“Para o prazer e para ser feliz, é que é preciso a gente saber tudo, formar alma, na consciência; para penar, não se carece.”

(Guimarães Rosa *in* **Grande Sertão: Veredas**,1956)

Introdução à Inteligência Artificial

Roteiro da aula:

- **IA Generativa**
- ***Large Language Models (LLM)***
- **Aprendizagem não supervisionada**
- **Modelagem Gerativa**
- **VAE (*Variational Autoencorders*)**
- **GAN (*Generative Adversarial Networks*)**

Com slides adaptados de A. Soleimany, L. Fei-Fei, R. Krishna, D. Xu, M. Lapata

Contexto

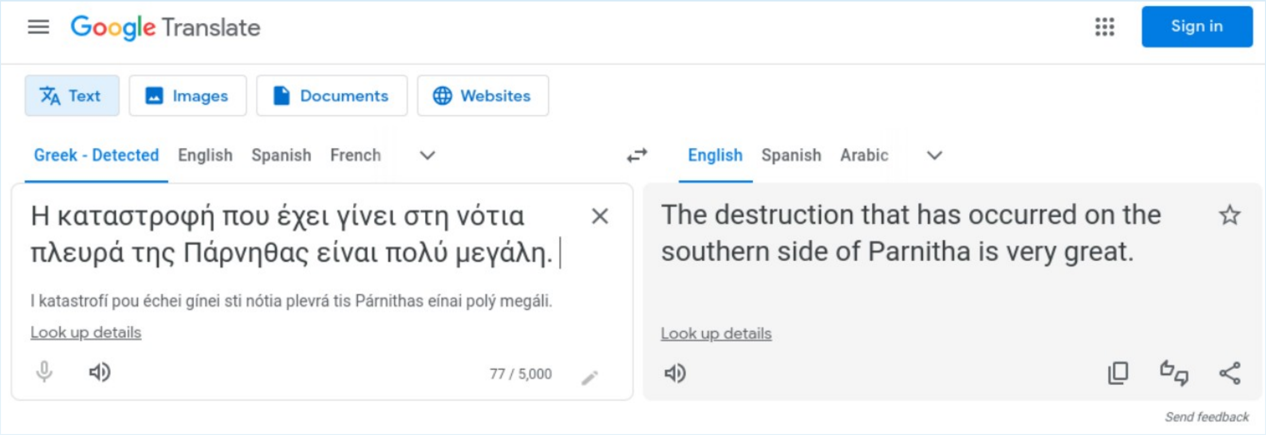
What is Generative AI?

Generative — create new content (audio, code, images, text, video)

Artificial Intelligence — automatically using a computer program

Crédito imagem/slide: M. Lapata

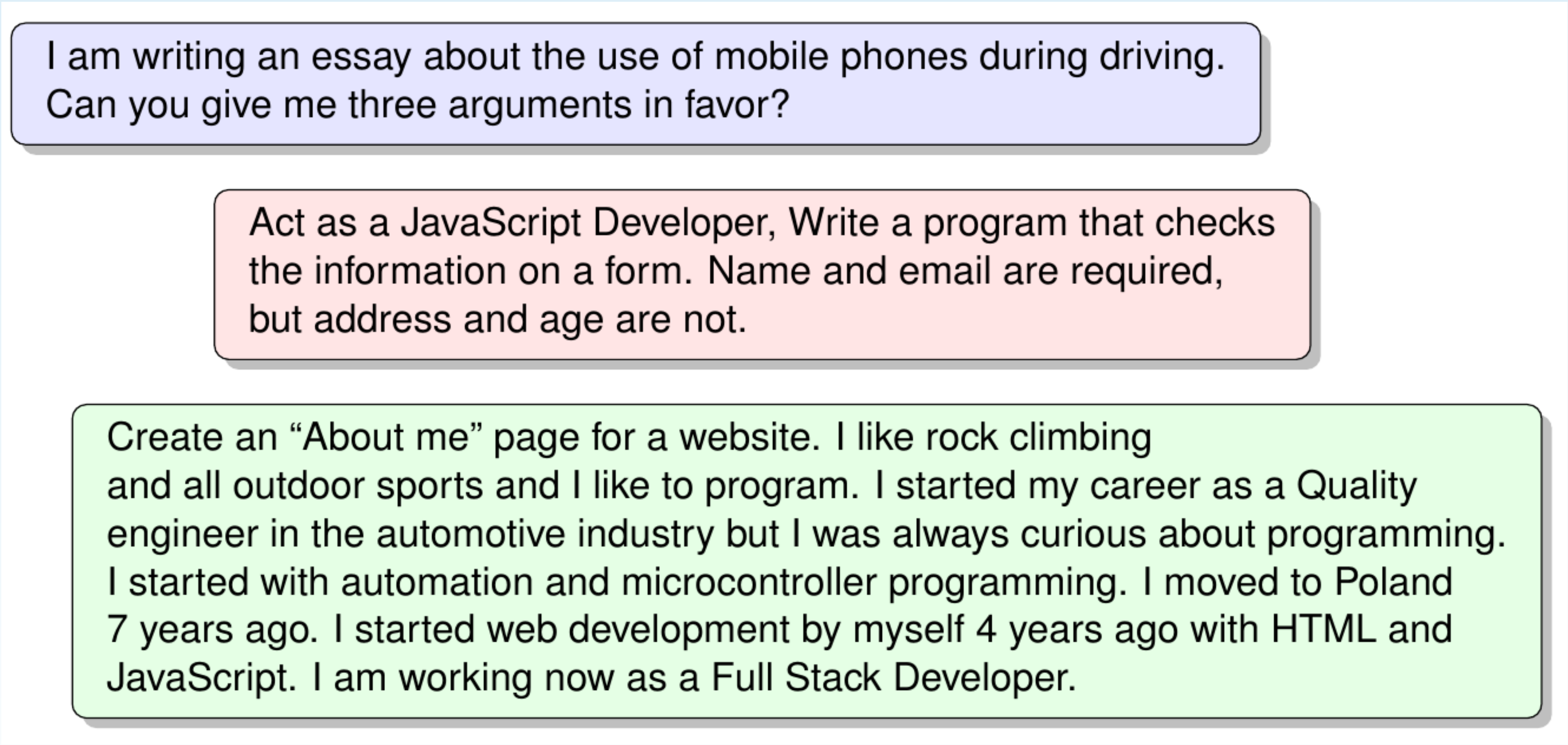
Ideia de IA generativa já existia...



Crédito imagem/slide: M. Lapata

O que mudou recentemente?

R: Escala. Detalhes e nuances da geração são imperceptíveis ao olhar comum.



Crédito imagem/slide: M. Lapata

Contexto

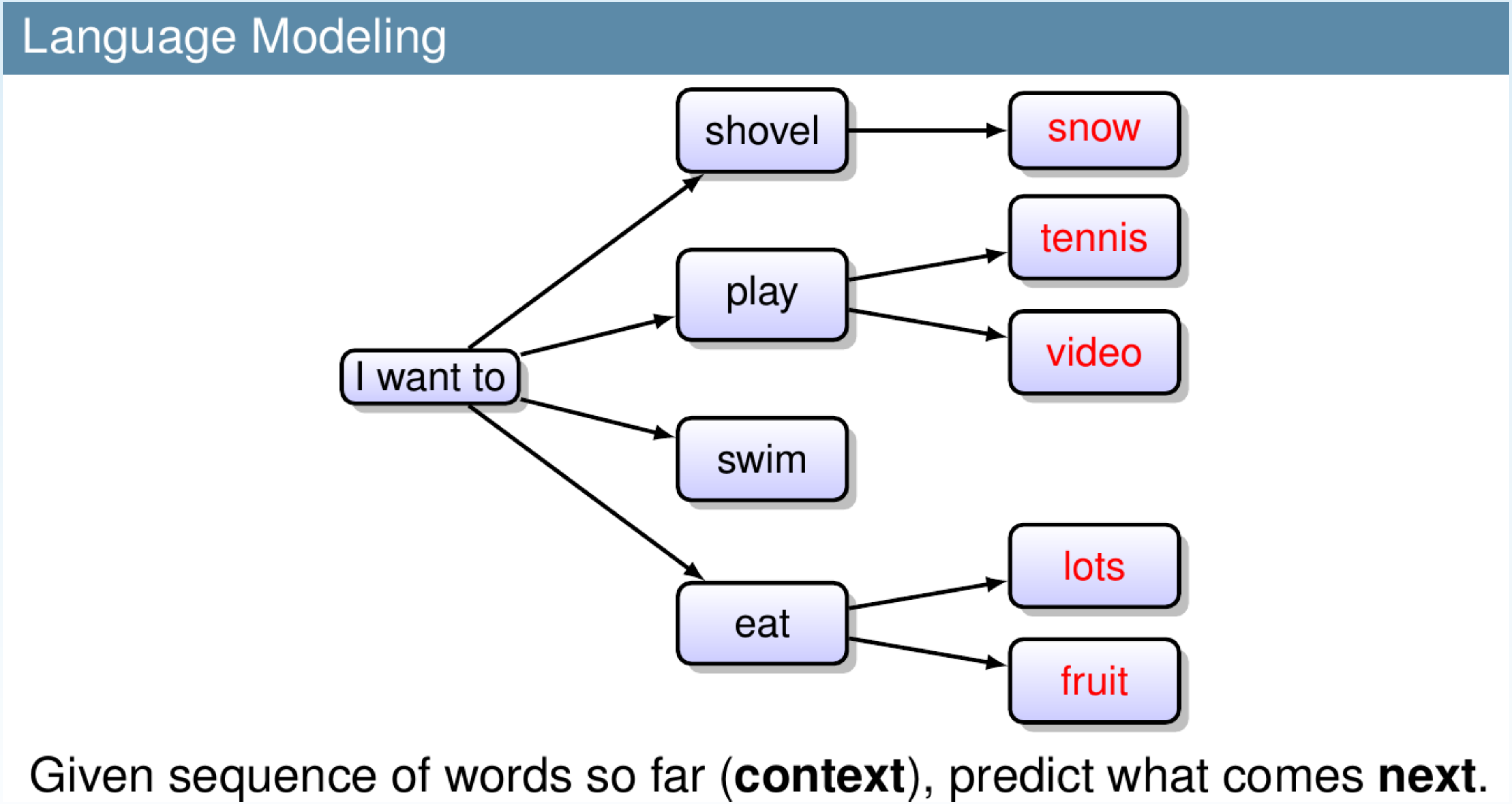
Crédito imagem/slide: A. Soleimany

Impacto...

Time to Reach 100M Users
Months to get to 100 million global Monthly Active Users
78
70
61
55
41
30
9
2
Google Translate
Uber
Telegram
Spotify
Pinterest
Instagram
TikTok
ChatGPT
Source: UBS / Yahoo Finance
@EconomyApp
APP ECONOMY INSIGHTS

IA Generativa

A partir de modelagem de linguagem natural...



Crédito imagem/slide: M. Lapata

IA Generativa

A partir de modelagem de linguagem natural...

A language model assigns probabilities to sequences of words, $\boldsymbol{w} = (w_1, w_2, \ldots, w_l)$.

It is convenient to decompose this probability using the **chain rule**:

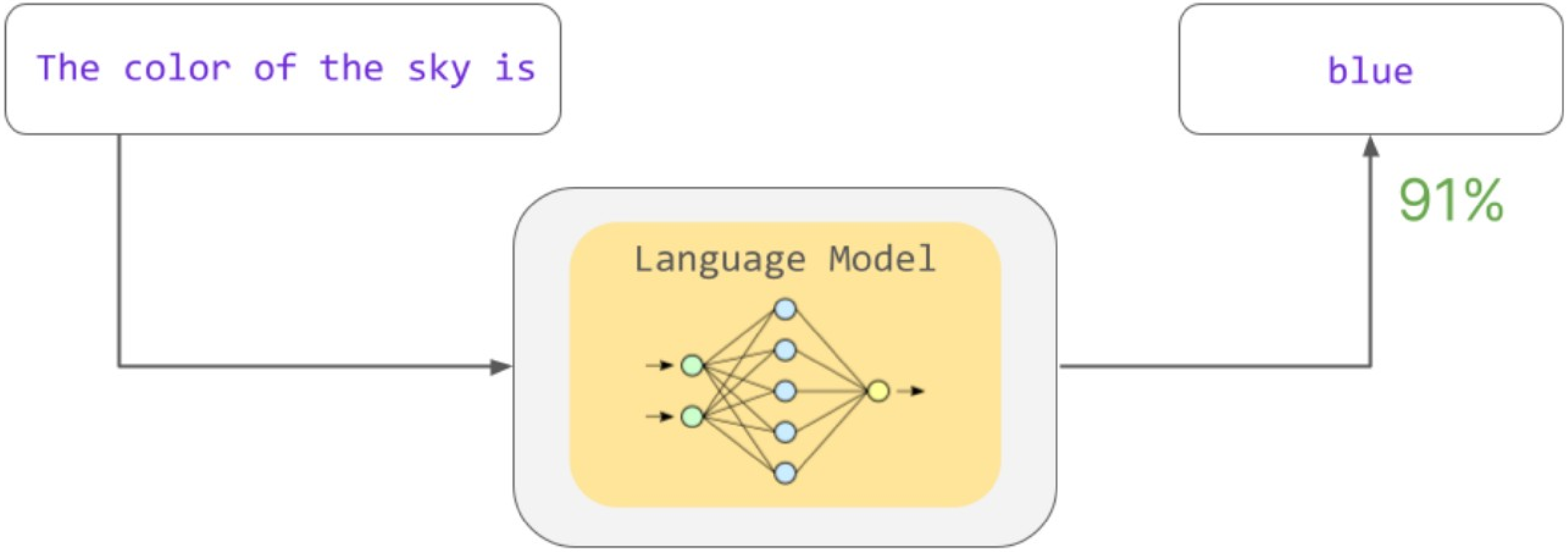
$$\begin{aligned} p(w) &= p(w_1) \times p(w_2|w_1) \times p(w_3|w_1, w_2) \times p(w_l|w_1, \ldots, w_{l-1}) \\ &= \prod_{t=1}^{|\boldsymbol{w}|} p(w_t|w_1, \ldots, w_{t-1}) \end{aligned}$$

This reduces the language modeling problem to **modeling the probability of the next word**, given the history of preceding words.

Crédito imagem/slide: M. Lapata

IA Generativa

A partir de modelagem de linguagem natural...



Given sequence of words so far (**context**), predict what comes **next**.

Crédito imagem/slide: M. Lapata

IA Generativa

Recipe for Creating your Own Language Model

Step 1: Collect a very large corpus:

- Wikipedia Books, StackOverflow
- Quora, Public Social media,
- Github, Reddit

Step 2: Ask LM to predict the next word in a sentence:

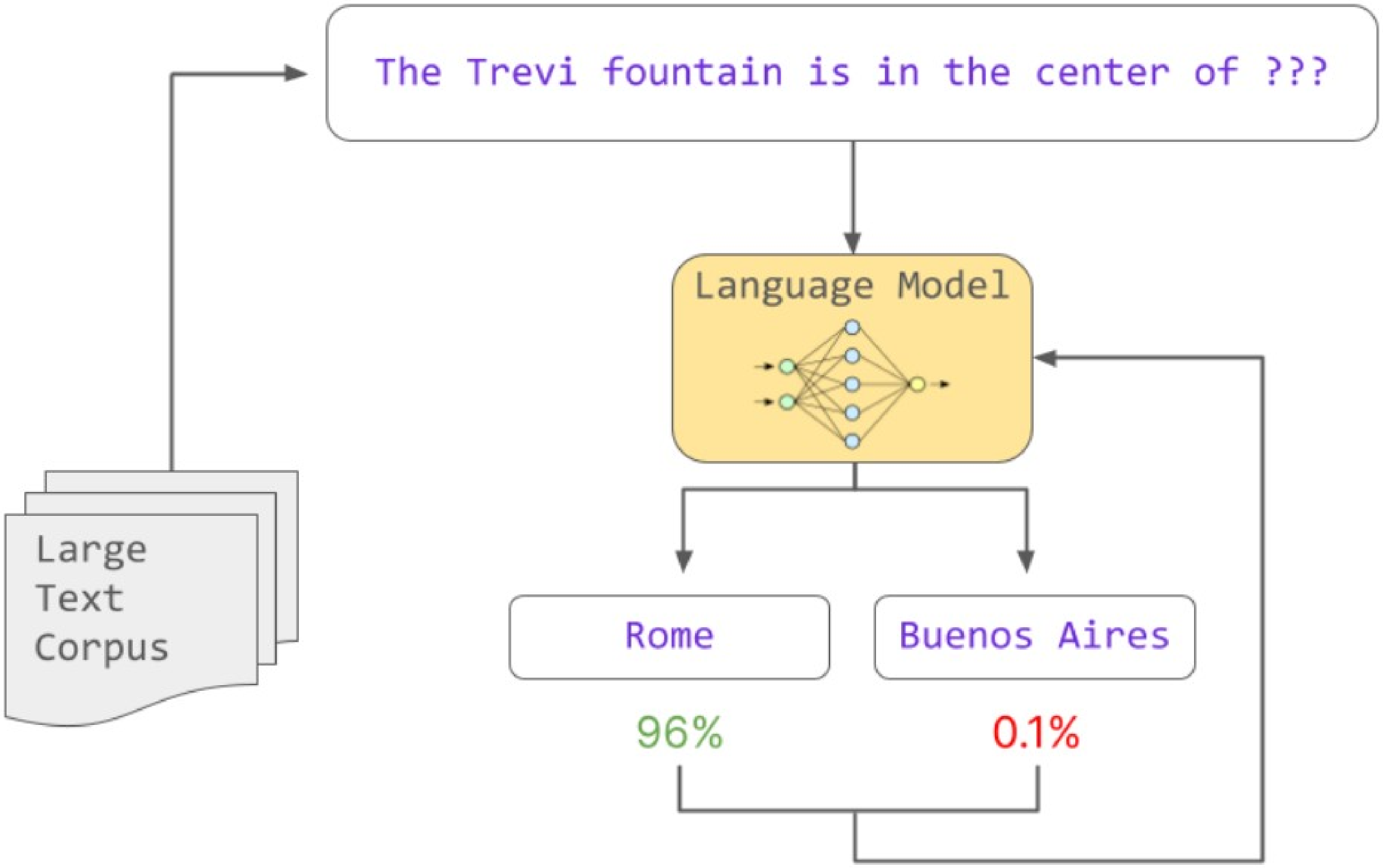
- randomly truncate parts of input sentence
- calculate probabilities of missing words
- adjust and feed back to the model to match the ground truth

Step 3: Repeat over the whole corpus.

