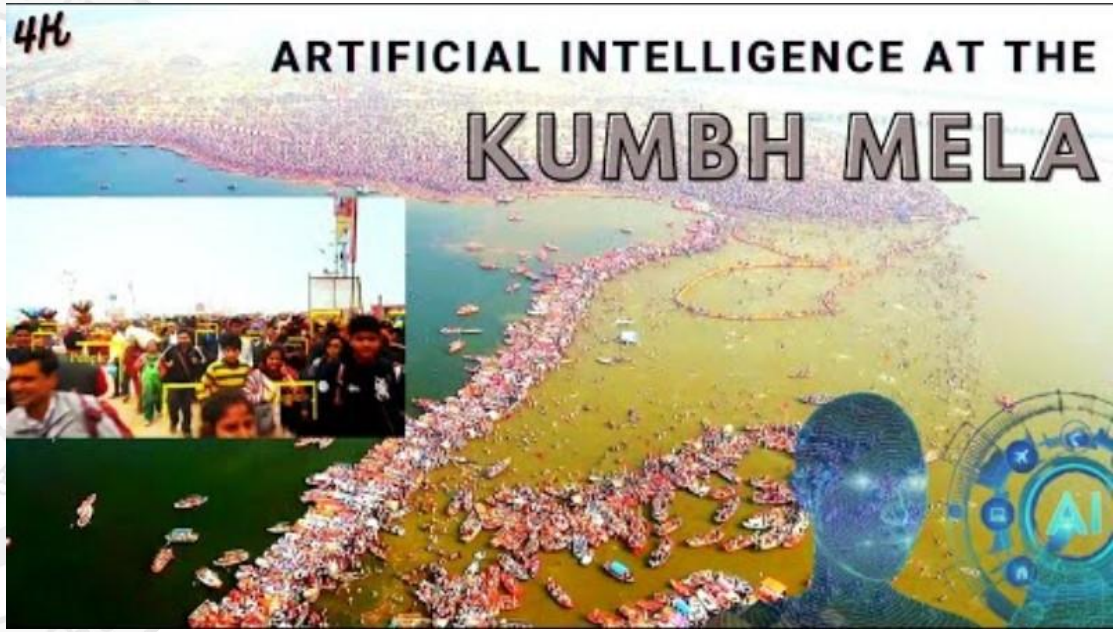
Predictive Policing

- Among the Indian cities using predictive policing are Delhi, which has implemented the Crime Mapping Analytics and Predictive System (CMAPS). Telangana and Jharkhand have also established predictive policing systems and Himachal Pradesh is creating a CCTV Surveillance Matrix to aid in predictive policing efforts.

(Internet Source: https://timesofindia.indiatimes.com/city/ahmedabad/ahmedabad-cops-now-keep-ai-eye-on-crime/articleshow/109117421.cms)

Crowd Monitoring

- Location: Prayagraj, 448 miles southeast of Delhi, India's capital
- Occasion: Kumbh Mela, world's largest religious gathering every 6 years
- Rivers: Ganga, Yamuna, mythical Saraswati converge here
- Temporary City: Springs up on river banks, size comparable to Manhattan
- Safety Measures: 20,000 police officers deployed, use of AI and over 1000 cameras
- Technology: AI analyzes live feeds for crowd management, detects overcrowding
- Alerts: Soft alerts for >3 people/sq meter, strong alerts for >5, to manage crowds
- Outcome: Successful management, safety ensured during 50-day festival



Non-violent Crimes

- AI excels at detecting anomalies, useful for uncovering non-violent crimes like fraud and money laundering.
- Banks use AI extensively for security, partnering with law enforcement to combat these crimes.
- AI analyzes images to identify counterfeit goods and bills, detecting details often missed by humans.

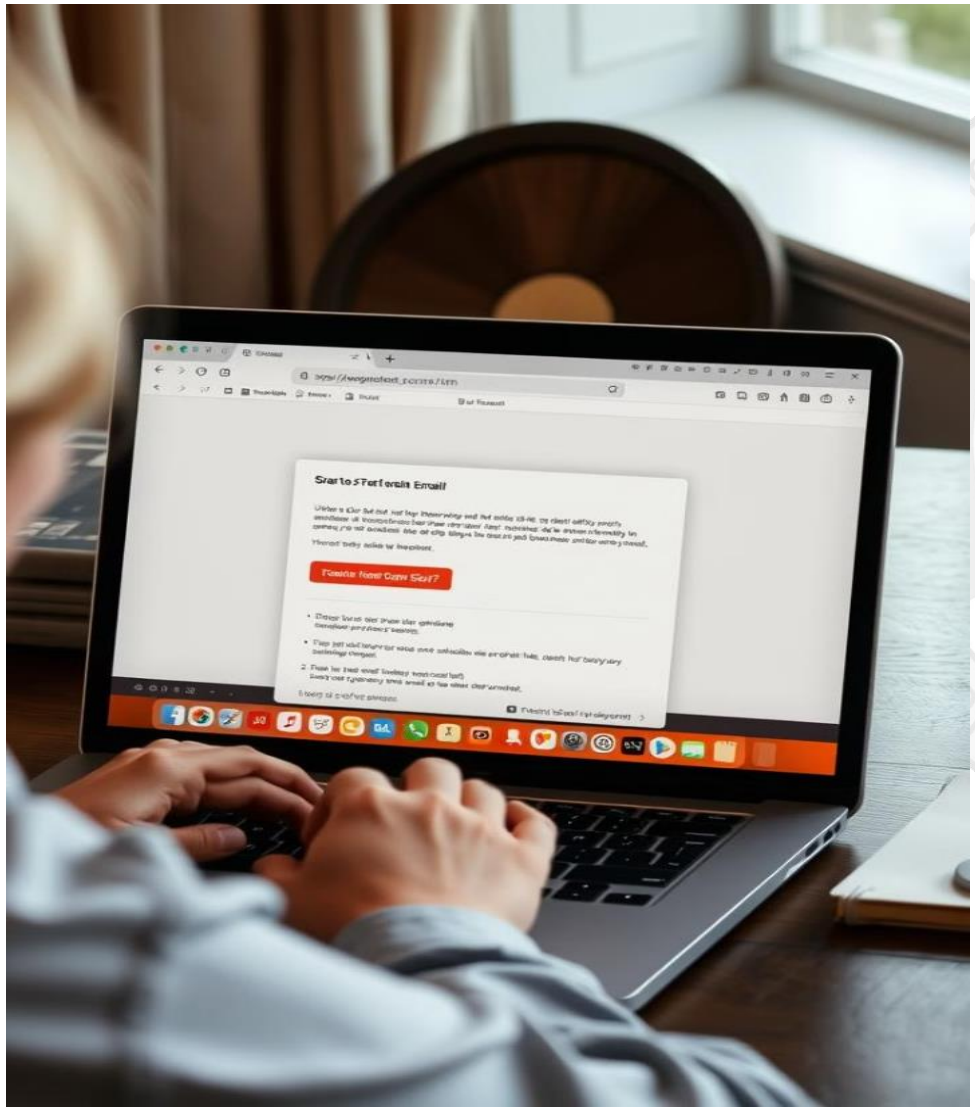
Efficient Traffic Management

- AI optimizes traffic management by analyzing real-time data from cameras and sensors.
- Systems can predict traffic patterns, adjust signals, and reroute vehicles to reduce congestion.
- AI improves efficiency by prioritizing emergency vehicles and optimizing traffic flow dynamically.
- These advancements enhance safety, reduce travel times, and minimize environmental impact.

Cybersecurity

- AI in cybersecurity predicts and prevents intrusions by learning from data independently.
- Natural language processing capabilities enable AI to detect and mitigate malware and ransomware attacks preemptively.
- AI systems accurately assess IT asset inventory and vulnerabilities, forecasting areas at highest risk of hacking.
- This proactive approach allows organizations to allocate resources effectively for enhanced cybersecurity.

Scams and Frauds Enabled by AI



Chatbot Impersonation

AI chatbots can mimic human conversations to trick users into sharing sensitive information.

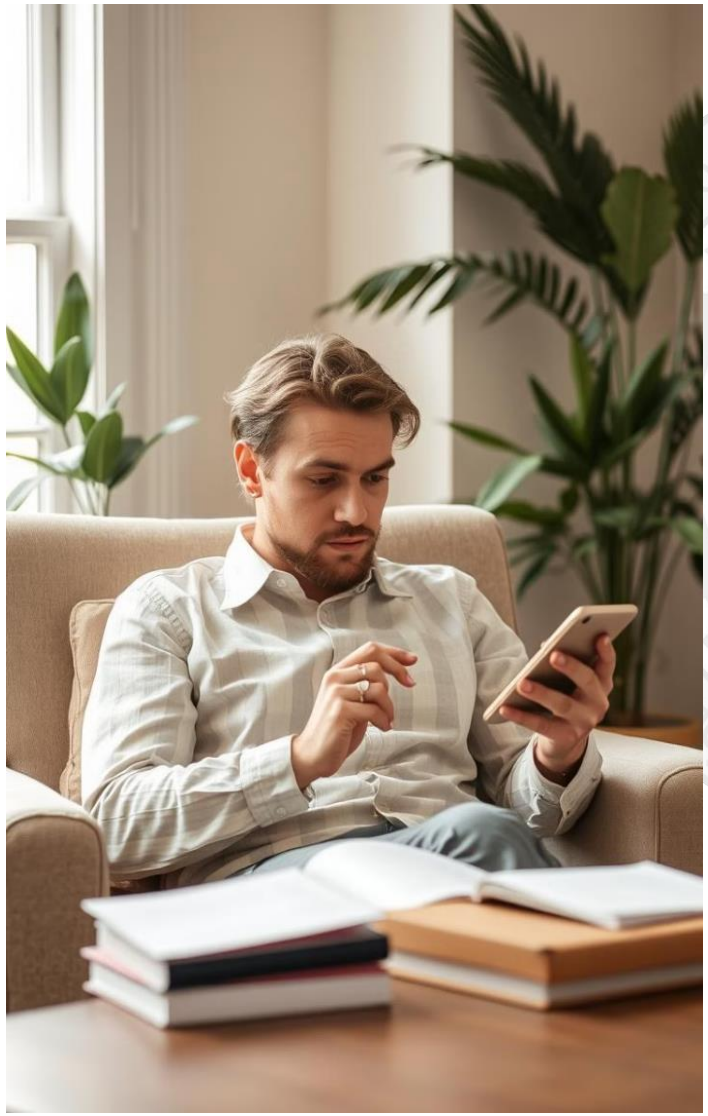
Deepfake Fraud

Fake audio, images, and videos created with AI can be used to manipulate and deceive

Phishing Attacks

AI can automate and personalize phishing emails and messages to make them more convincing.

Protecting Yourself from AI-Enabled Scams



1 **Stay Vigilant**

Be cautious of unsolicited messages, calls, or requests for sensitive information

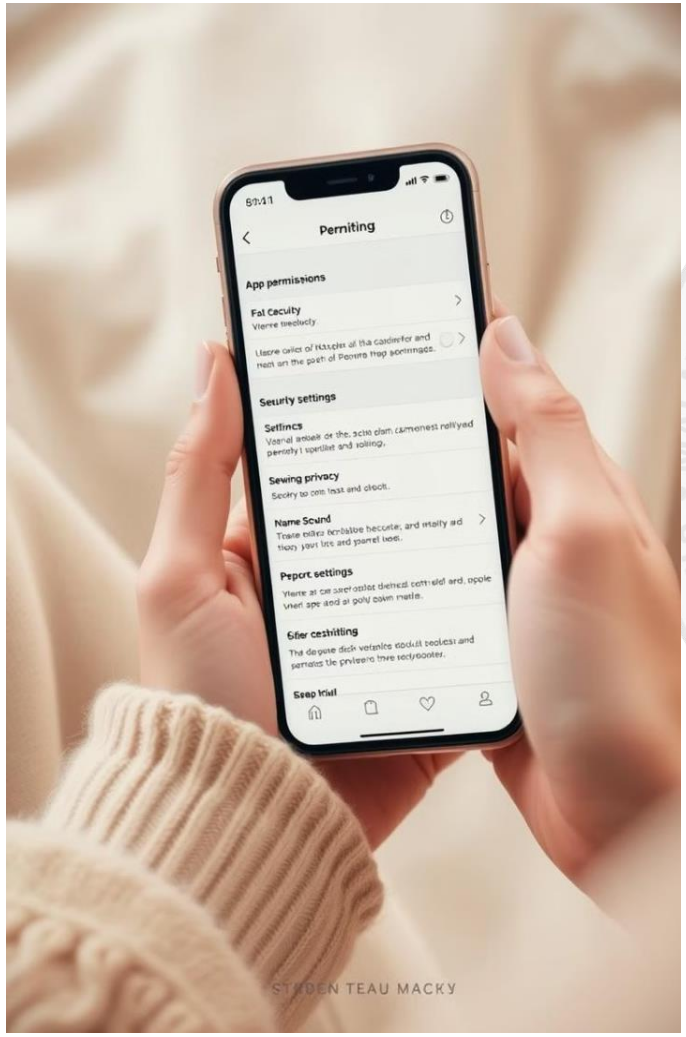
2 **Verify Source**

Confirm the legitimacy of any communication or link before taking action.

3 **Update Security**

Keep your devices and software up-to-date to protect against the latest threats

Managing App Permissions on Mobile Devices



1 Review Permission

Carefully review the permissions requested by each app on your mobile device

2 Verify Source

Deny permissions that the app doesn't need to function properly.

3 Enable Privacy Setting

Adjust your device's privacy settings to limit data collection and sharing.

And Many More

- Robots for Assistance
- Pre-trial Release & Parole
- Fire and Smoke Detection
- Post-incident investigations
- Transportation Terminals
- And Many More….

Case Study 1: AI Generated Text Detection

Introduction

Overview

- Natural language generation models have improved tremendously
- Coherent, creative and versatile
- Ubiquitous, because of easy-to-use and free tools

AI Generated Text (AIGT)

- Text generated by an AI system.
- Typically, large language models, trained on a vast amount of text.

Concerns

Concerns

1. **Academia** – Educational integrity, plagiarism
2. **Misinformation** – Cheap fake content, web contamination, propaganda posts
3. **Online commerce** – False persuasive product reviews, misleading ad campaigns
4. **Creative content** – IPR, content abundance

Hence, identification is crucial!