Crédito imagem/slide: M. Lapata

IA Generativa

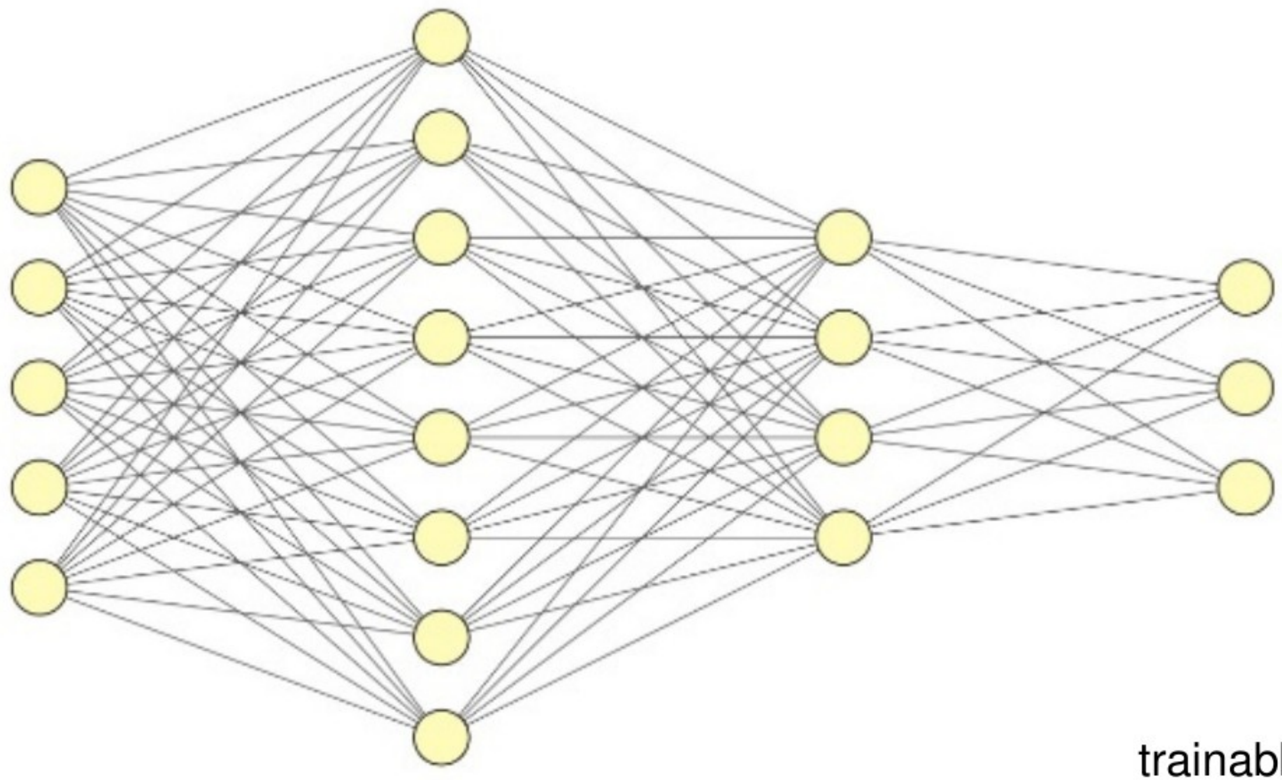
Self-supervised Learning



Crédito imagem/slide: M. Lapata

IA Generativa

Neural Network Language Models



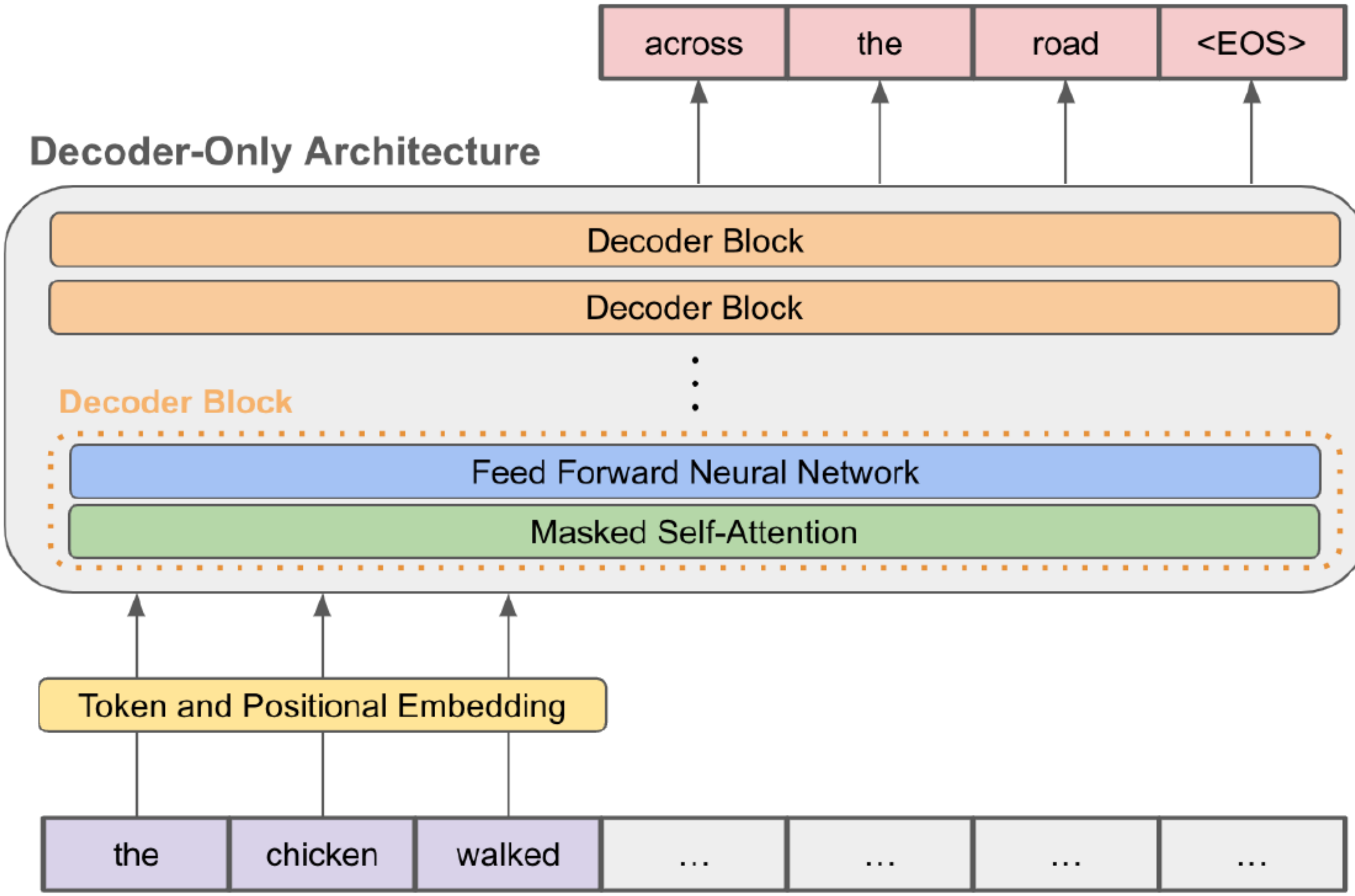
trainable parameters

$5 \times 8 + 8 \quad + \quad 8 \times 4 + 4 \quad + \quad 4 \times 3 + 3 \quad = \quad 99$

Crédito imagem/slide: M. Lapata

IA Generativa

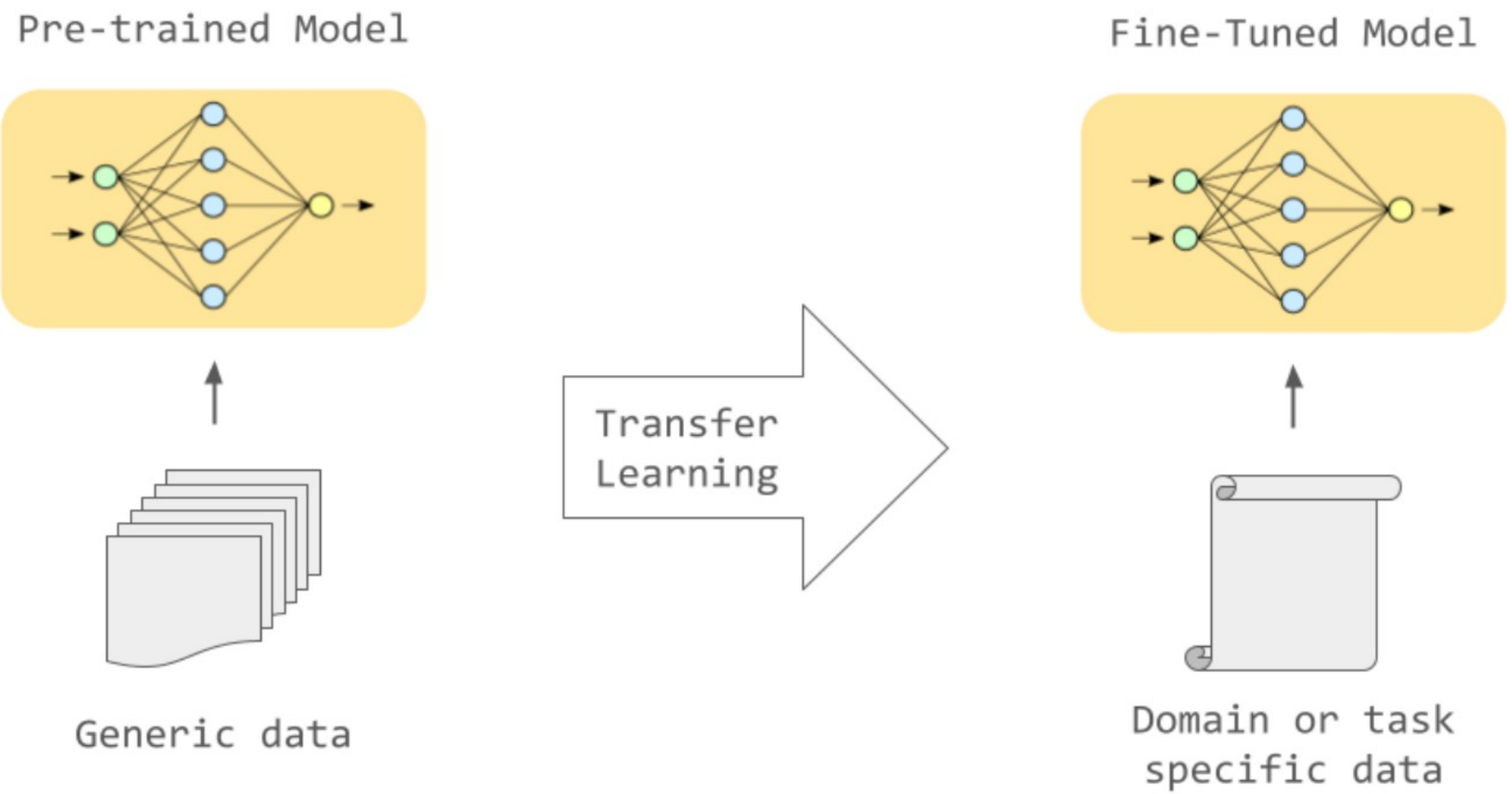
Transformers: The King of AI Architectures



Crédito imagem/slide: M. Lapata

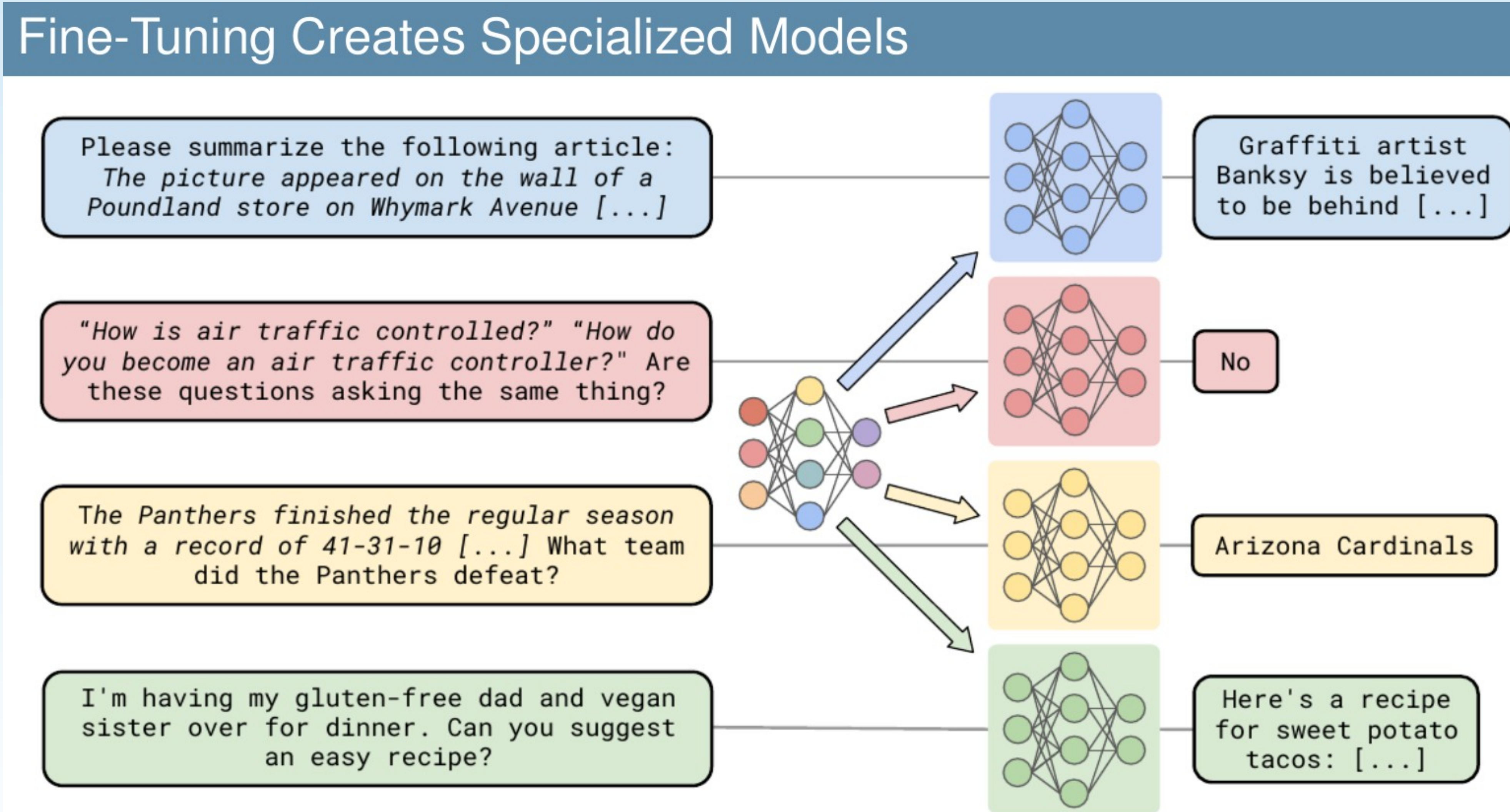
IA Generativa

Exploiting a Pre-trained Model: Fine-Tuning



Crédito imagem/slide: M. Lapata

IA Generativa



Crédito imagem/slide: M. Lapata

IA Generativa

From Language Models to Large Language Models

How good can a language model become?

Crédito imagem/slide: M. Lapata

IA Generativa

From Language Models to Large Language Models

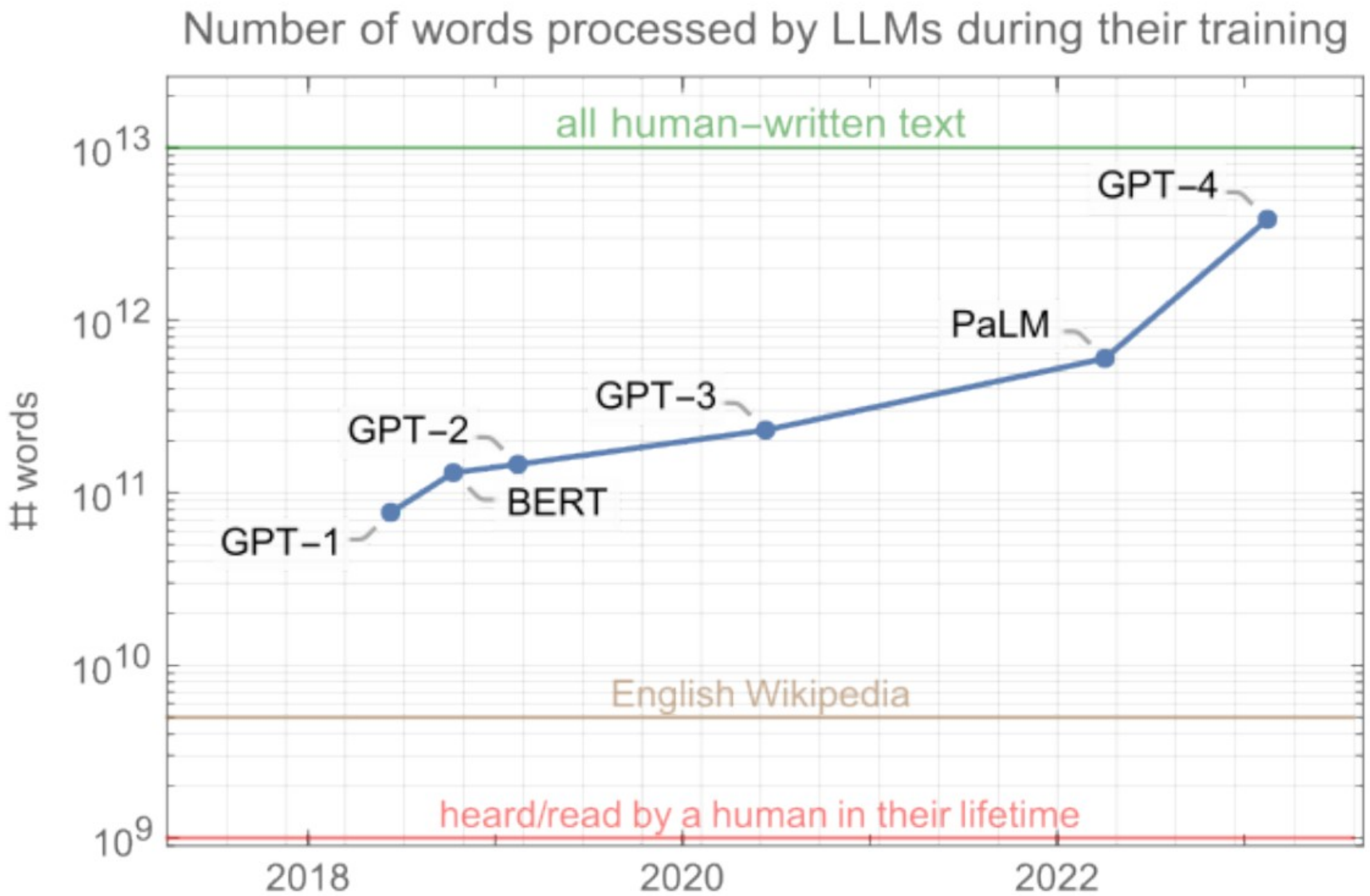
How good can a language model become?

The bigger the better!