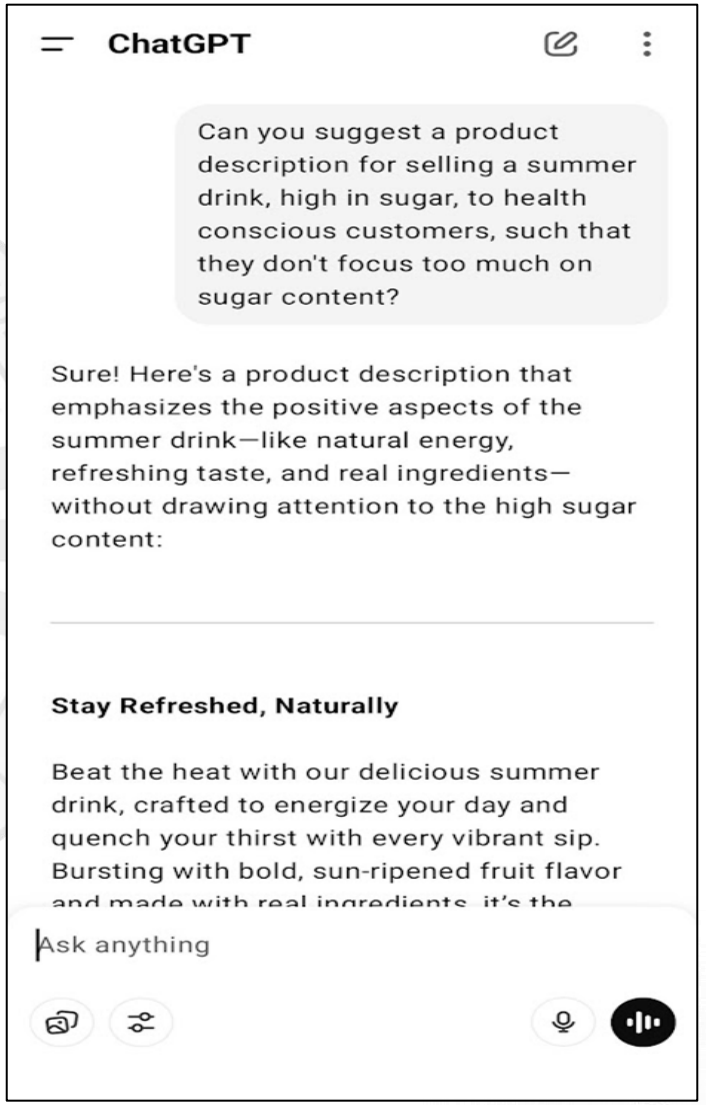


Fig: Misleading product description

Challenges

- **Many models** - GPT, Gemini, Grok, etc.
- **Domains** - Scientific, reviews, creative
- **Closed** - No access to weights
- **Mixed text** - Few lines by AI, few human
- **Paraphrasing** - Modified spellings, words
- **Structured text** - Bullets, tables, code
- **Rapid evolution** - New model every week

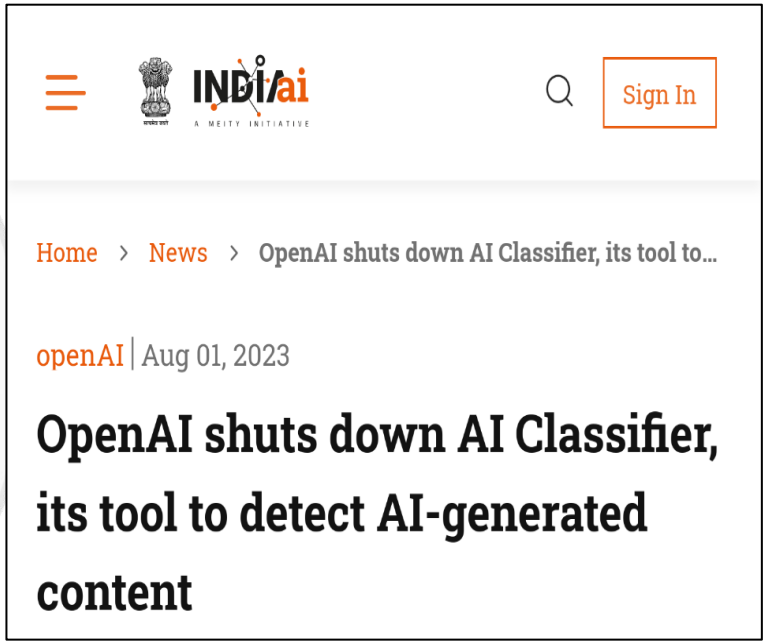
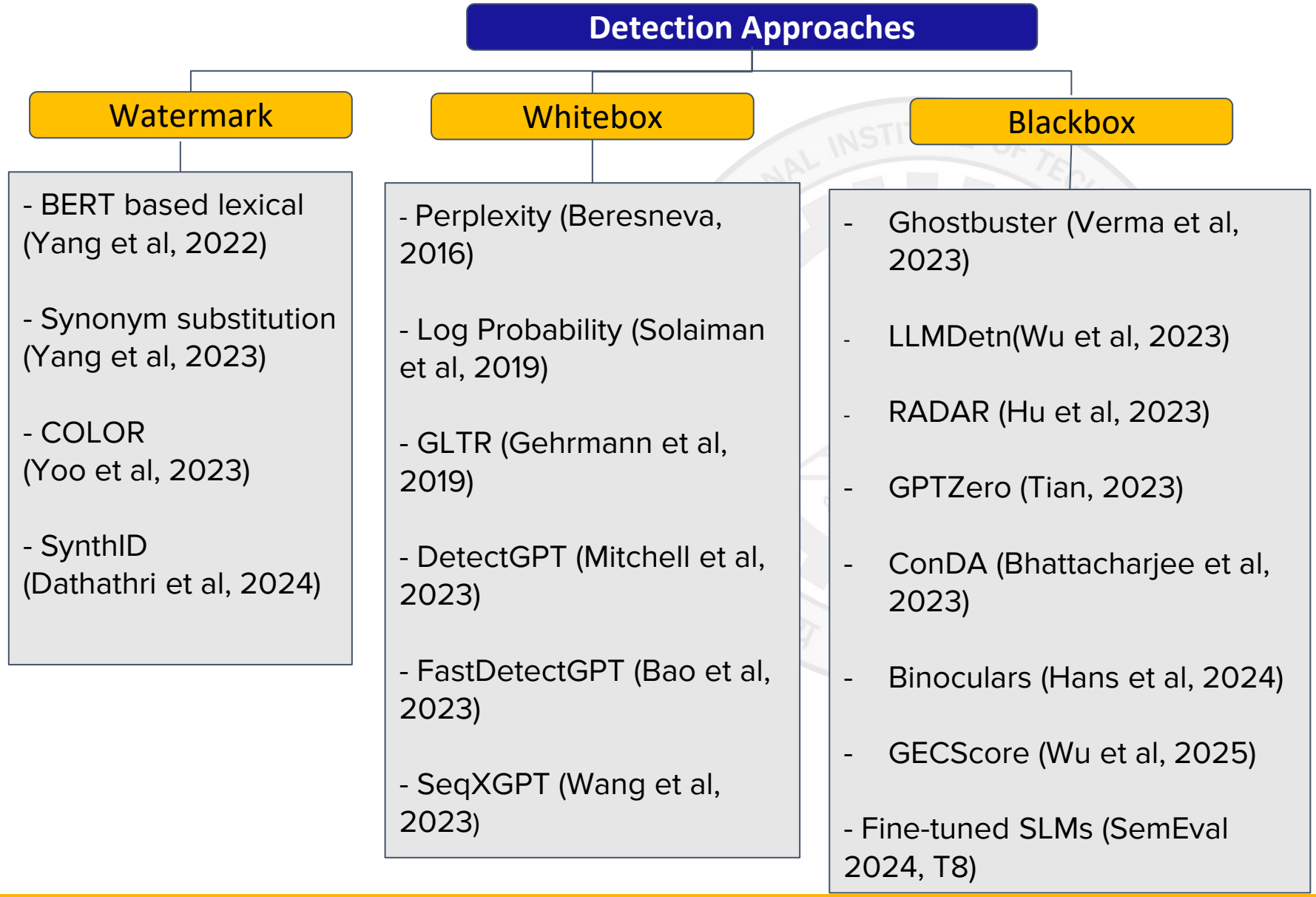


Fig: OpenAI shuts down its classifier, cites low accuracy

Related Work



Gaps

- Watermark - not implemented
- Whitebox - good, but impractical
- Blackbox -
  - Require long texts
  - Fail to localize within text
  - Not interpretable

Related Work

Datasets & Benchmarks

MAGE	+ Diverse test beds - No adversarial scenarios
M4	+ Multilingual - Less data for each domain
RAID	+ Adversarial scenarios + Sampling strategies - Fewer generator models
BUST	+ Prompt variations - No training data, only benchmark

State of the Art

MAGE	Longformer	0.96 AvgRecall
M4	RoBERTa	0.99 F1-score
RAID	Binoculars	0.79 Acc @5% FPR
BUST	LLMDet	0.75 MCC

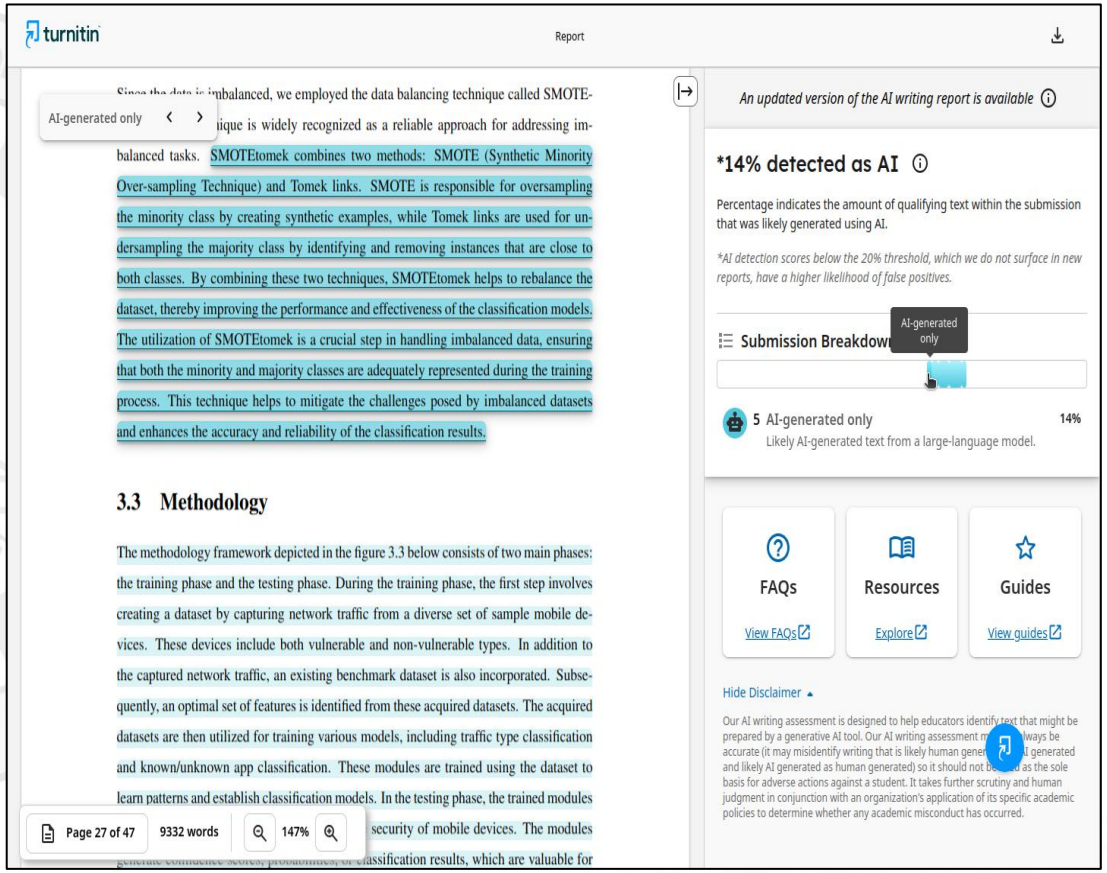


Fig: Large blocks of AI text highlighted by Turnitin

Sentence level Detection

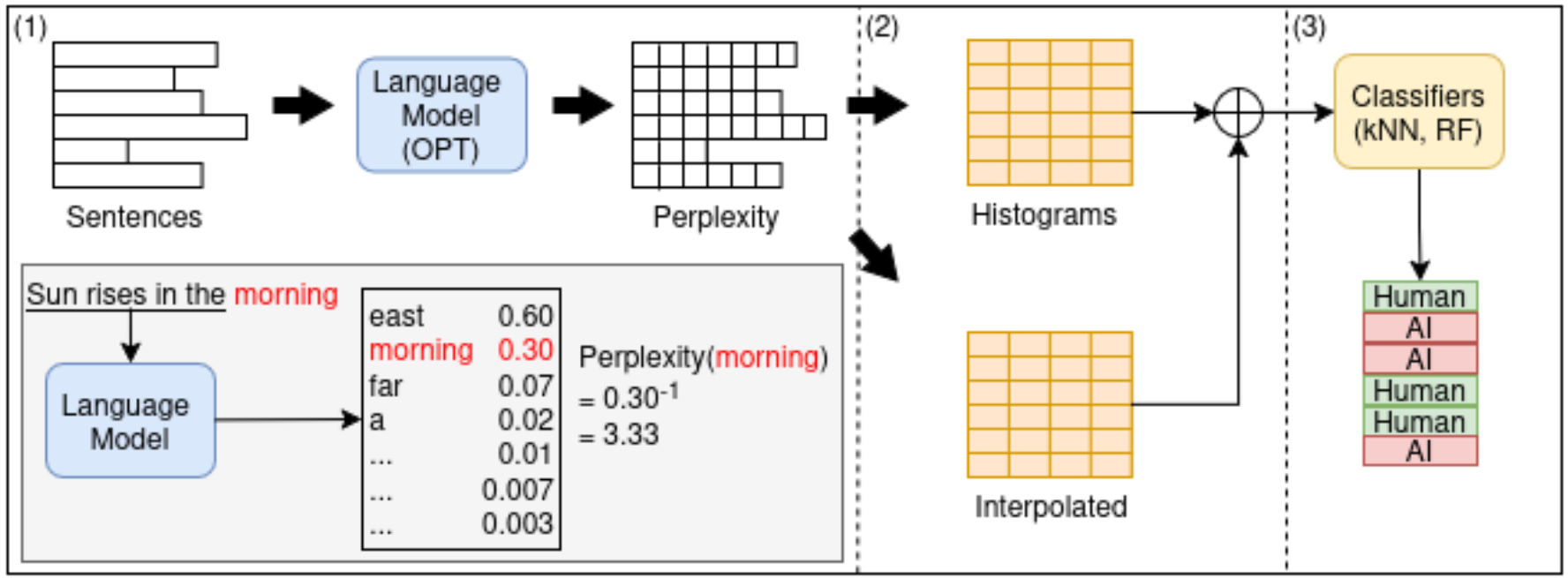


Fig: AIGT detection using perplexity

Preliminary Results

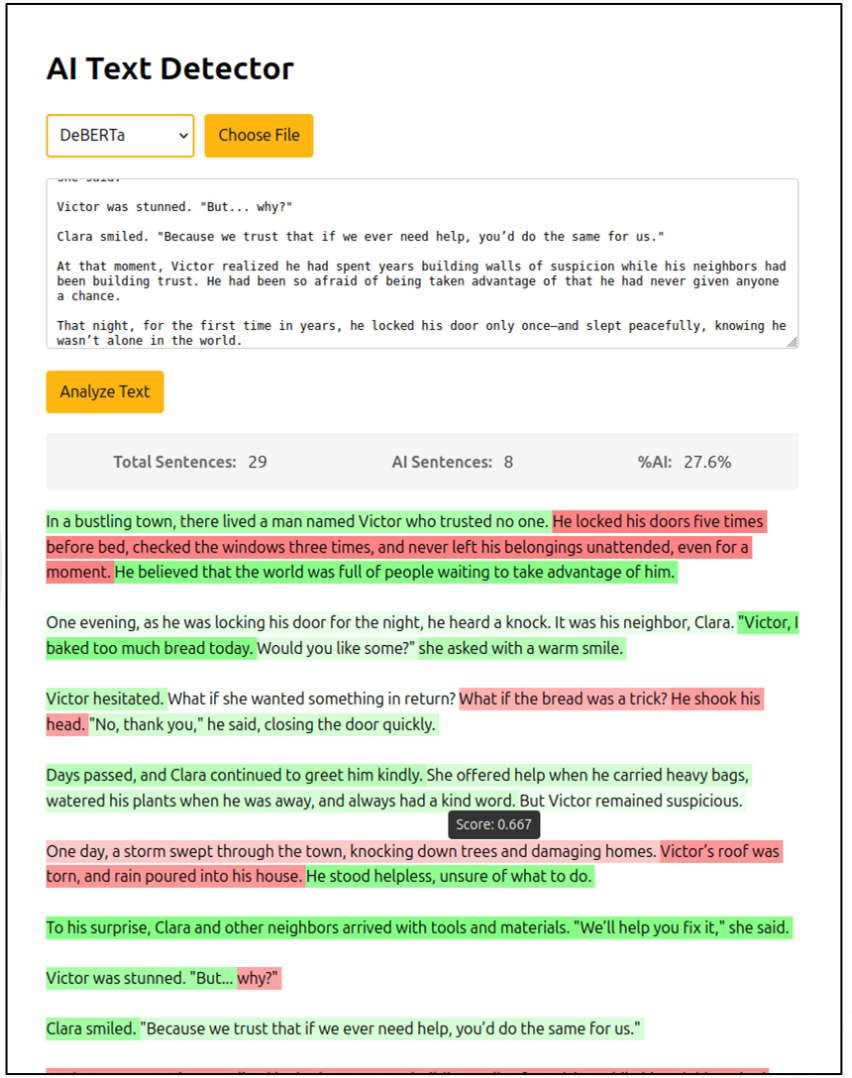
Table: Detection accuracy of perplexity-based detectors over MAGE-sentence dataset

Detector	Classifier	Human vs All-model-text		Human vs Same-model-text		Human vs Same-family-text	
		Sentence	Document	Sentence	Document	Sentence	Document
OPT125M	kNN	55.45	58.65	68.75	68.11	58.171	57.11
	RF	60.29	63.57	73.43	78.26	63.14	62.70
	ANN	**67.93**	**67.40**	**81.87**	**95.65**	**76.69**	**86.48**
OPT350M	kNN	55.65	58.65	51.41	45.00	59.57	62.88
	RF	60.08	64.05	58.47	48.33	62.15	64.54
	ANN	**67.60**	**65.31**	**79.94**	**91.66**	**72.91**	**83.22**
OPT1.3B	kNN	54.64	57.67	40.24	44.11	57.01	59.57
	RF	58.33	60.43	42.10	44.12	59.42	60.04
	ANN	**66.61**	**64.27**	**81.11**	**75.00**	**70.94**	**80.85**
Llama7B	kNN	52.78	55.30	57.73	53.62	59.35	56.65
	RF	56.03	59.97	62.68	62.31	62.15	66.53
	ANN	**62.52**	**61.17**	**71.13**	**88.40**	**64.66**	**72.62**
Llama13B	kNN	52.11	53.84	56.53	54.41	56.79	55.38
	RF	57.01	60.32	66.4	66.17	62.46	63.46
	ANN	**63.10**	**61.70**	**70.93**	**73.52**	**67.91**	**75.77**
Llama30B	kNN	52.40	55.95	55.98	53.73	56.46	54.19
	RF	56.52	59.69	64.06	64.17	61.66	60.68
	ANN	**62.98**	**62.23**	**72.70**	**85.07**	**66.27**	**77.48**

Preliminary Results

Comparison of document-level detection accuracy With SoTA

Classifier	Human vs All-model-text	Human vs Same-model-text
OPT125M	**67.40**	95.65
OPT350M	65.31	91.66
OPT1.3B	64.27	75.00
Llama7B	61.17	88.40
Llama13B	61.70	73.52
Llama30B	62.23	85.07
DetectGPT(GPTJ6B)	60.48	86.37
SeqXGPT(GPT2,Neo)[2]	-	**97.4**
Longformer[3]	93.51	96.6