Crédito imagem/slide: M. Lapata

IA Generativa

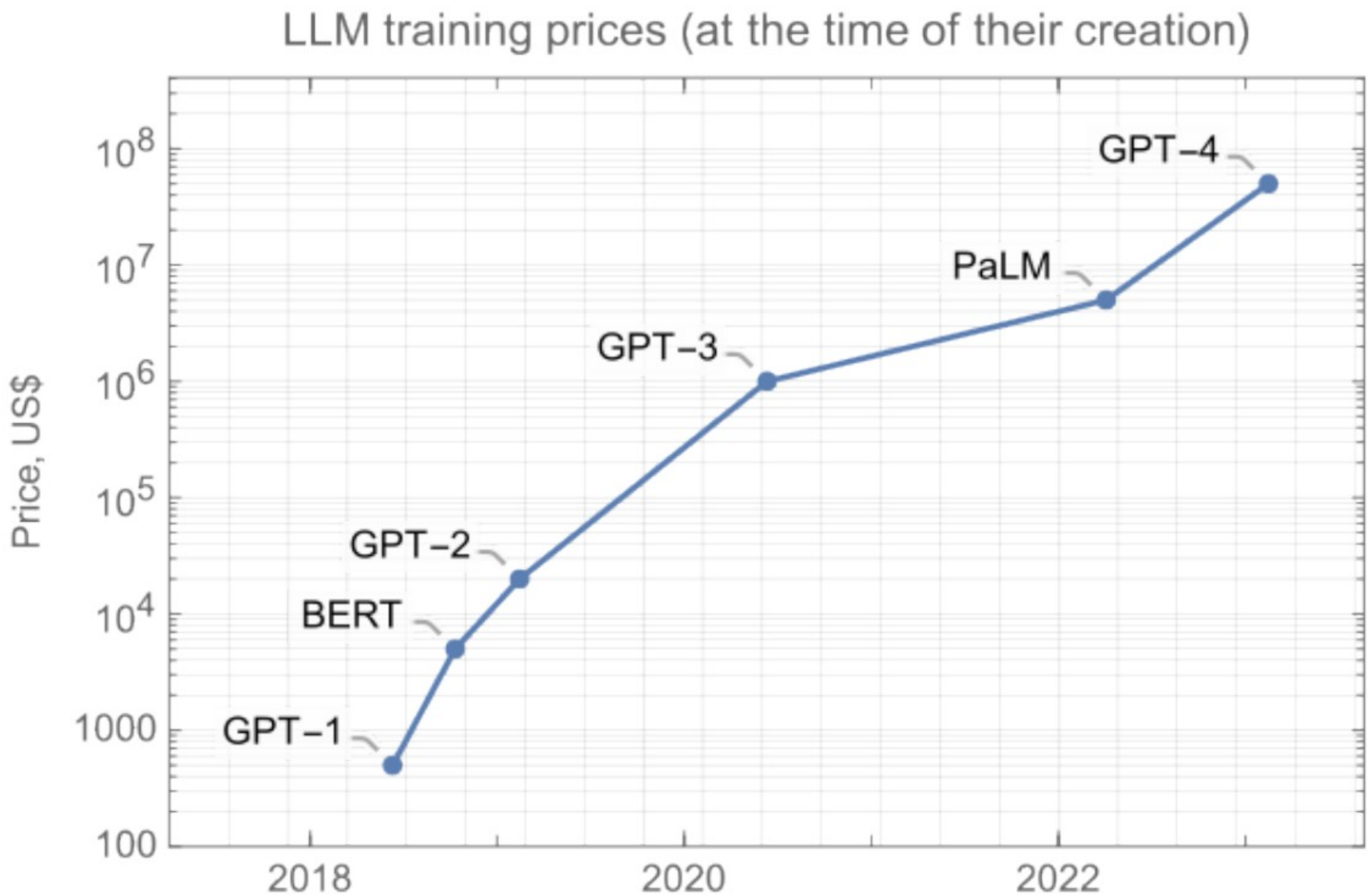
From Language Models to Large Language Models



Crédito imagem/slide: M. Lapata

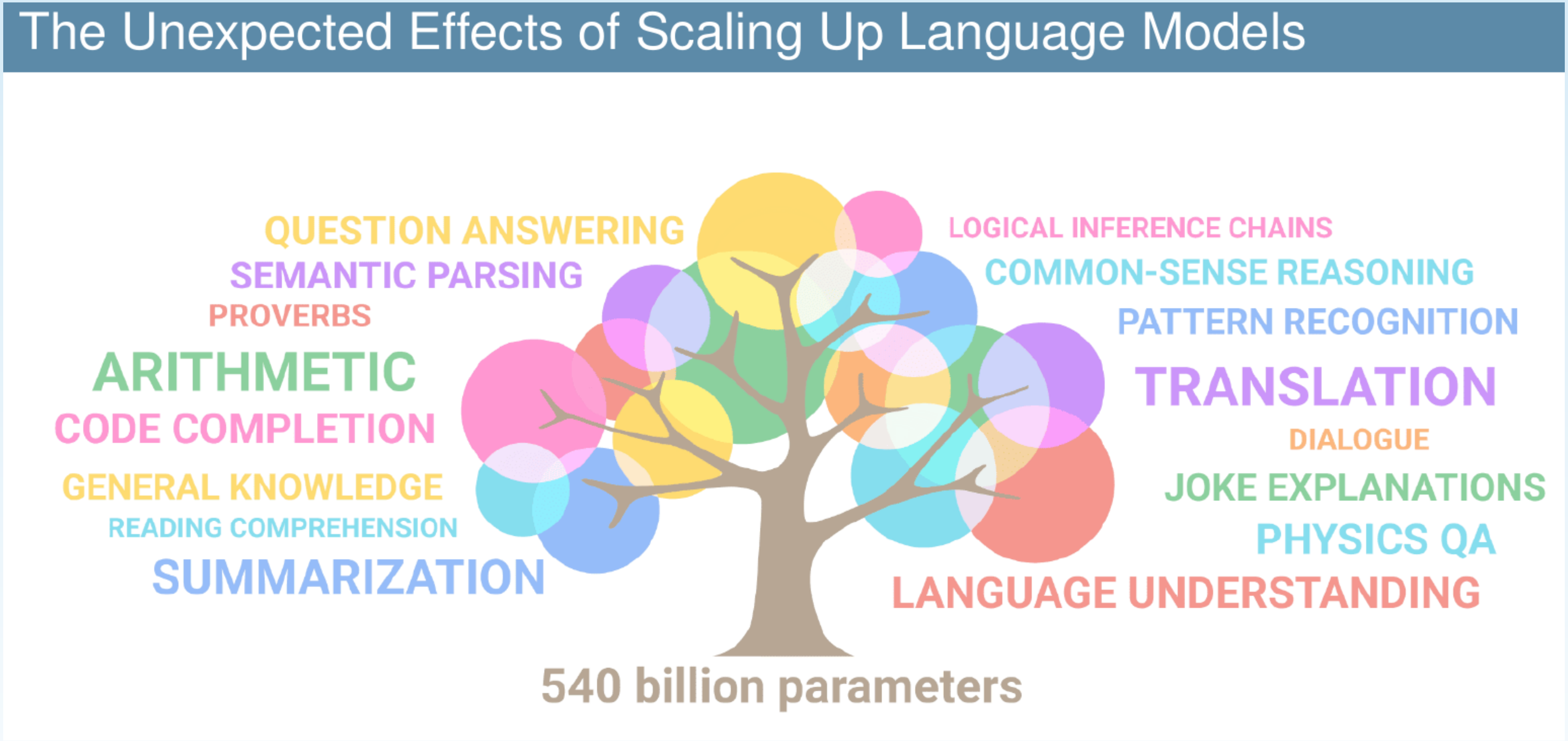
IA Generativa

From Language Models to Large Language Models



Crédito imagem/slide: M. Lapata

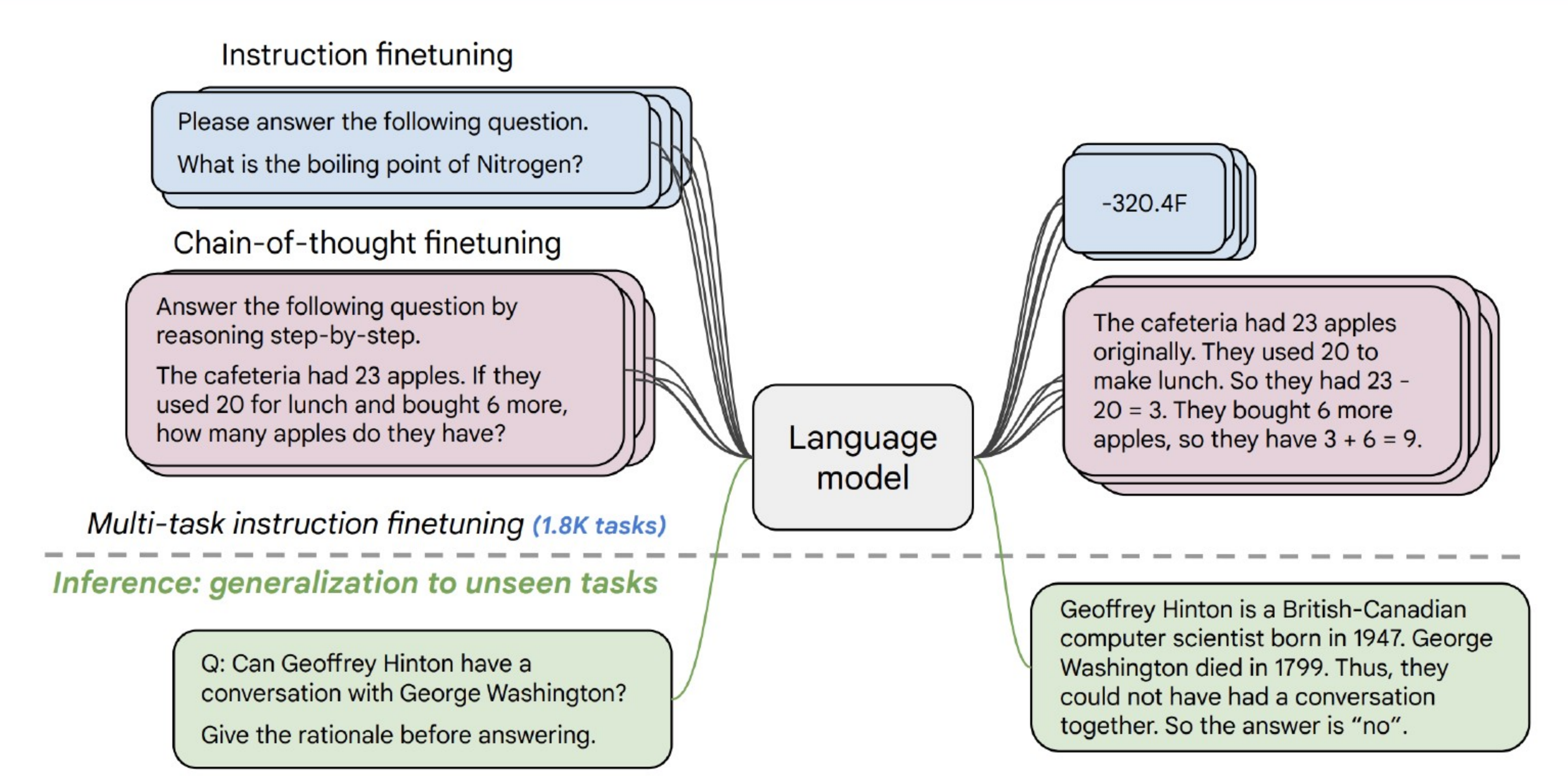
IA Generativa



Crédito imagem/slide: M. Lapata

IA Generativa

Prompting and Instruction Tuning



Crédito imagem/slide: M. Lapata

IA Generativa

Are Language Models Always Right or Fair?

- It is virtually impossible to regulate the content LLMs are exposed to during training.
- Because LLMs are trained on the web, they'll always encode historical biases and may reproduce harmful content.
- They generate hallucinations by fabricating nonexistent or false facts.
- LLMs may occasionally exhibit various types of undesirable behavior (we'll see examples).

Crédito imagem/slide: M. Lapata

IA Generativa

Impact on the Enviroment

A ChatGPT query takes 100x more energy to execute than a Google Search query

Llama 2 (a ChatGPT-like model from Meta) training produced 539 metric tons of $C0_2$

Larger models use more energy during their deployment!

Crédito imagem/slide: M. Lapata

IA Generativa

Impact on Society

A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it.

"It was super easy actually," he says, "which was the scary part."

An AI that writes convincing prose risks mass-producing fake news

AI-Generated Fake 'Drake'/'Weeknd' Collaboration, 'Heart on My Sleeve,' Delights Fans and Sets Off Industry Alarm Bells

Crédito imagem/slide: M. Lapata

IA Generativa

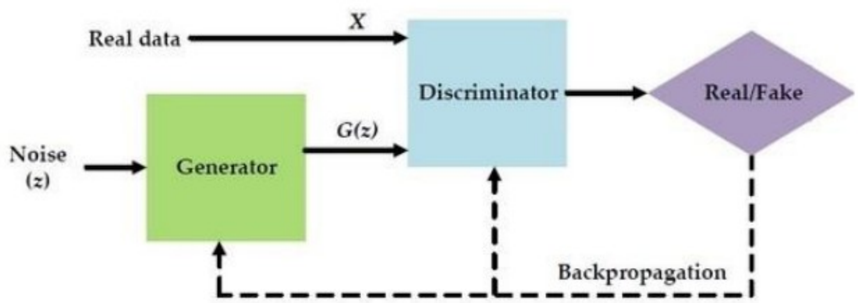
- What is a bigger threat to mankind, AI or climate change?
- Who is in control of AI and who benefits from it?
- Does the benefit outweigh the risk?
- All risky technology has historically been strongly regulated.

Crédito imagem/slide: M. Lapata

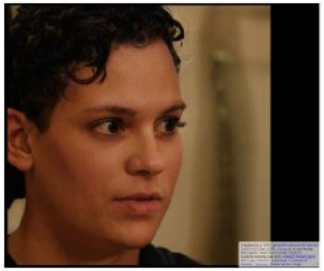
IA Generativa

Where it all comes from: GANs, LLMs & Diffusers

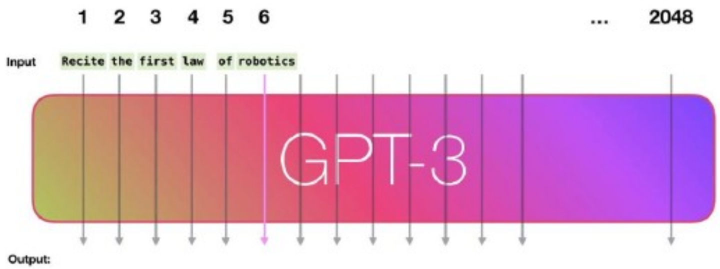
Generative Adversarial Networks (2014)
The Generator creates deep fakes to train the Discriminator.



This Person Does Not Exist



Large language models perform sequence-to-sequence prediction and generate the next word in a sentence. (Transformer networks, since 2018)



Diffusers (since 2020) – Example from DALLE2 (Open AI, 2022)
"*Teddy bears mixing sparkling chemicals as mad scientists in a steampunk style*" – create images from text and by de-noising.

Aprendizagem não supervisionada

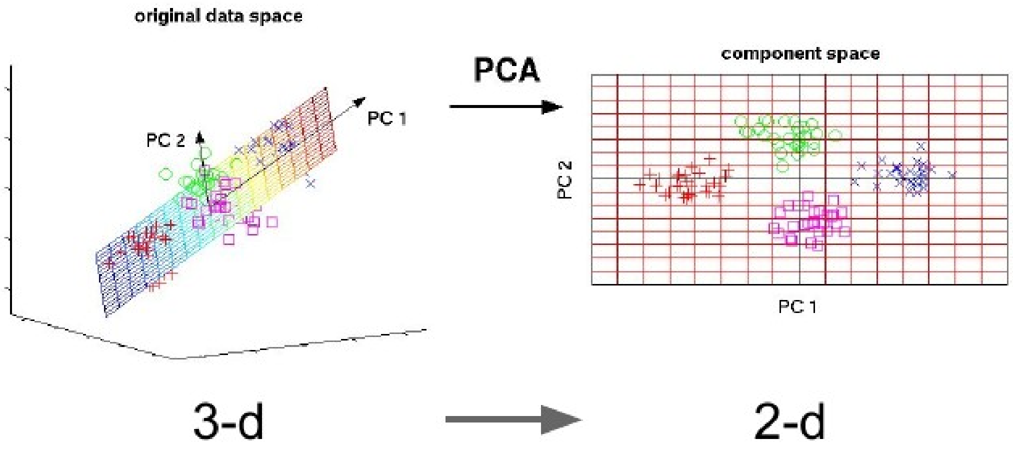
Supervised vs Unsupervised Learning

Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, density estimation, etc.



Principal Component Analysis
(Dimensionality reduction)

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Aprendizagem não supervisionada

Supervised vs Unsupervised Learning

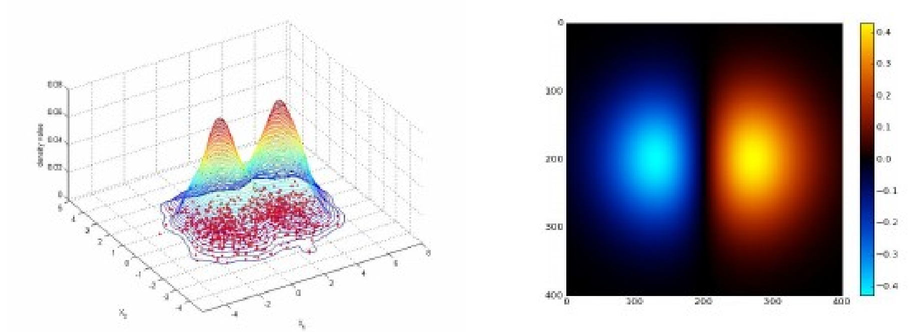
Unsupervised Learning

Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, density estimation, etc.