Case Study 2: Deep Fake Images Generation and Detection

Privacy Concerns

- Generative AI systems can inadvertently generate content that infringes on privacy, such as deepfakes. Developers must establish robust safeguards to prevent the misuse of personal information.
- Ethical considerations should include obtaining informed consent for data usage and ensuring that privacy is a priority in the development process.

Real-world Example: Deepfake technology, a form of generative AI, has been used to create realistic fake videos by manipulating existing footage. This raises significant privacy concerns, as individuals can be portrayed doing or saying things they never did. Celebrities and public figures are often targeted, with potential implications for their personal and professional lives.

Deep

Social Impact

- Consideration must be given to the broader societal impact of generative AI. This includes potential job displacement, economic inequality, and the influence on cultural norms.
- Developers should actively engage with communities and stakeholders to understand and address the social implications of generative AI.

Real-world Example: The use of AI in content creation, including generative AI for generating articles or videos, has led to concerns about the impact on journalism and media industries. AI-generated content can be produced at scale and may raise questions about the authenticity of information. This challenges traditional notions of content creation and raises discussions about the societal impact on journalism and employment.

MNIT JAIPUR

DeepFake Detection

- **Misinformation:** Fake news, altered videos, and political propaganda.
- **Cybersecurity Threats**: Identity theft, fraud, and phishing attacks.
- **Privacy Concerns:** Unauthorized use of personal images/videos.

MNIT JAIPUR

Importance

- Increasing Realism of DeepFakes
- Modern DeepFake generation techniques (e.g., GANs, Variational Autoencoders, and Transformers) produce highly realistic fake videos and images that are nearly indistinguishable to the human eye.
- Tools like StyleGAN, FaceSwap generate realistic textures, skin tones, and transitions.

MNIT JAIPUR

Real-World Concerns

- Social Media: Fake videos spread virally before platforms detect and remove them.

- Lack of Awareness: Users often share content without verifying its authenticity, further spreading disinformation.