1-d density estimation



2-d density estimation

Modeling p(x)

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Aprendizagem não supervisionada

Supervised vs Unsupervised Learning

Supervised Learning

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

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

Unsupervised Learning

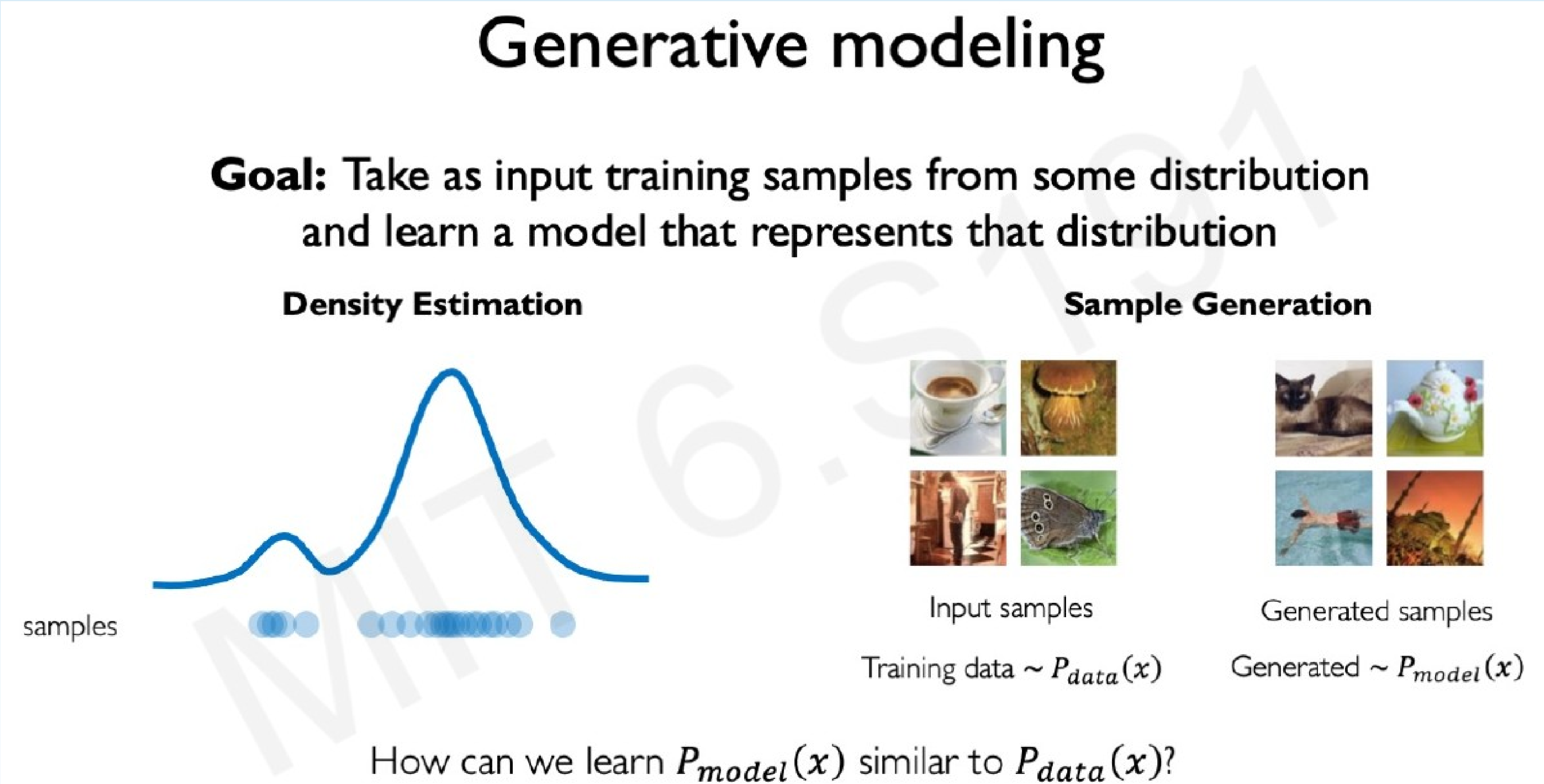
Data: x
Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, density estimation, etc.

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Modelagem Generativa

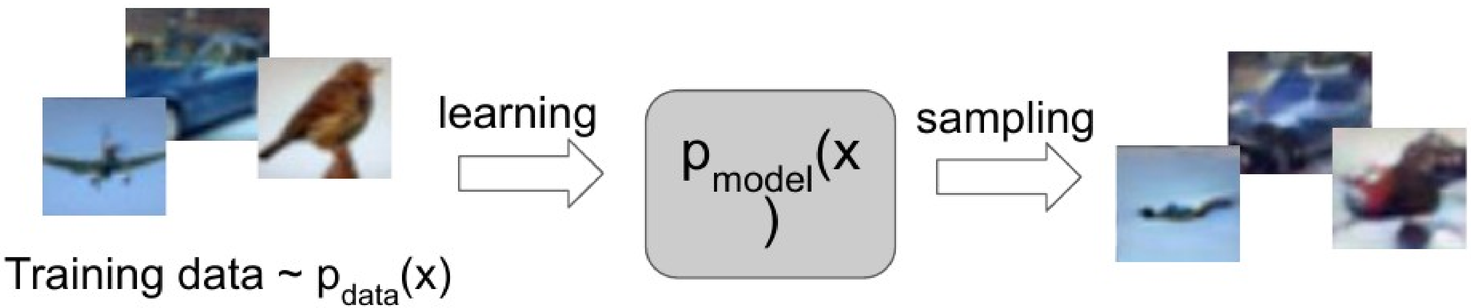


Crédito imagem/slide: A. Soleimany

Modelagem Generativa

Generative Modeling

Given training data, generate new samples from same distribution



Formulate as density estimation problems:

- **Explicit density estimation**: explicitly define and solve for $p_{model}(x)$
- **Implicit density estimation**: learn model that can sample from $p_{model}(x)$ **without explicitly defining it.**

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Modelagem Generativa (Por que?)

Why generative models? Debiasing

Capable of uncovering **underlying features** in a dataset

Homogeneous skin color, pose

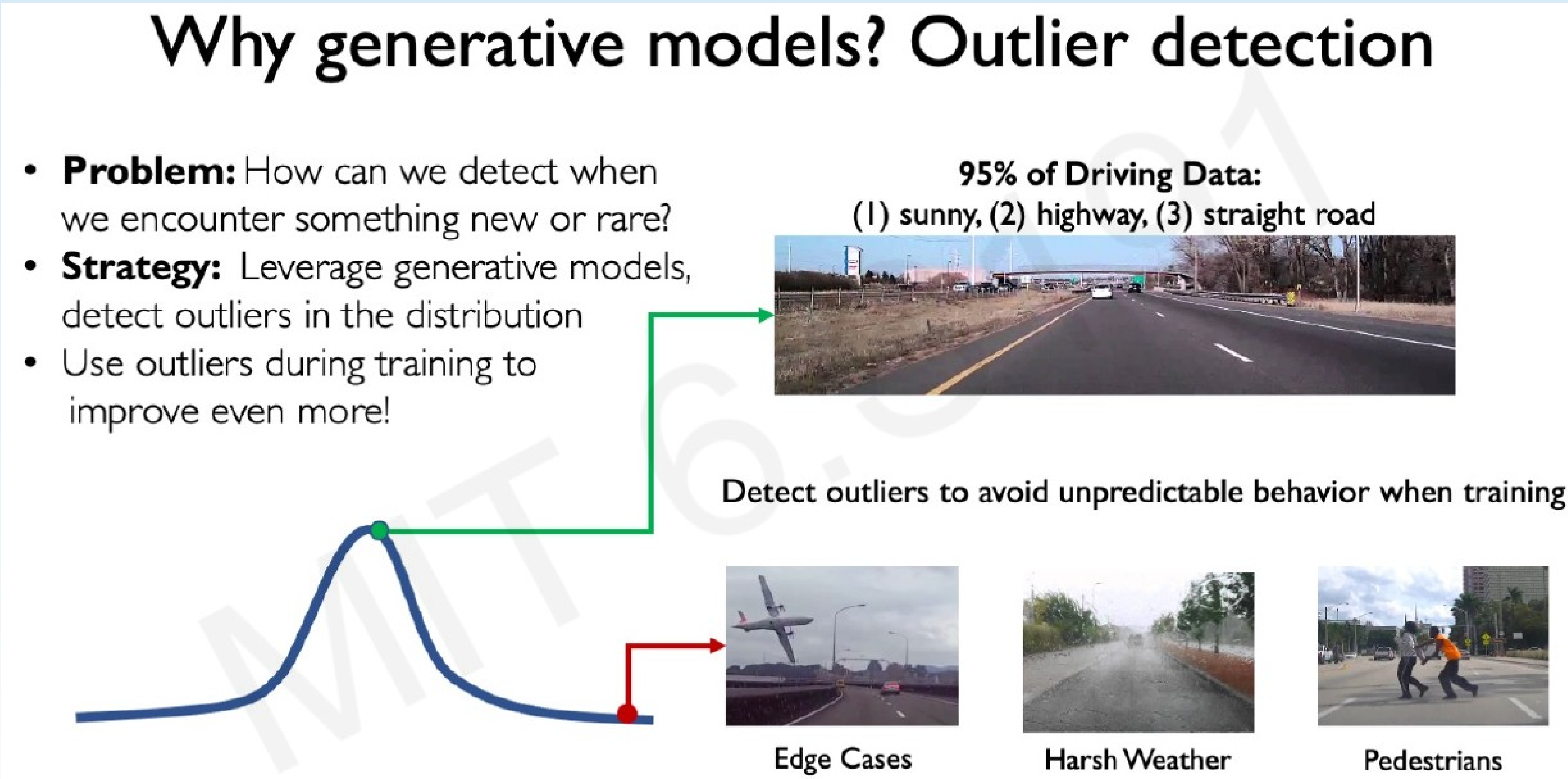
vs

Diverse skin color, pose, illumination

How can we use this information to create fair and representative datasets?

Crédito imagem/slide: A. Soleimany

Modelagem Generativa (Por que?)



Crédito imagem/slide: A. Soleimany

Modelagem Generativa (métodos)

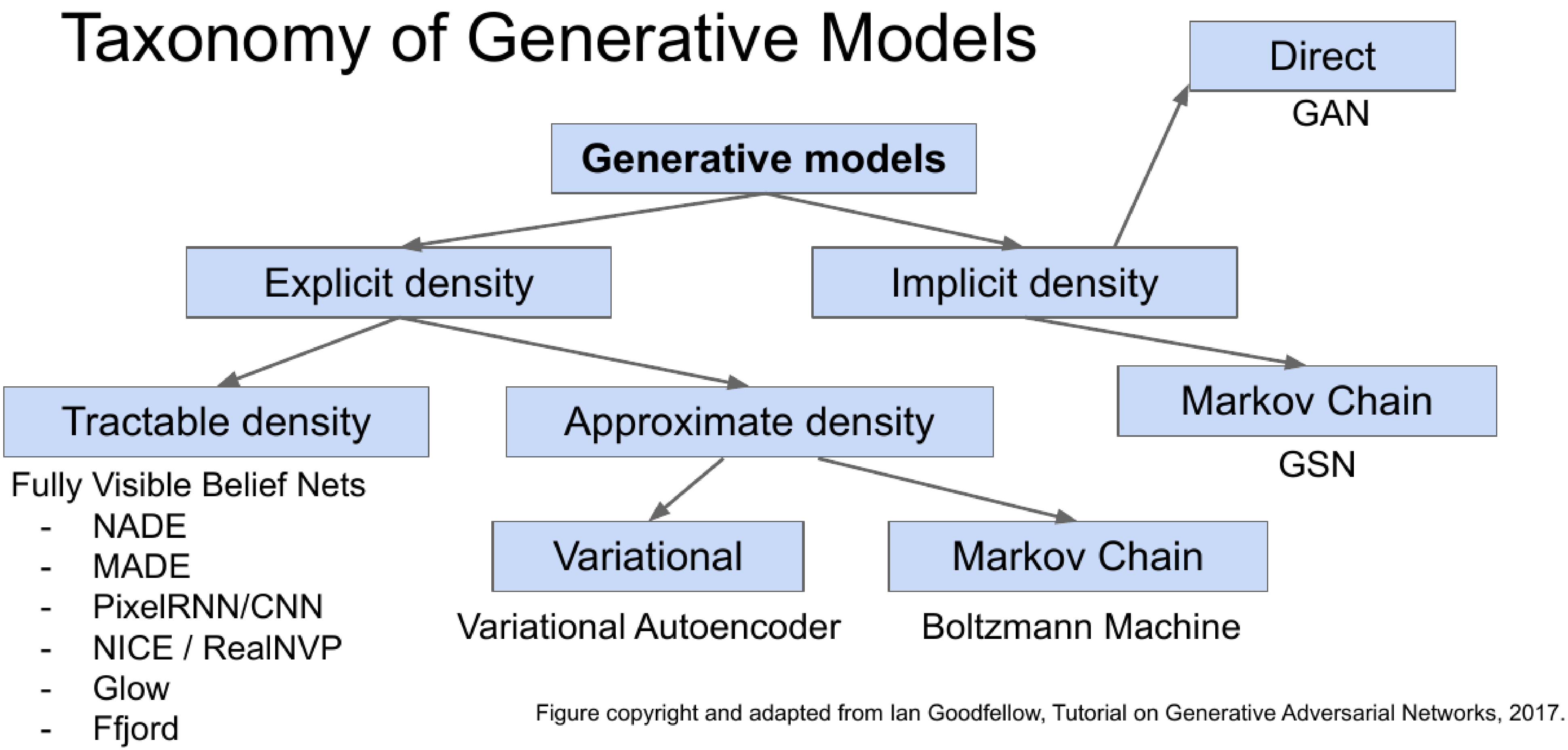
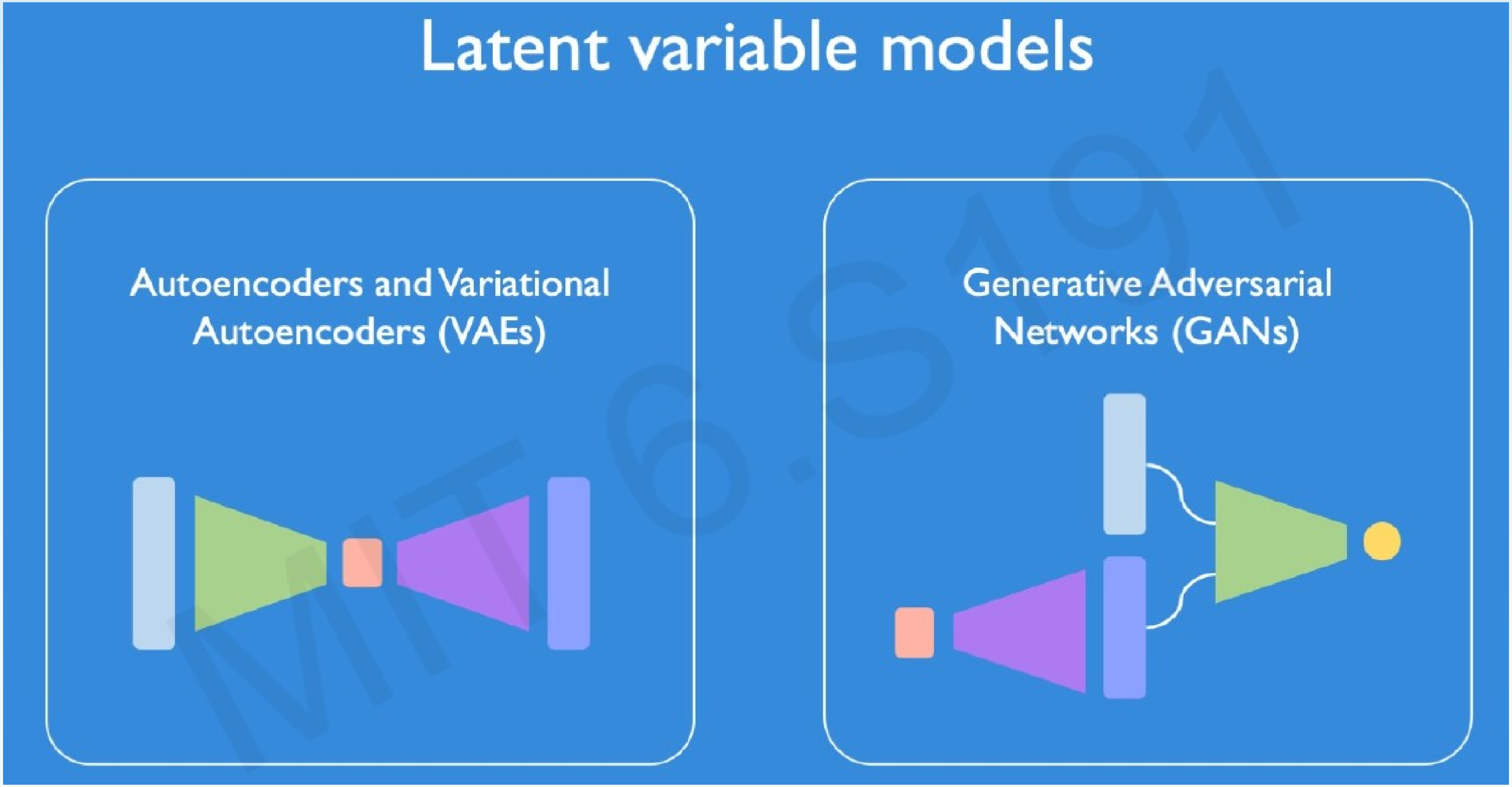


Figure copyright and adapted from Ian Goodfellow, Tutorial on Generative Adversarial Networks, 2017.

Modelagem Gerativa (métodos)



Crédito imagem/slide: A. Soleimany

Variáveis latentes

What is a latent variable?

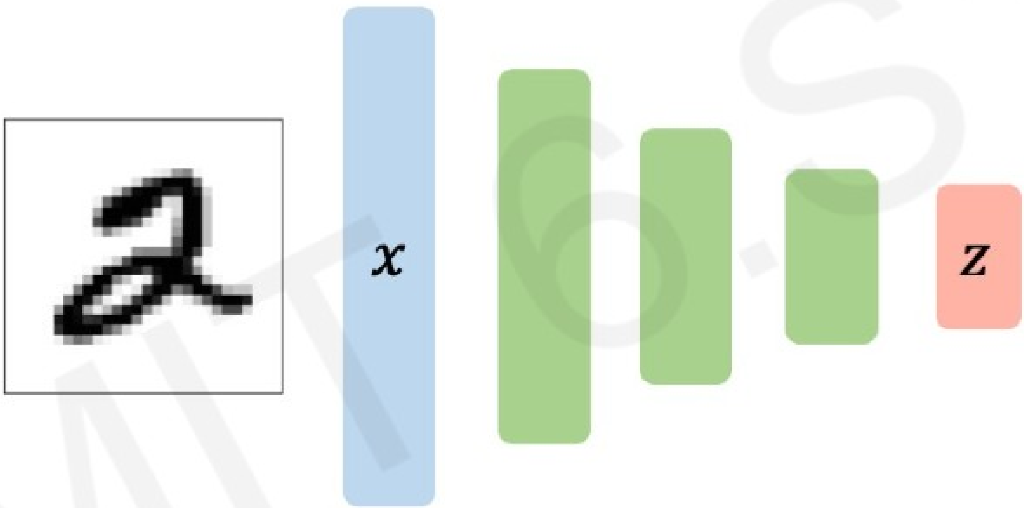
Can we learn the **true explanatory factors**, e.g. latent variables, from only observed data?