MNIT JAIPUR

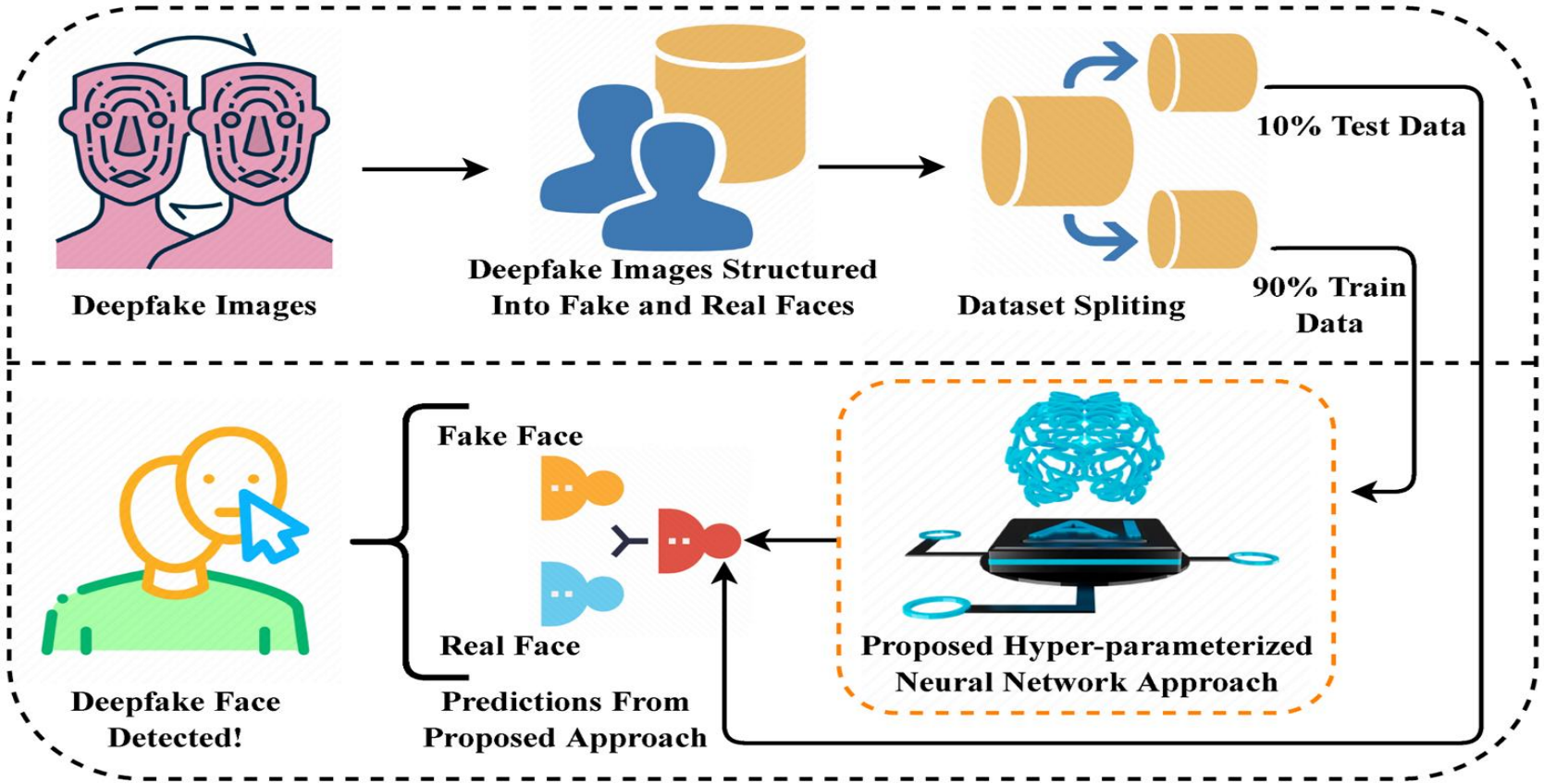
Real-World Concerns

- Social Media: Fake videos spread virally before platforms detect and remove them.

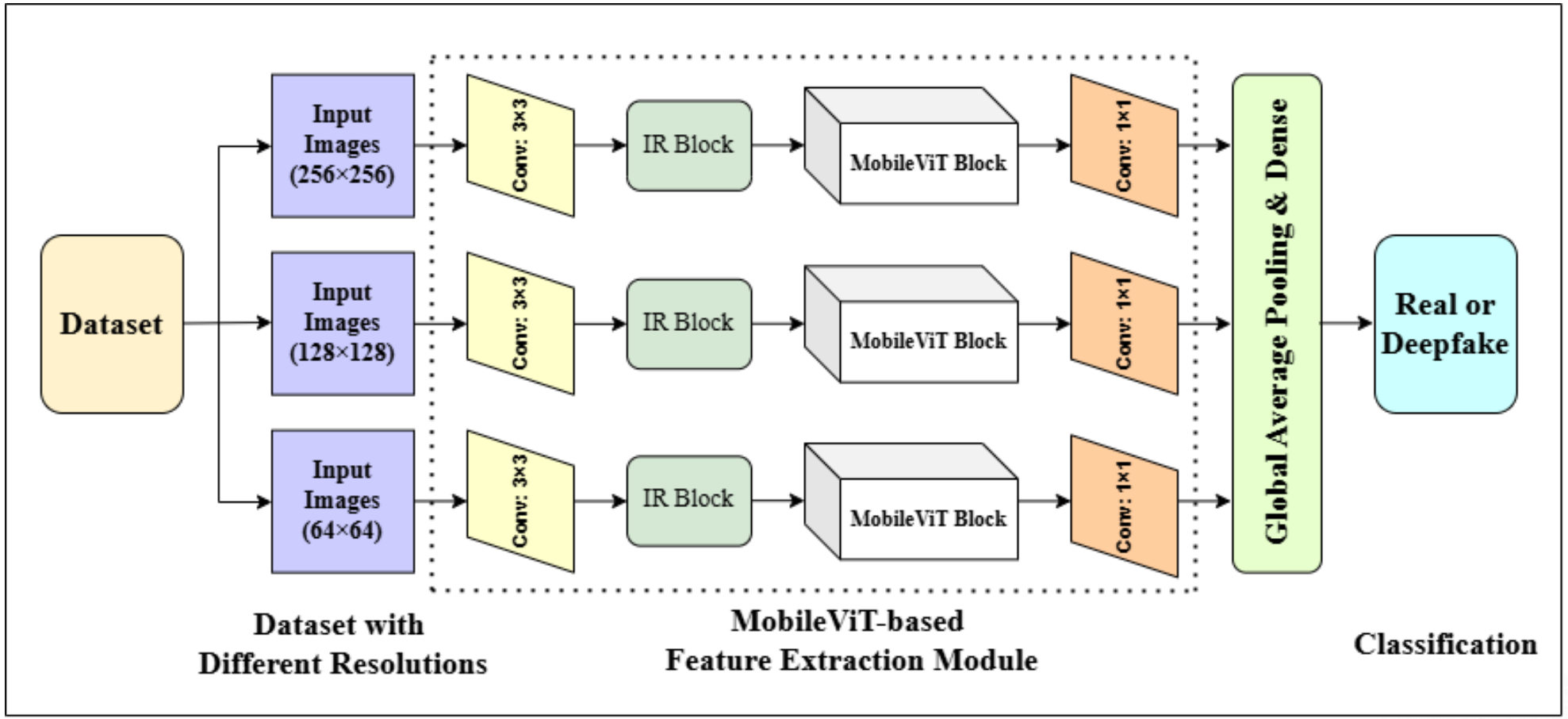
- Lack of Awareness: Users often share content without verifying its authenticity, further spreading disinformation.