Crédito imagem/slide: A. Soleimany

Autocodificadores

Autoencoders: background

Unsupervised approach for learning a **lower-dimensional** feature representation from unlabeled training data

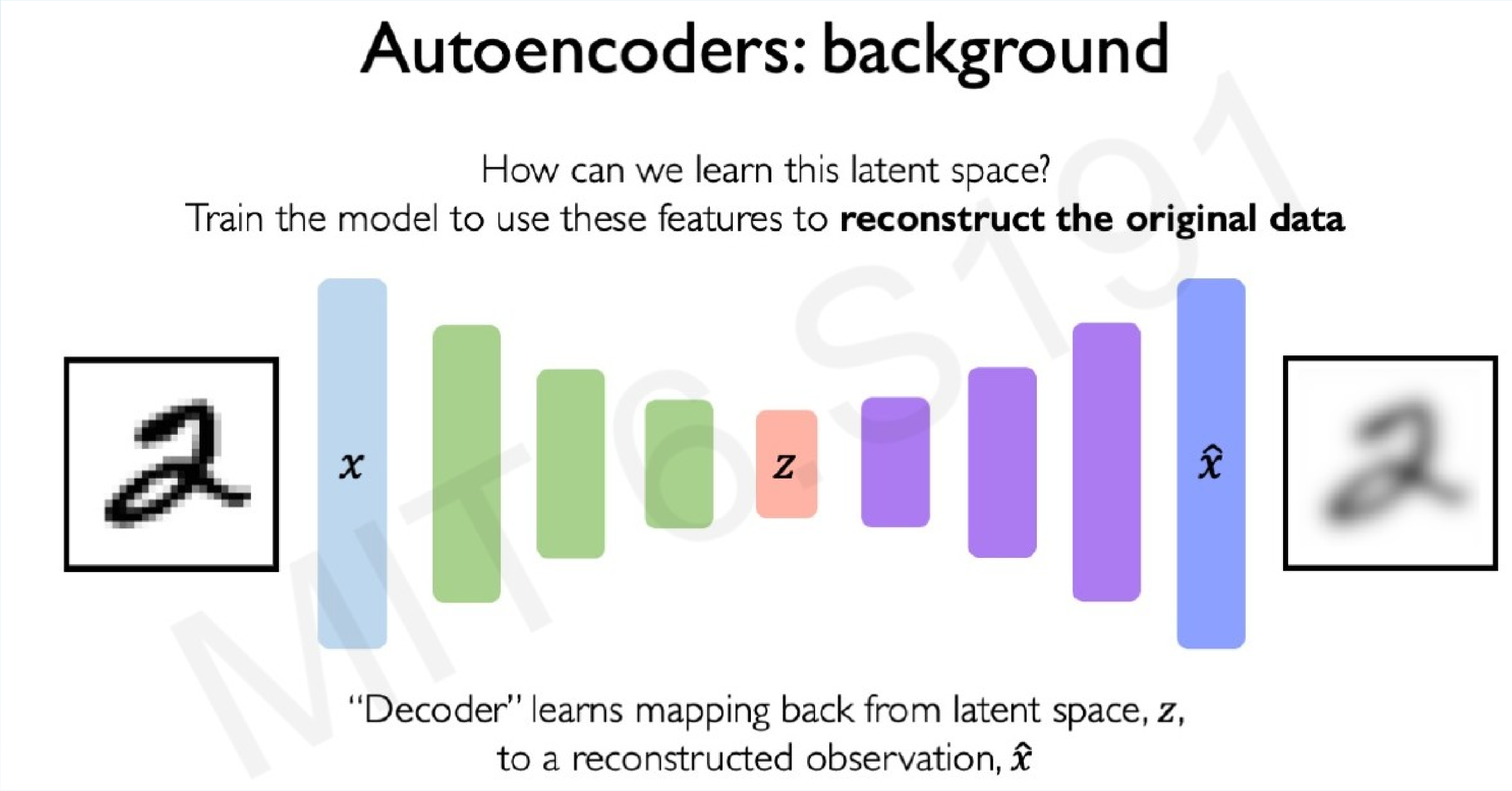


Why do we care about a low-dimensional z?

"Encoder" learns mapping from the data, x, to a low-dimensional latent space, z

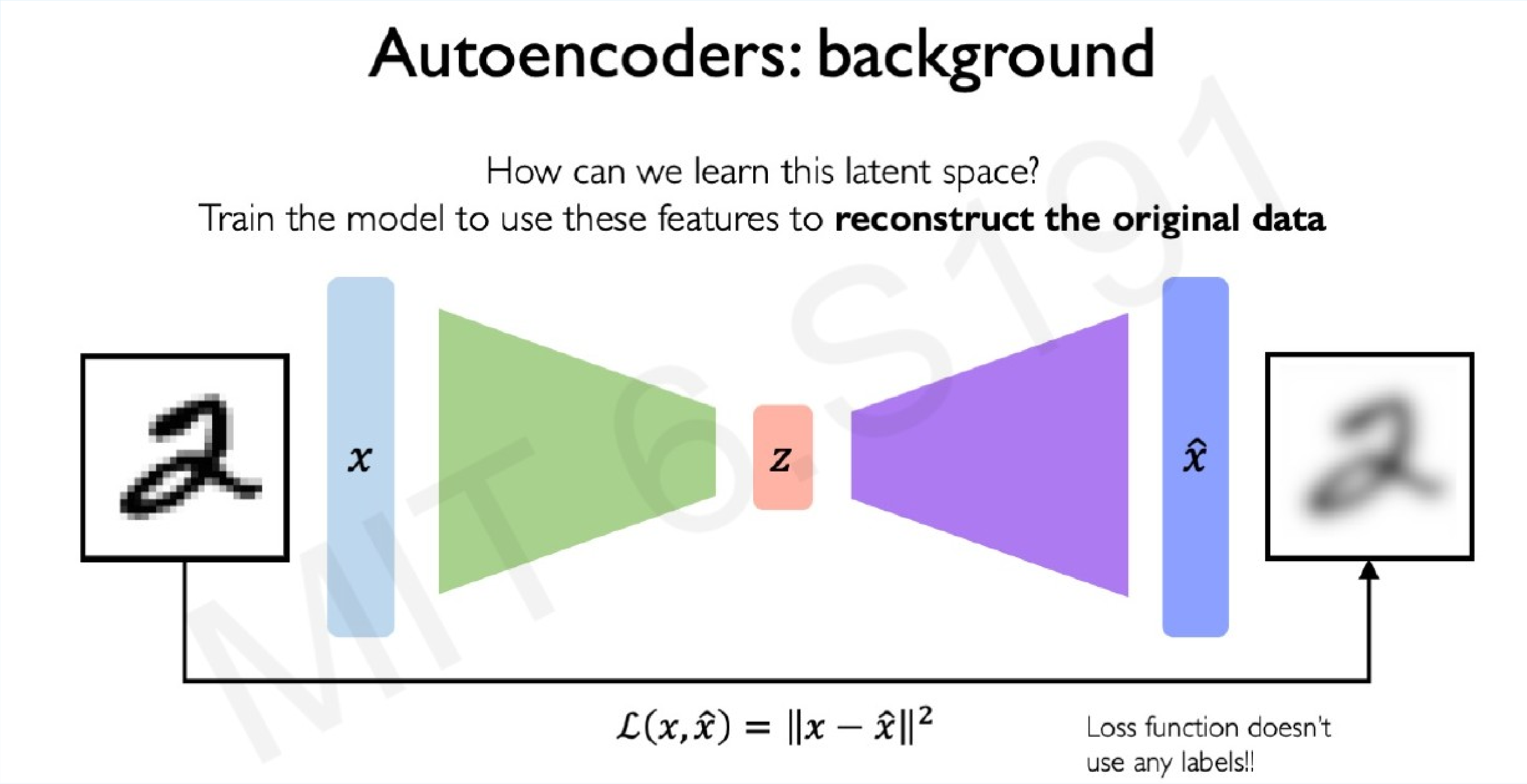
Crédito imagem/slide: A. Soleimany

Autocodificadores



Crédito imagem/slide: A. Soleimany

Autocodificadores

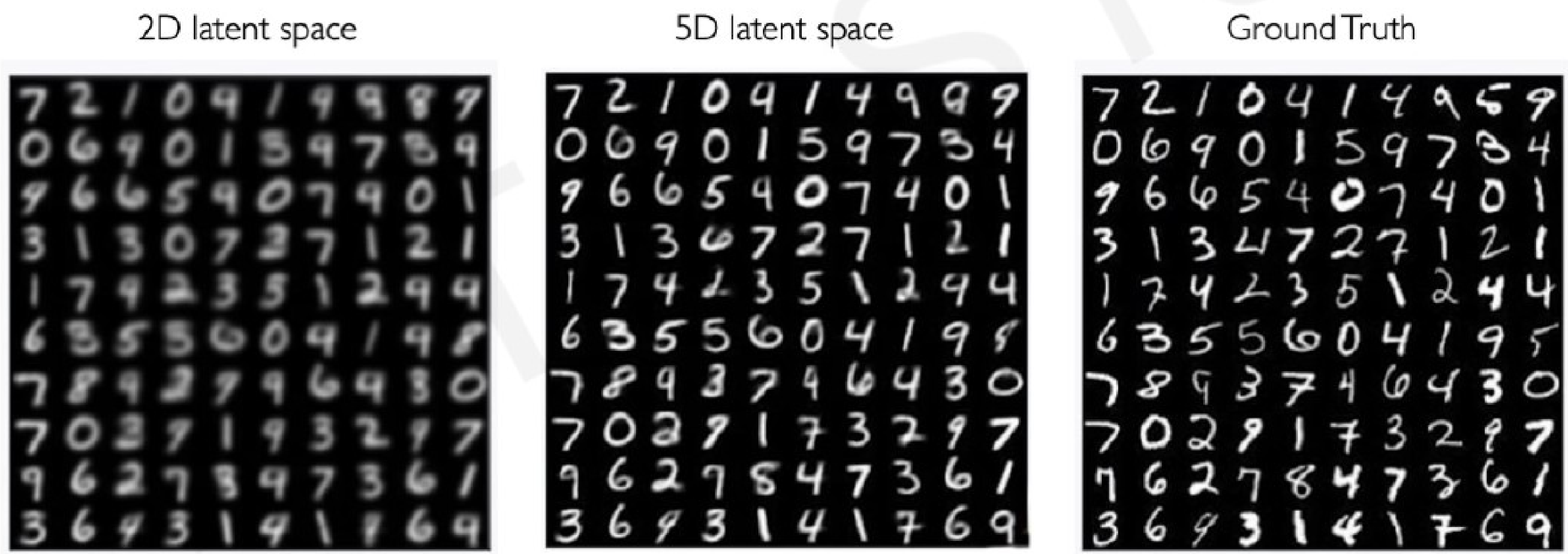


Crédito imagem/slide: A. Soleimany

Autocodificadores

Dimensionality of latent space → reconstruction quality

Autoencoding is a form of compression!
Smaller latent space will force a larger training bottleneck



Crédito imagem/slide: A. Soleimany

Autocodificadores

Autoencoders for representation learning

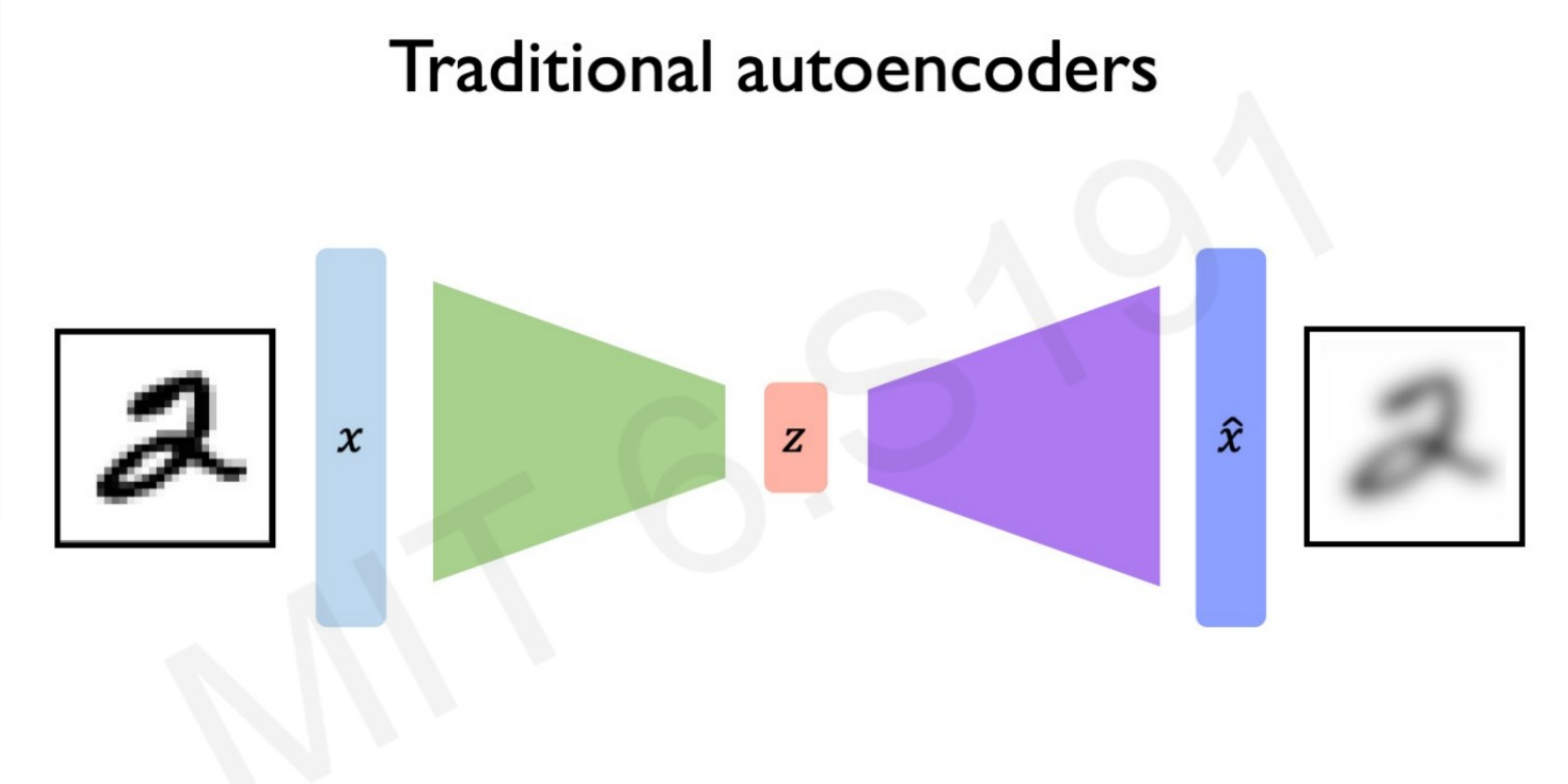
Bottleneck hidden layer forces network to learn a compressed latent representation

Reconstruction loss forces the latent representation to capture (or encode) as much "information" about the data as possible

Autoencoding = **Auto**matically **encoding** data

Crédito imagem/slide: A. Soleimany

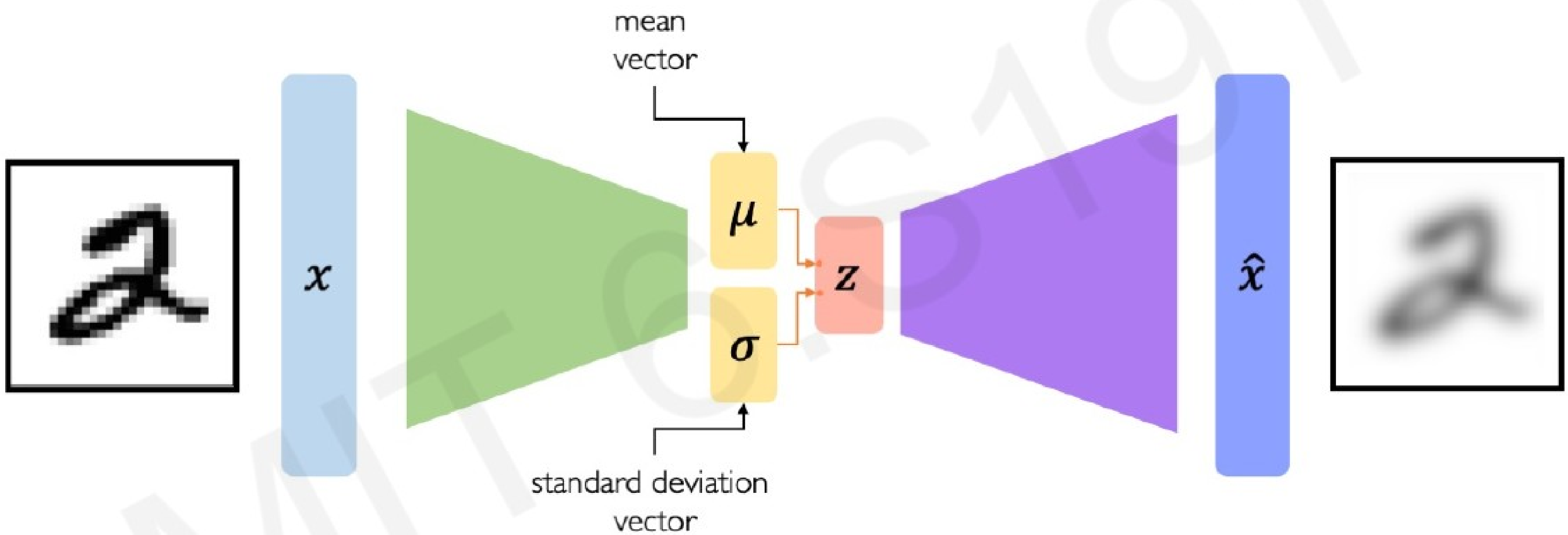
Autocodificadores



Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

VAEs: key difference with traditional autoencoder

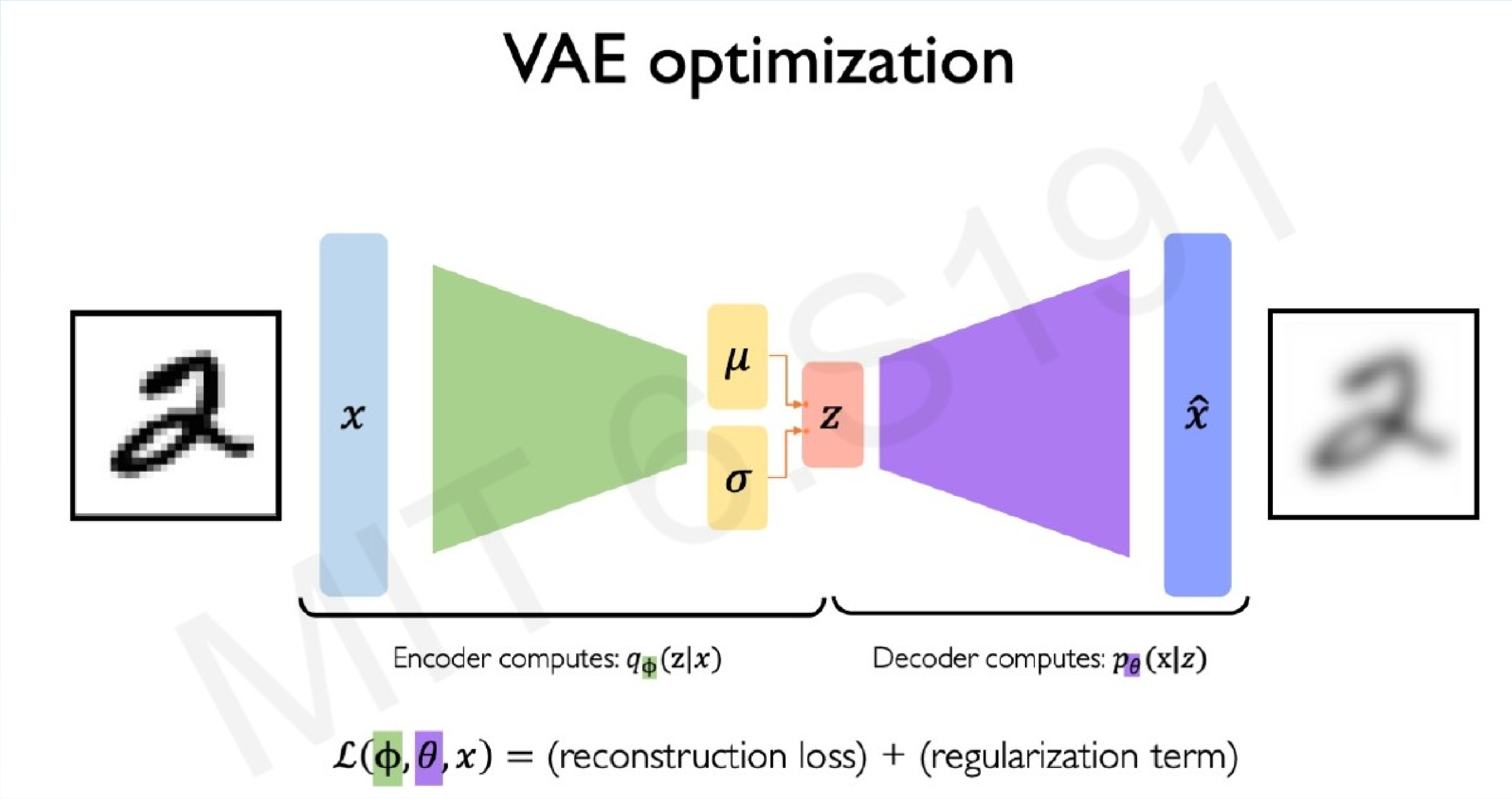


Variational autoencoders are a probabilistic twist on autoencoders!

Sample from the mean and standard deviation to compute latent sample

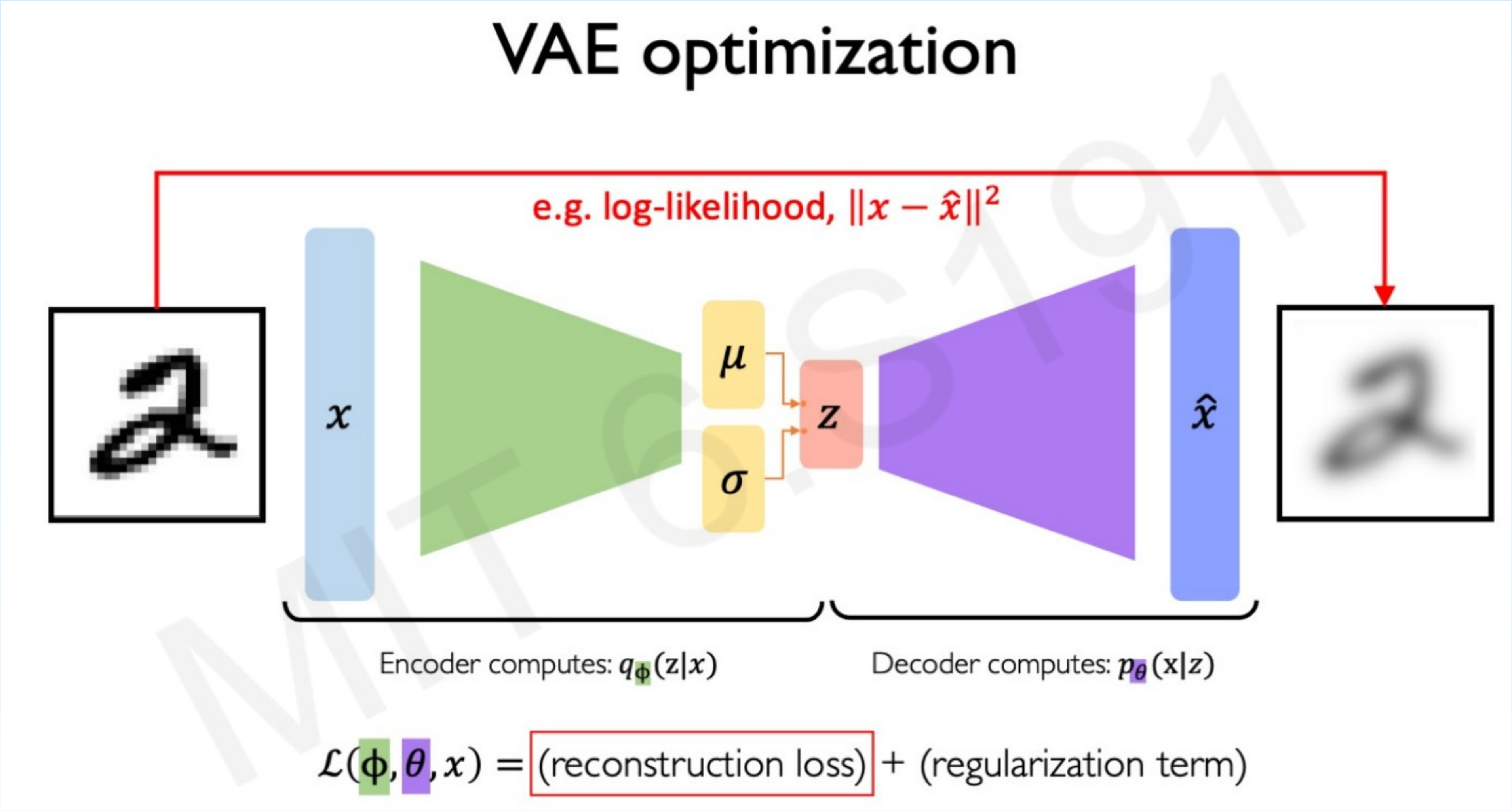
Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais



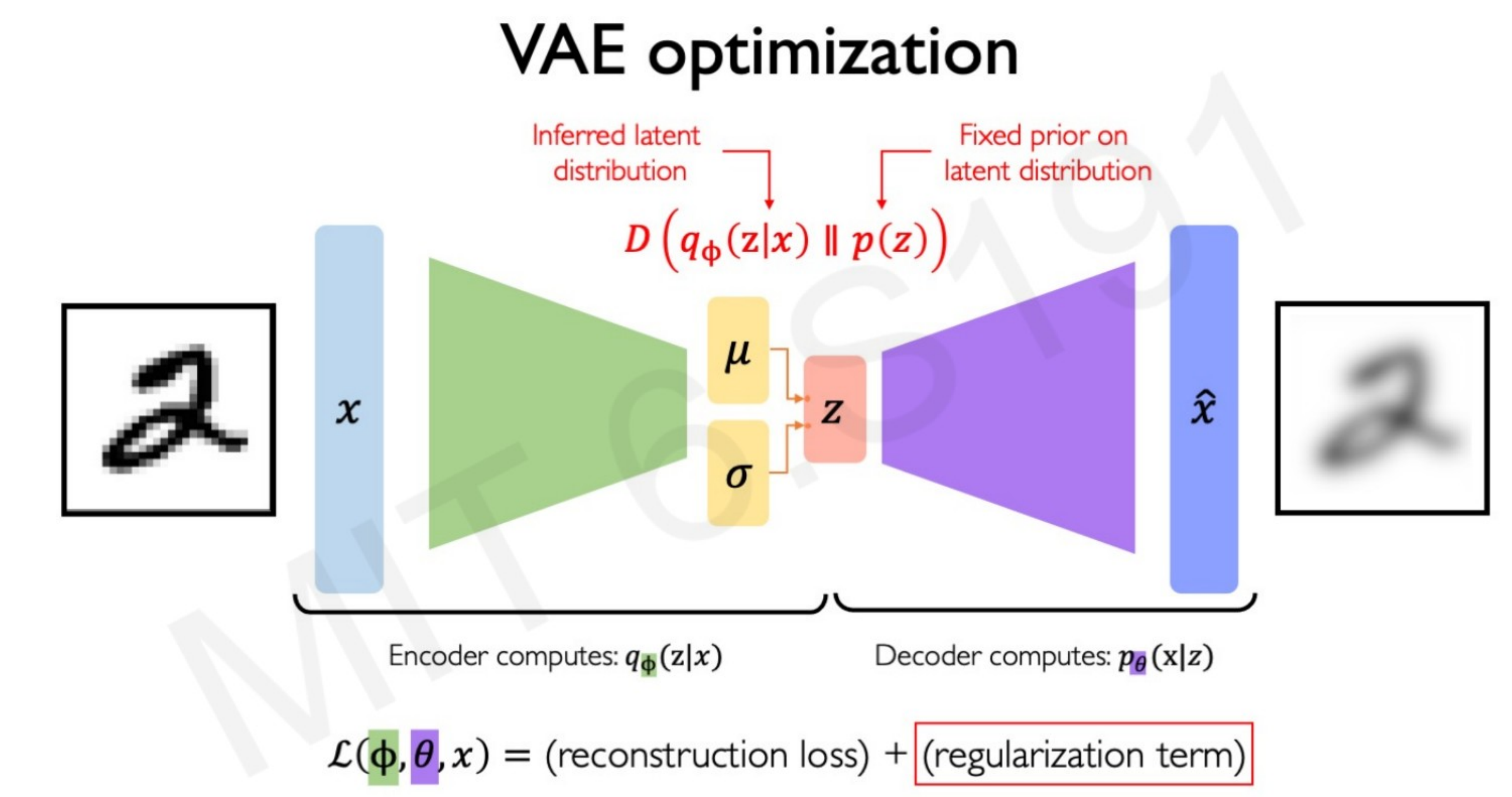
Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais



Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais



Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

VAEs: Latent perturbation

Slowly increase or decrease a **single latent variable**
Keep all other variables fixed

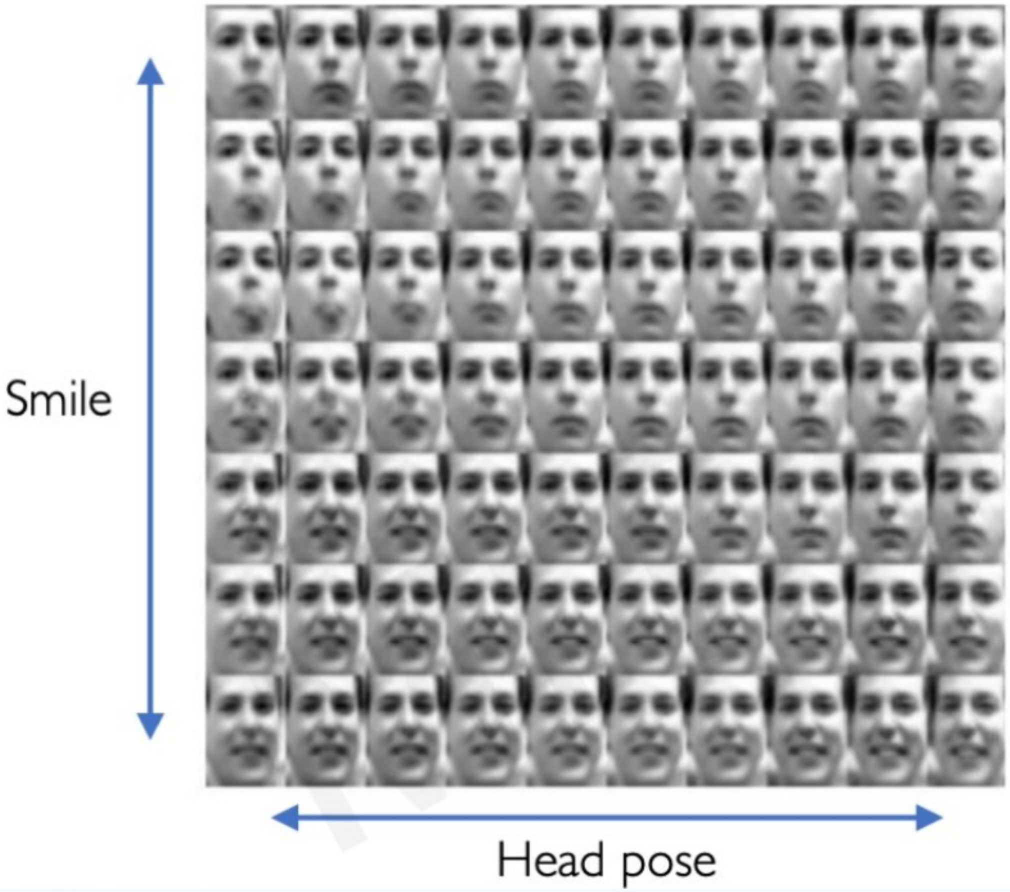
Head pose

Different dimensions of z encodes **different interpretable latent features**

Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

VAEs: Latent perturbation



Ideally, we want latent variables that are uncorrelated with each other

Enforce diagonal prior on the latent variables to encourage independence

Disentanglement

Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

Why latent variable models? Debiasing

Capable of uncovering **underlying latent variables** in a dataset

Homogeneous skin color, pose

vs

Diverse skin color, pose, illumination

How can we use latent distributions to create fair and representative datasets?

Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

Variational Autoencoders: Generating Data!

Diagonal prior on **z**
=> independent
latent variables

Different
dimensions of **z**
encode
interpretable factors
of variation

Also good feature representation that
can be computed using $q_\phi(z|x)$!

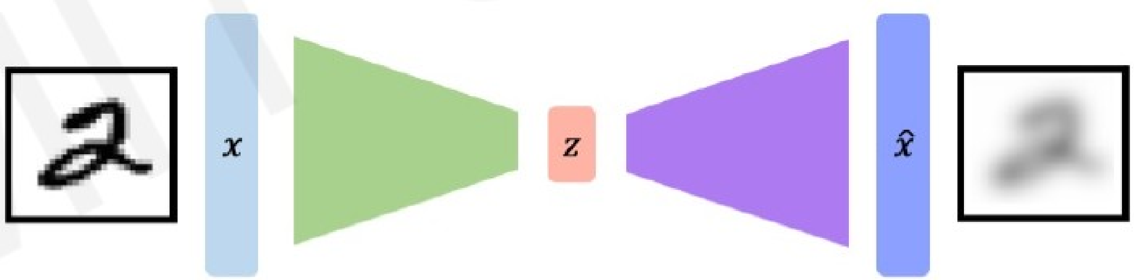
Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014

Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

VAE summary

1. Compress representation of world to something we can use to learn
2. Reconstruction allows for unsupervised learning (no labels!)
3. Reparameterization trick to train end-to-end
4. Interpret hidden latent variables using perturbation
5. Generating new examples



Crédito imagem/slide: A. Soleimany

Autocodificadores Variacionais

Variational Autoencoders

Probabilistic spin to traditional autoencoders => allows generating data
Defines an intractable density => derive and optimize a (variational) lower bound

Pros:

- Principled approach to generative models
- Interpretable latent space.
- Allows inference of q(z|x), can be useful feature representation for other tasks

Cons:

- Maximizes lower bound of likelihood: okay, but not as good evaluation as PixelRNN/PixelCNN
- Samples blurrier and lower quality compared to state-of-the-art (GANs)

Active areas of research:

- More flexible approximations, e.g. richer approximate posterior instead of diagonal Gaussian, e.g., Gaussian Mixture Models (GMMs), Categorical Distributions.
- Learning disentangled representations.

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Redes Adversárias Generativas

Generative Adversarial Networks

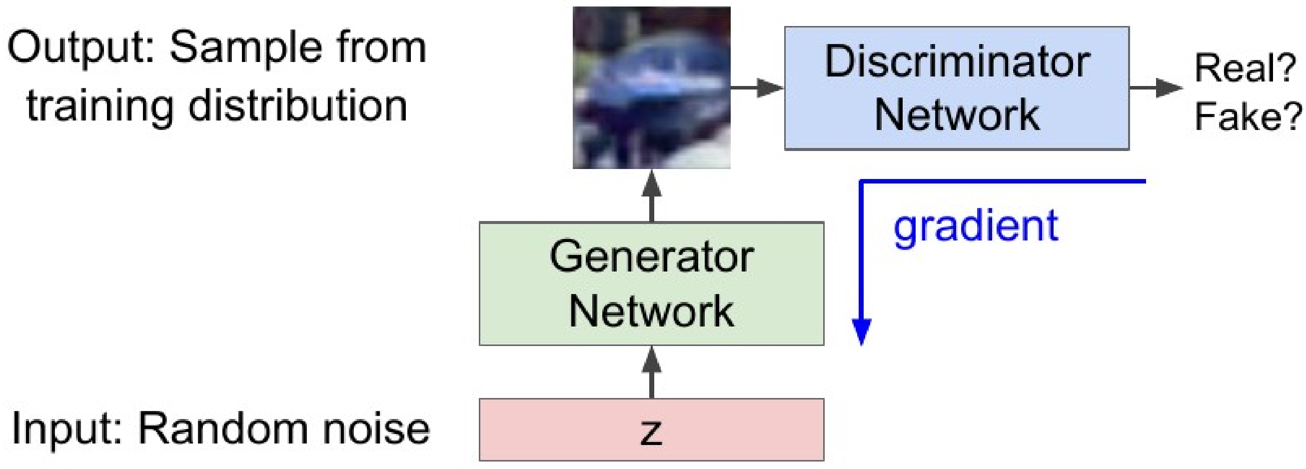
Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Problem: Want to sample from complex, high-dimensional training distribution. No direct way to do this!

Solution: Sample from a simple distribution we can easily sample from, e.g. random noise. Learn transformation to training distribution.

But we don't know which sample z maps to which training image -> can't learn by reconstructing training images

Solution: Use a discriminator network to tell whether the generate image is within data distribution ("real") or not



Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

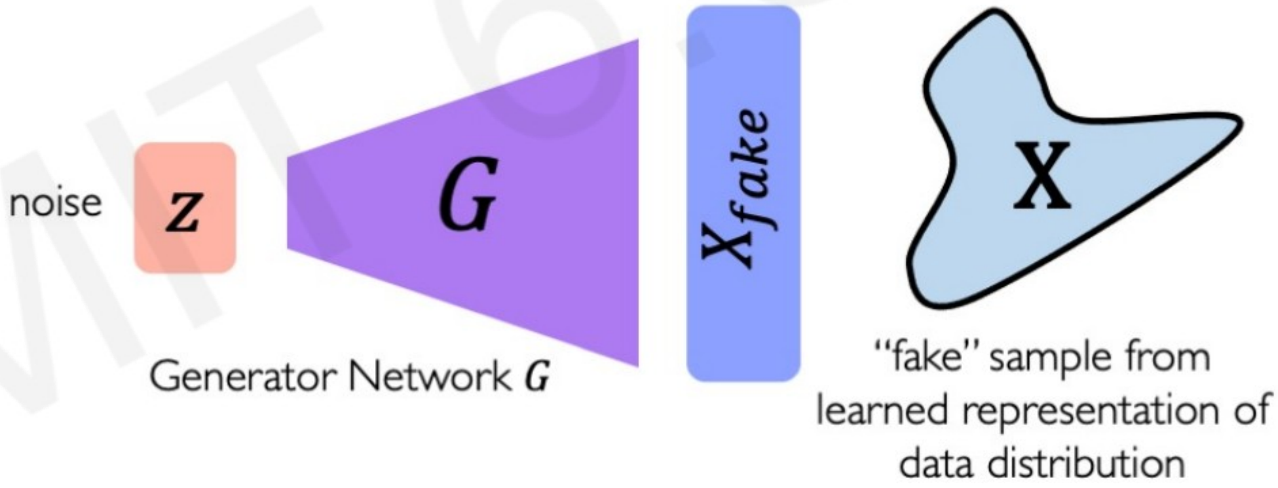
Redes Adversárias Generativas

What if we just want to sample?