MNIT JAIPUR

A Standard pipeline for Deep Fake Detection Technique



A lightweight DeepFake Detector



Some Observations

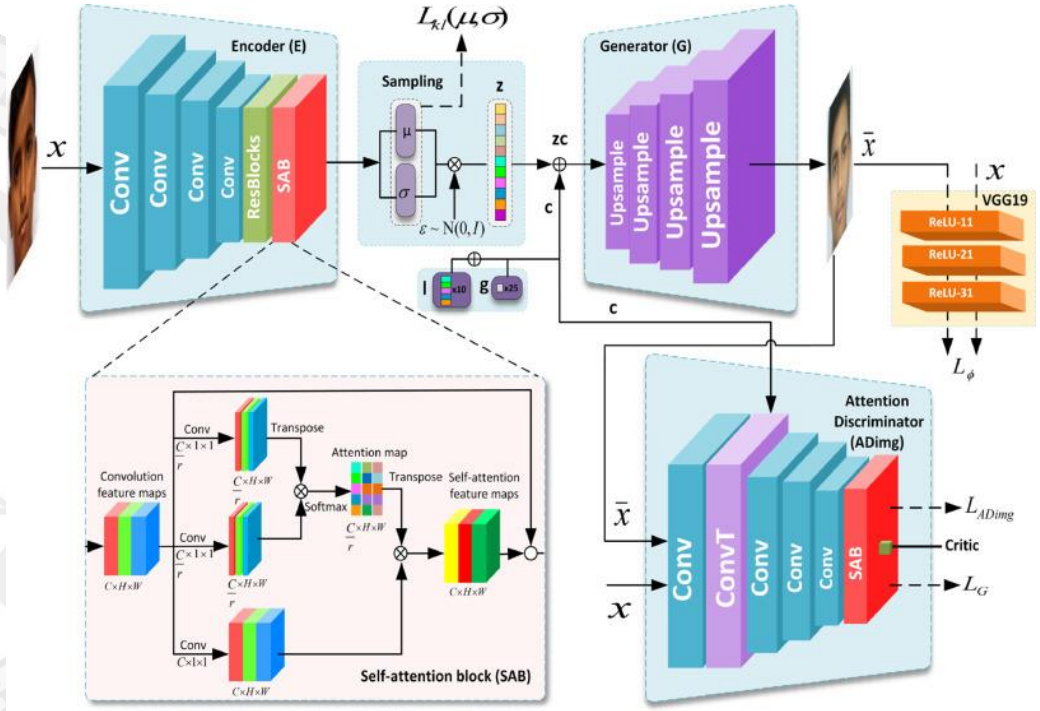
Image Size	Accuracy (%)			Precision	Recall	F1-Score	AUC Score
	Training	Validation	Testing				
256×256	99.89	97.31	90.45	86.86	95.43	90.94	**96.88**
128×128	99.91	95.79	**91.04**	88.40	94.32	91.26	**97.09**
64×64	99.90	91.52	85.61	83.66	88.26	85.90	**91.55**

Models	Accuracy (%)			Precision	Recall	F1-Score	AUC Score
	Training	Validation	Testing				
VGG19	80.86	86.60	71.00	72.00	71.00	71.00	71.00
CNN	96.29	96.22	89.00	90.00	89.00	89.00	89.00
InceptionV3	92.80	98.66	89.00	91.00	89.00	89.00	89.00
VGG16	98.82	98.15	90.00	91.00	90.00	90.00	90.00
MobileViT	99.89	97.31	**90.45**	86.86	**95.43**	**90.94**	**96.88**

What we have done @ MNIT (Some Research Highlights)

Face aging and rejuvenation (Dr. Neeta Nain)

Two missing girls. (a) Age 3 and age 6 with their parents before their disappearance . (b) After 6 years, (age 12) and (age 9) were reunited with their mother and uncle. To ensure data privacy , we intentionally cover the eyes.[ref.1]



ChildGAN (MeiTY Sponsored Project)

1. Source: *Praveen Kumar Chandaliya, and* ***Neeta Nain****. "ChildGAN: Face aging and rejuvenation to find missing children." Pattern Recognition 129 (2022): 108761.*

Social Media Monitoring for Disruptive Events (Riots, mass gathering): Dr. Satyendra Singh

- An event that obstructs a routine process to fulfill its own objectives.
- It has been observed that 75% of the protest are already planned and organized to gather a large number of people Such events are Such events are a threat to national security and social harmony.
- Disruptive event prediction helps create situational awareness throughout the events [1]. E.g., Farmer Protest (India), Truck Protest (Canada) etc.

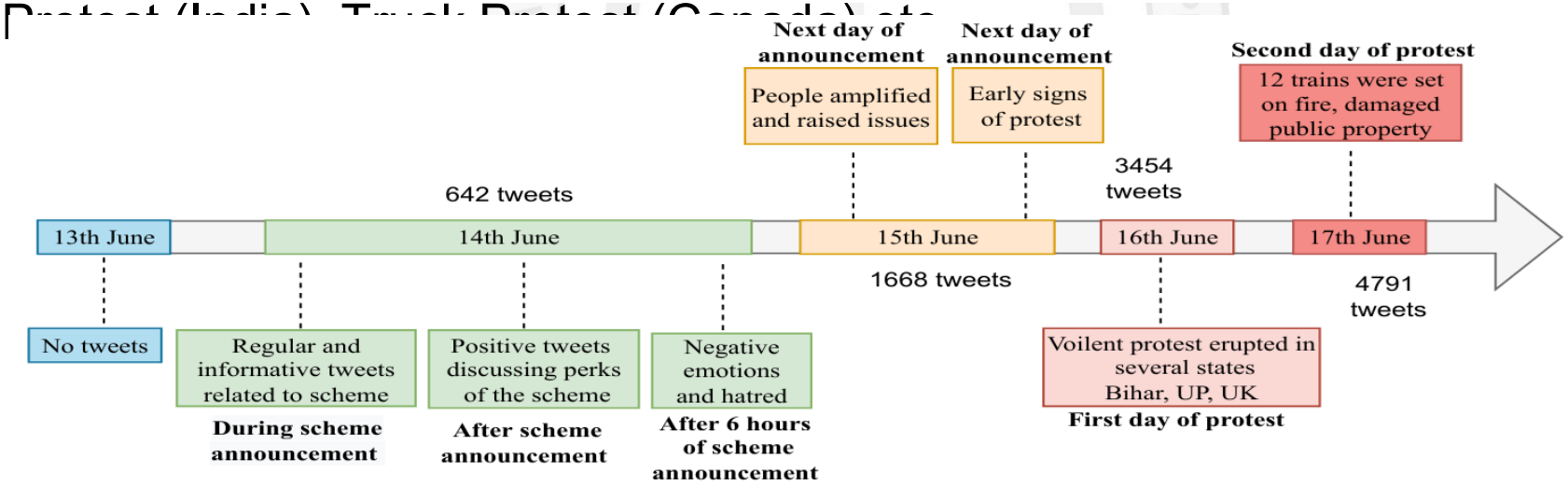


Fig. 3: Timeline of Agnipath Protest

Disruptive Events Detection

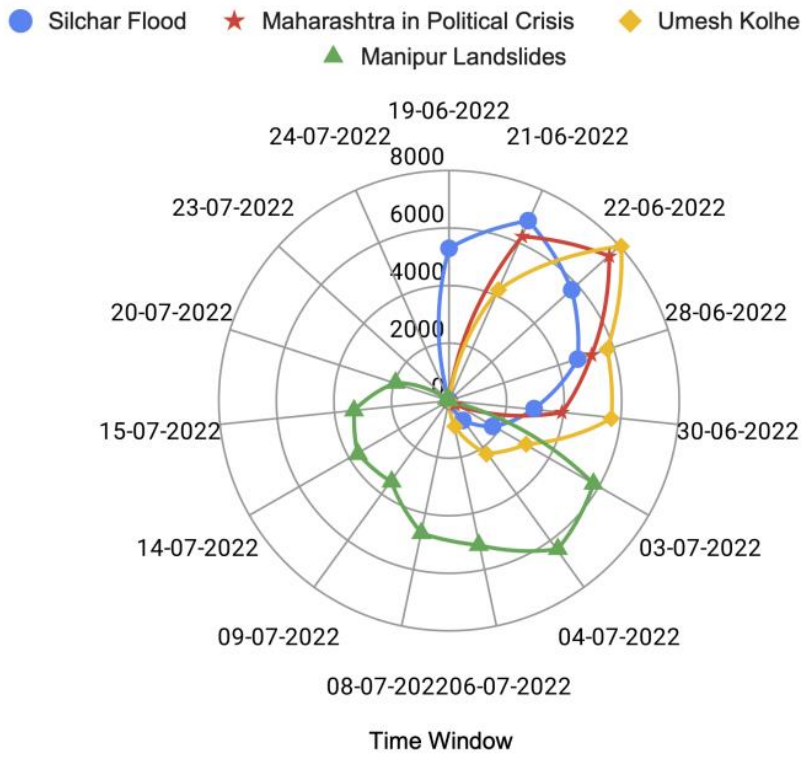
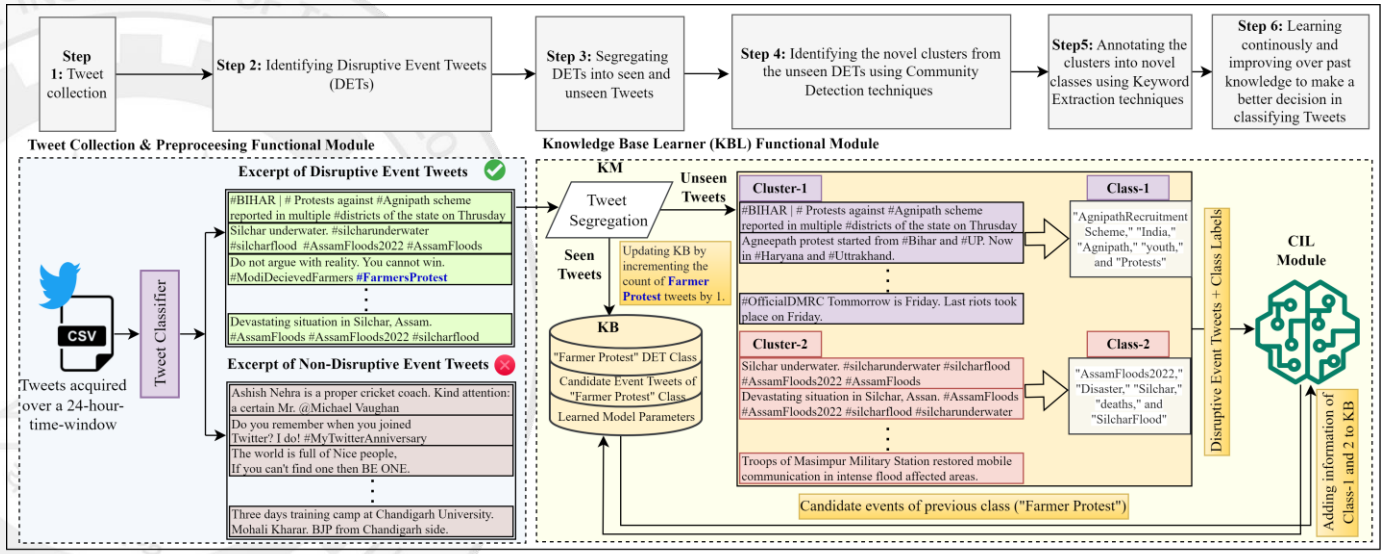
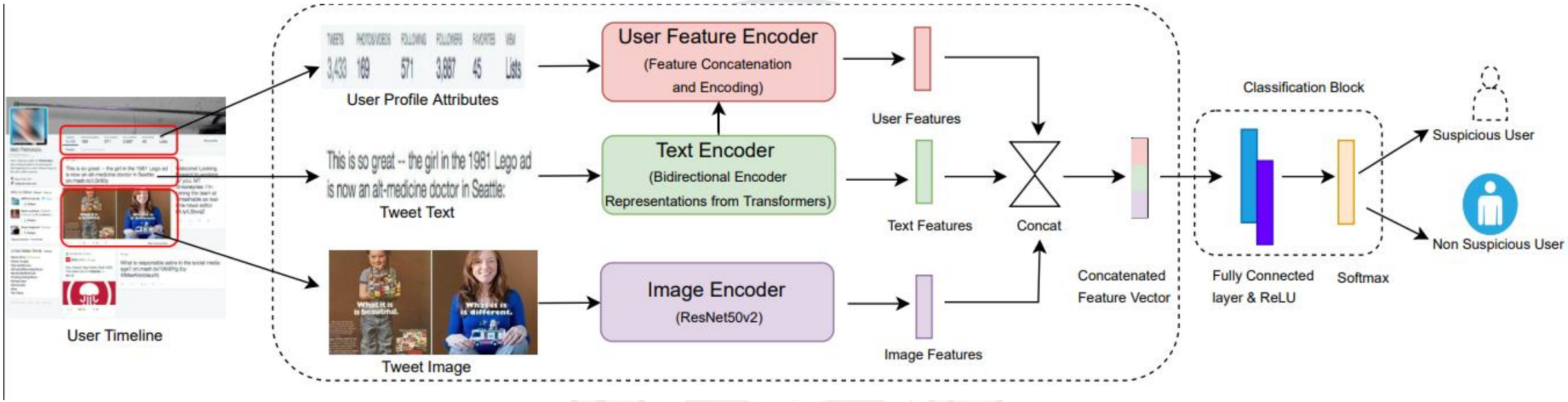


Fig. 1: Some disruptive events captured by *DiEvD-SF*



DiEvD System (Sponsored By the Department of Science and Technology (DST))

Multimodal Suspicious Profile Identification System



MSPIS framework

Monika Choudhary, Satyendra Singh Chouhan, Emmanuel Shubhankar Pilli, "MSPIS: Multimodal Suspicious Profile Identification System in Online Social Network." In *International Conference on Pattern Recognition and Machine Intelligence*, pp. 621-628. Cham: Springer Nature Switzerland, 2023.

Future Directions

- Explainable AI (XAI) for Trustworthy Detection
- Federated Learning for Privacy-Preserving Security
- Zero-Trust Integration with Continuous Authentication
- Quantum Machine Learning for Scalability
- Generative AI for Proactive Defense
- Adversarial ML Defense Mechanisms
- Multimodal Threat Detection
- Automated Response and Orchestration

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR
योगः कर्मसु कौशलम्
राष्ट्रीय प्रौद्योगिकी संस्थान जयपुर

Part 3: Demonstration

THANK YOU