Idea: don't explicitly model density, and instead just sample to generate new instances.

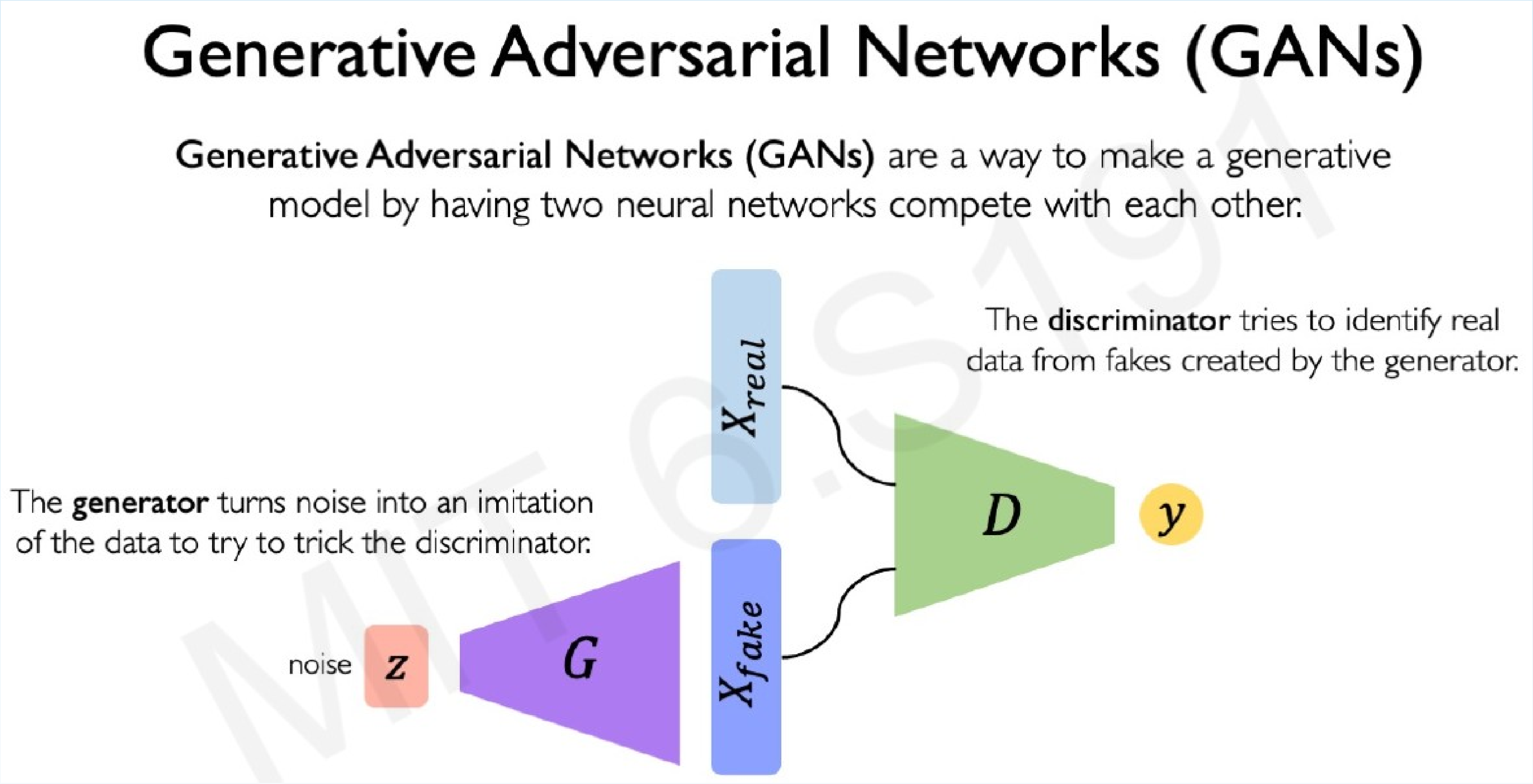
Problem: want to sample from complex distribution – can't do this directly!

Solution: sample from something simple (e.g., noise), learn a transformation to the data distribution.



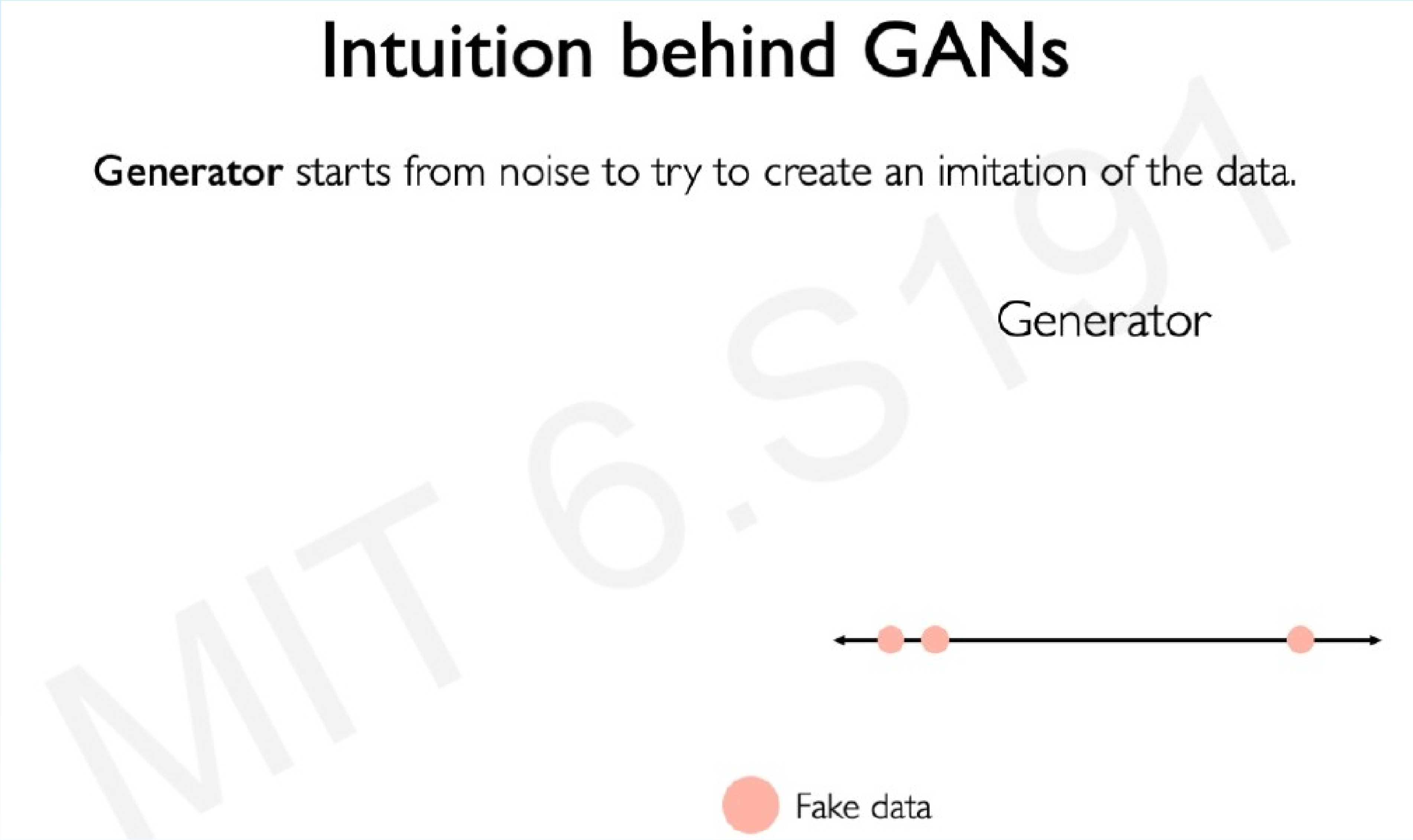
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



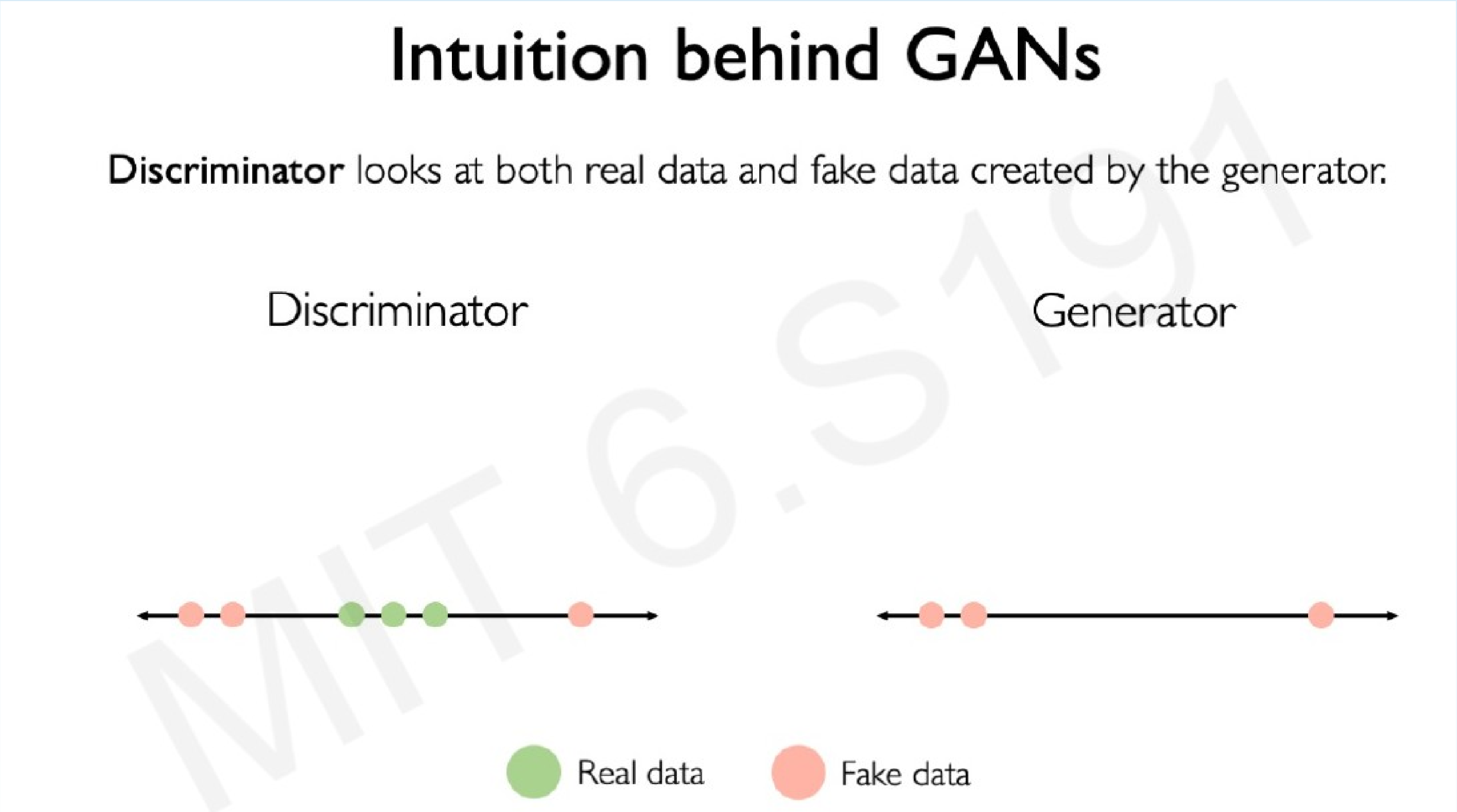
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



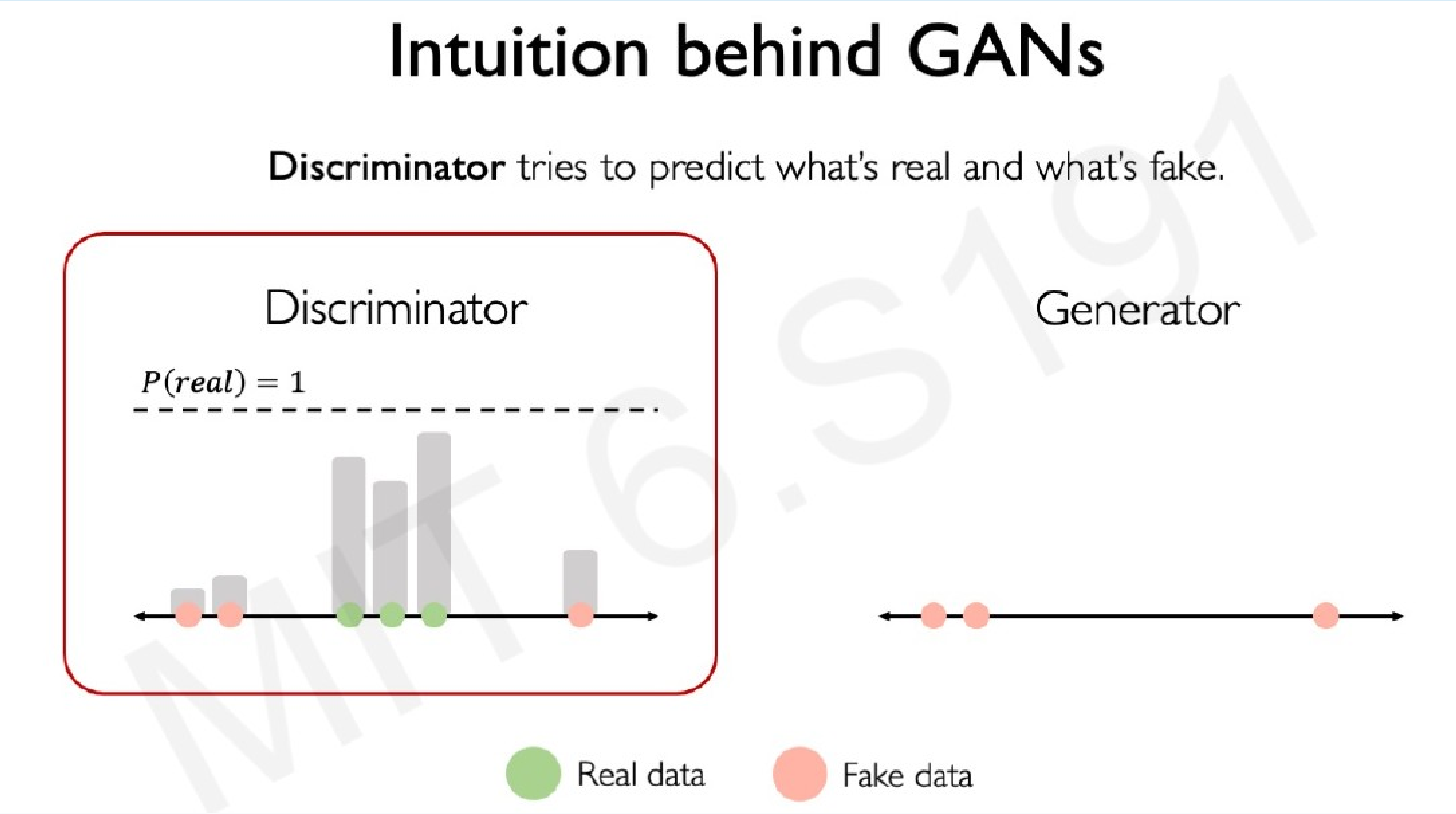
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



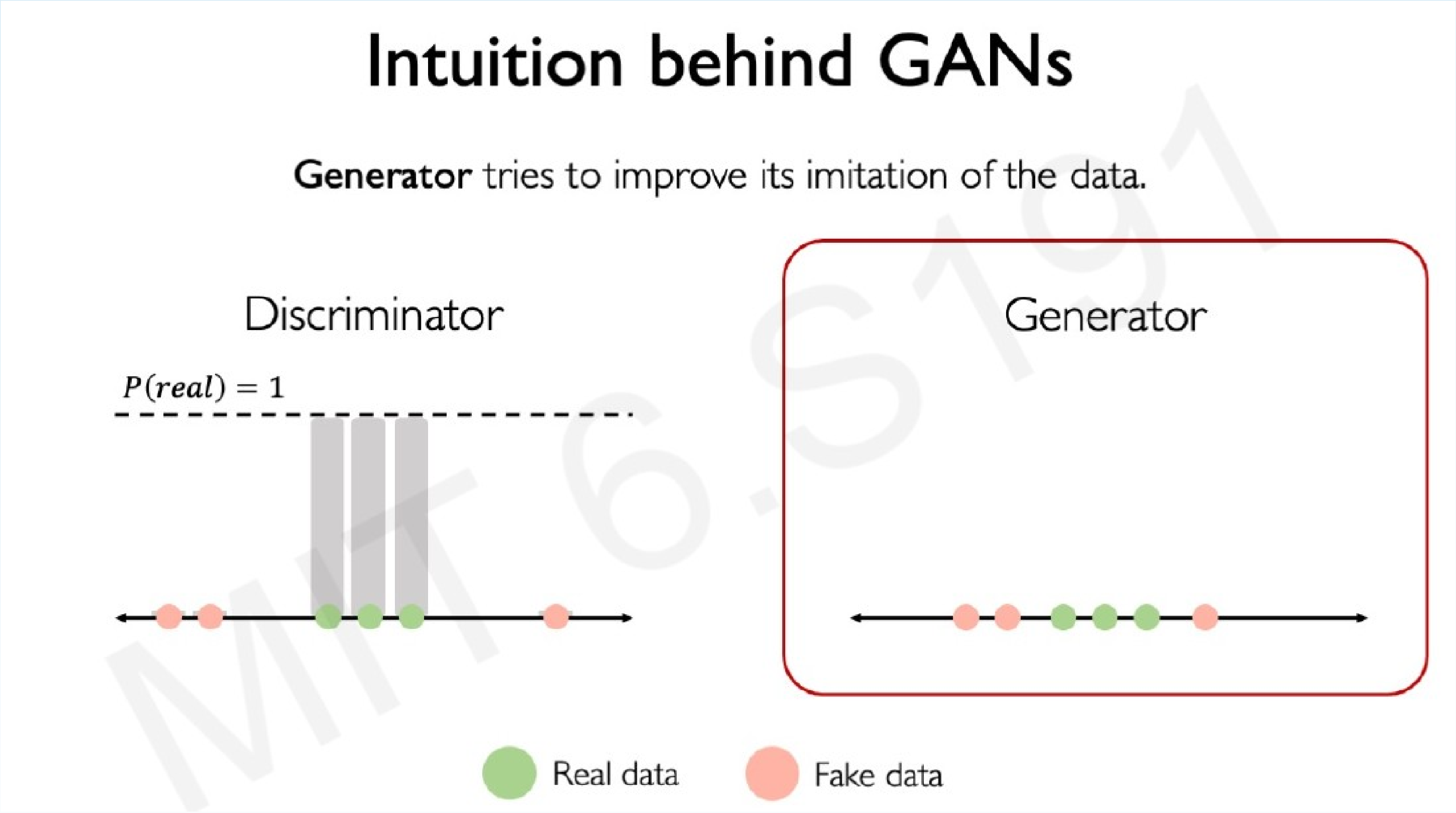
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



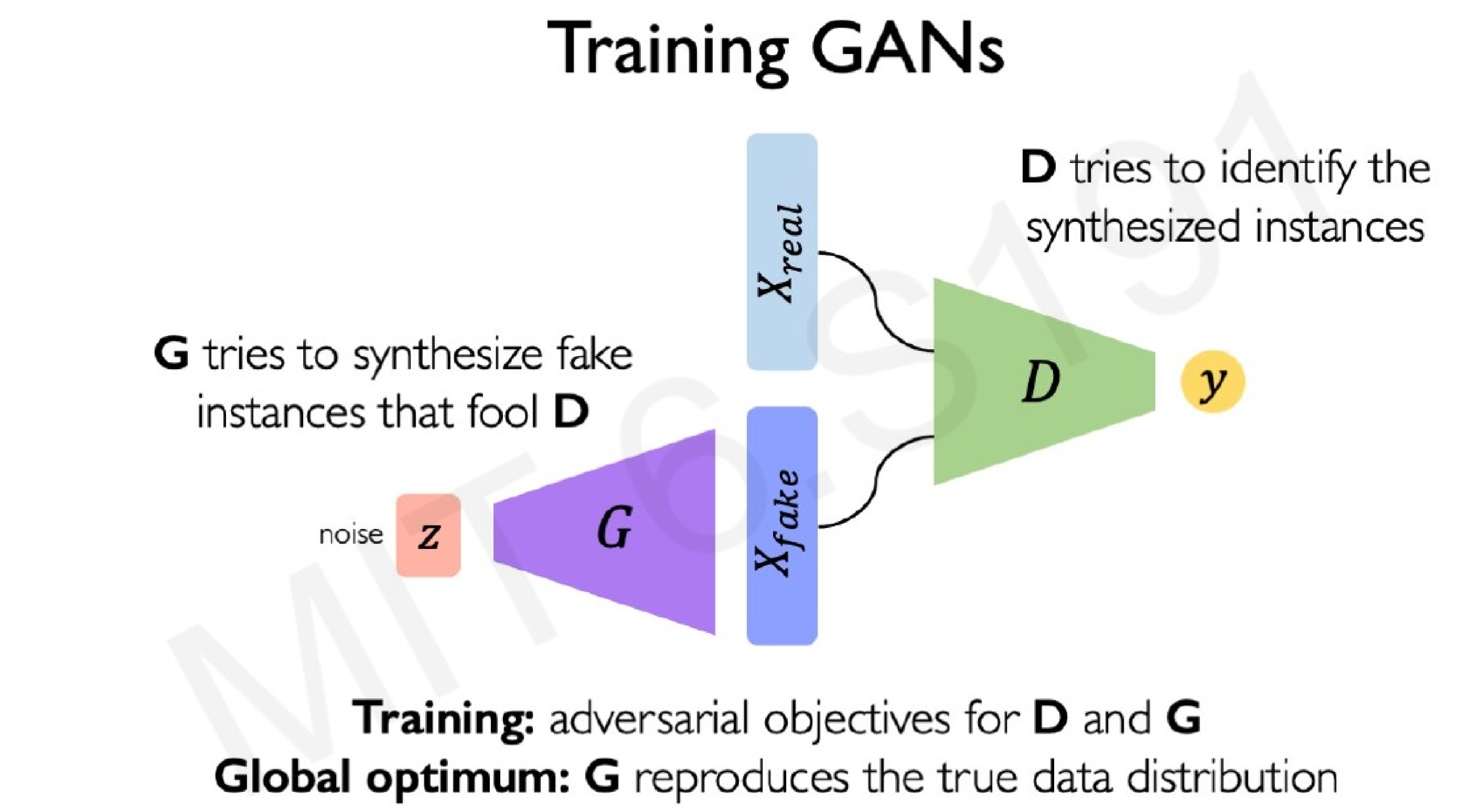
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



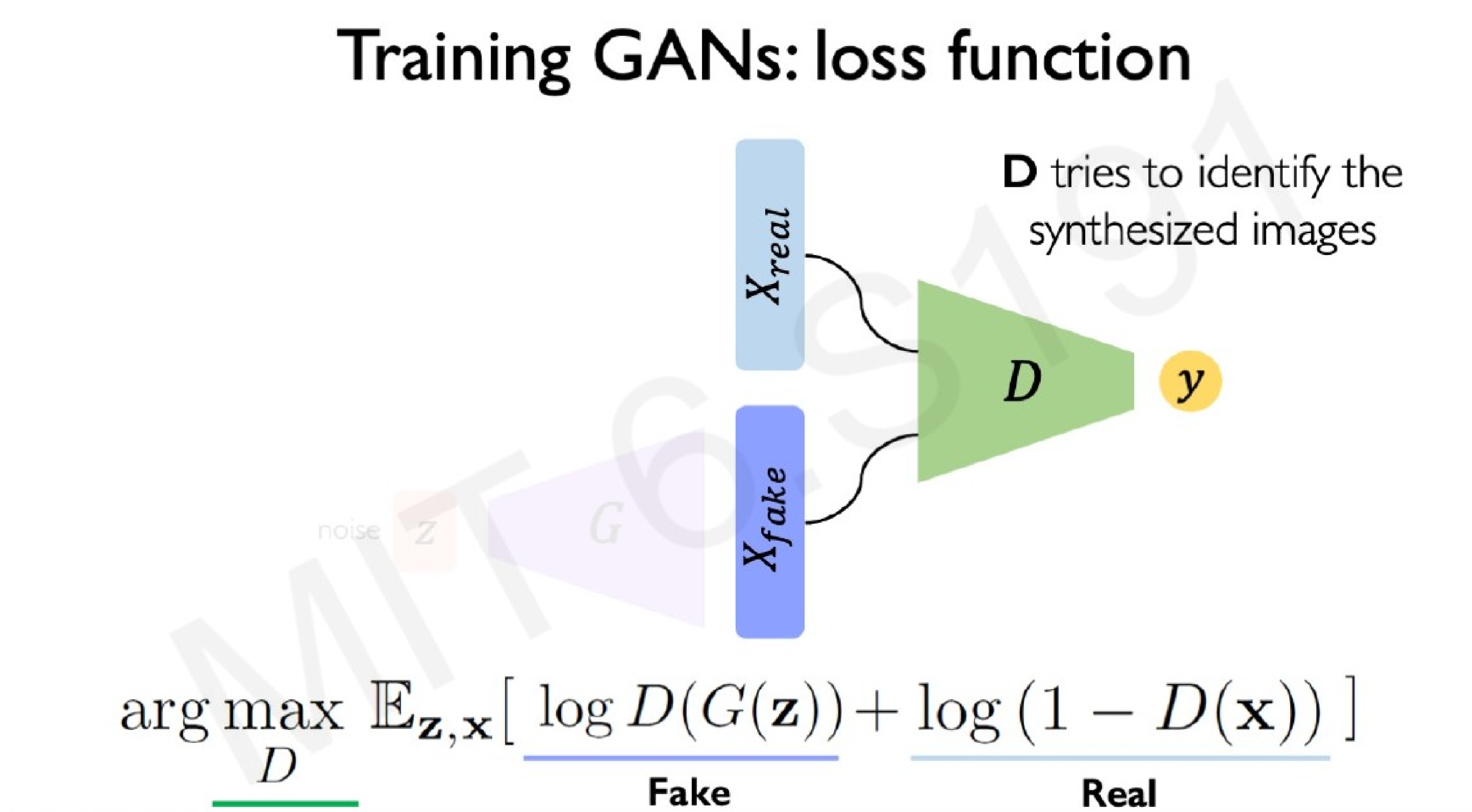
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



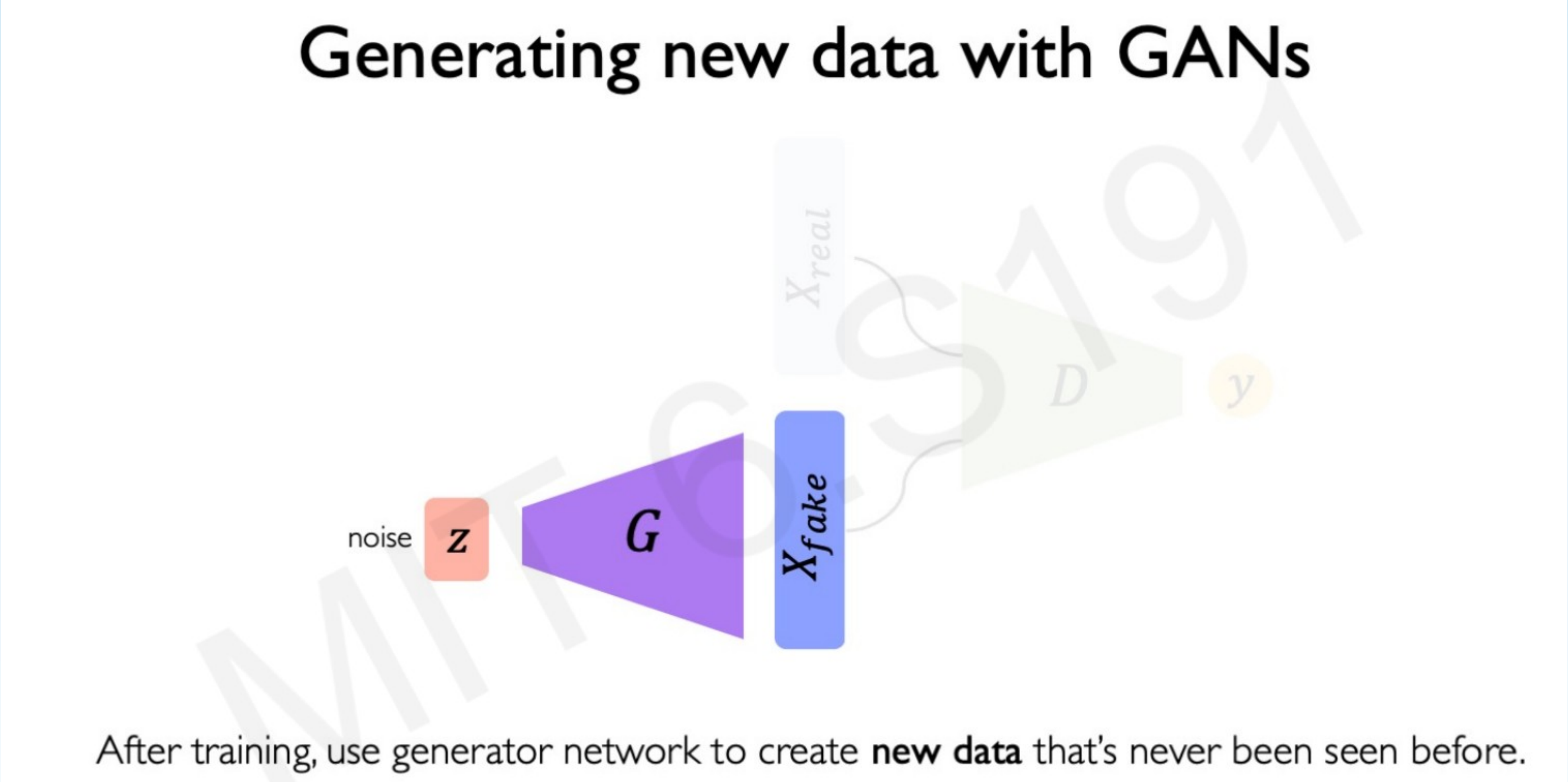
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



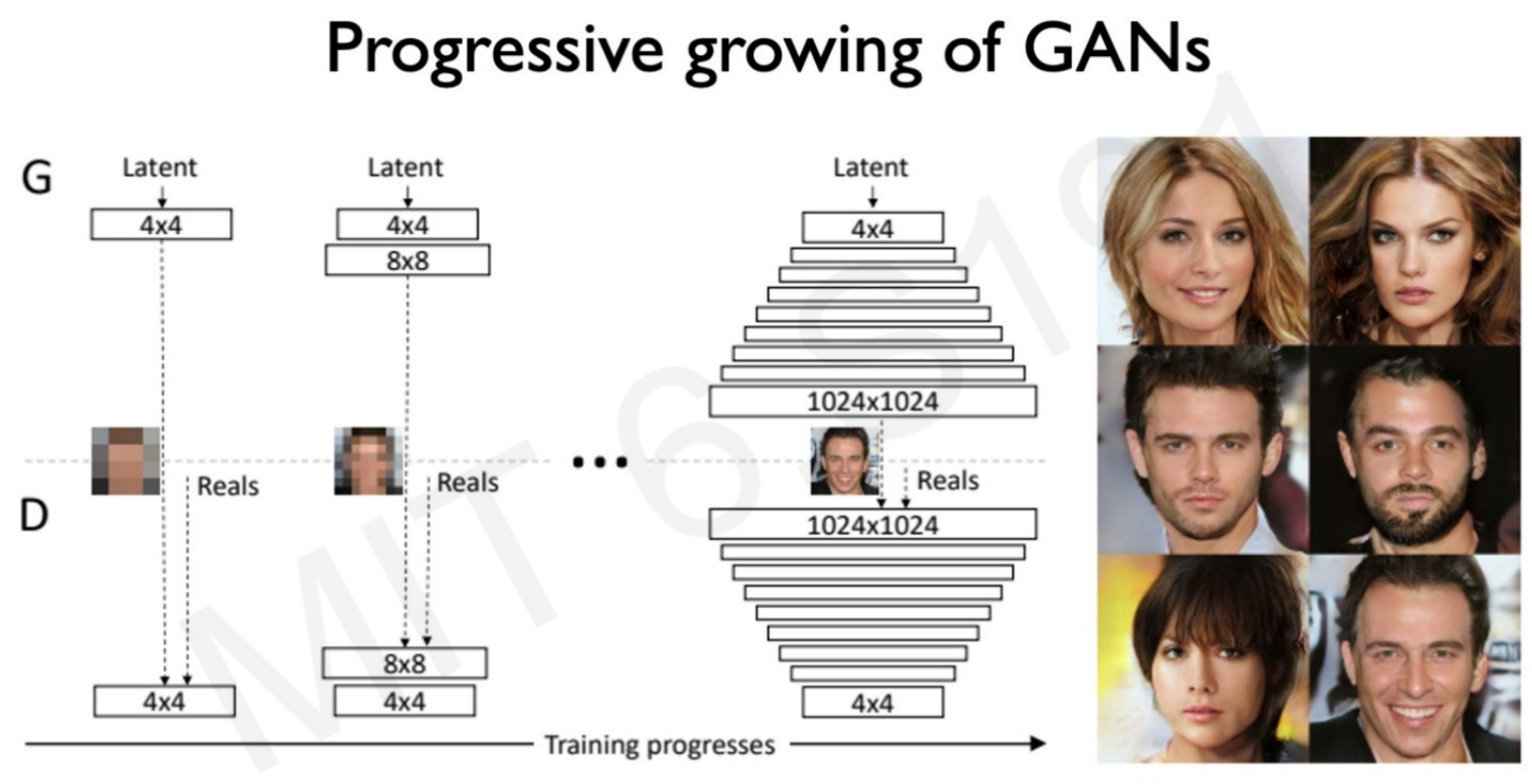
Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas



Crédito imagem/slide: A. Soleimany

Redes Adversárias Generativas

Training GANs: Two-player game

Ian Goodfellow et al., "Generative Adversarial Nets", NIPS 2014

Putting it together: GAN training algorithm

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.
- Sample minibatch of m examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data generating distribution $p_{\text{data}}(x)$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)}))) \right]$$

end for

- Sample minibatch of m noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$.
- Update the generator by ascending its stochastic gradient (improved objective):

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$$

end for

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Redes Adversárias Generativas

2017: Explosion of GANs

See also: https://github.com/soumith/ganhacks for tips and tricks for trainings GANs

"The GAN Zoo"

- GAN - Generative Adversarial Networks
- 3D-GAN - Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling
- acGAN - Face Aging With Conditional Generative Adversarial Networks
- AC-GAN - Conditional Image Synthesis With Auxiliary Classifier GANs
- AdaGAN - AdaGAN: Boosting Generative Models
- AEGAN - Learning Inverse Mapping by Autoencoder based Generative Adversarial Nets
- AffGAN - Amortised MAP Inference for Image Super-resolution
- AL-CGAN - Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts
- ALI - Adversarially Learned Inference
- AM-GAN - Generative Adversarial Nets with Labeled Data by Activation Maximization
- AnoGAN - Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery
- ArtGAN - ArtGAN: Artwork Synthesis with Conditional Categorial GANs
- b-GAN - b-GAN: Unified Framework of Generative Adversarial Networks
- Bayesian GAN - Deep and Hierarchical Implicit Models
- BEGAN - BEGAN: Boundary Equilibrium Generative Adversarial Networks
- BiGAN - Adversarial Feature Learning
- BS-GAN - Boundary-Seeking Generative Adversarial Networks
- CGAN - Conditional Generative Adversarial Nets
- CaloGAN - CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks
- CCGAN - Semi-Supervised Learning with Context-Conditional Generative Adversarial Networks
- CatGAN - Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks
- CoGAN - Coupled Generative Adversarial Networks
- Context-RNN-GAN - Contextual RNN-GANs for Abstract Reasoning Diagram Generation
- C-RNN-GAN - C-RNN-GAN: Continuous recurrent neural networks with adversarial training
- CS-GAN - Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets
- CVAE-GAN - CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training
- CycleGAN - Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks
- DTN - Unsupervised Cross-Domain Image Generation
- DCGAN - Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
- DiscoGAN - Learning to Discover Cross-Domain Relations with Generative Adversarial Networks
- DR-GAN - Disentangled Representation Learning GAN for Pose-Invariant Face Recognition
- DualGAN - DualGAN: Unsupervised Dual Learning for Image-to-Image Translation
- EBGAN - Energy-based Generative Adversarial Network
- f-GAN - f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
- FF-GAN - Towards Large-Pose Face Frontalization in the Wild
- GAWWN - Learning What and Where to Draw
- GeneGAN - GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data
- Geometric GAN - Geometric GAN
- GoGAN - Gang of GANs: Generative Adversarial Networks with Maximum Margin Ranking
- GP-GAN - GP-GAN: Towards Realistic High-Resolution Image Blending
- IAN - Neural Photo Editing with Introspective Adversarial Networks
- iGAN - Generative Visual Manipulation on the Natural Image Manifold
- IcGAN - Invertible Conditional GANs for image editing
- ID-CGAN - Image De-raining Using a Conditional Generative Adversarial Network
- Improved GAN - Improved Techniques for Training GANs
- InfoGAN - InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets
- LAGAN - Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis
- LAPGAN - Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks

https://github.com/hindupuravinash/the-gan-zoo

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Redes Adversárias Generativas

2017: Explosion of GANs

Better training and generation

LSGAN, Zhu 2017.

Wasserstein GAN, Arjovsky 2017. Improved Wasserstein GAN, Gulrajani 2017.

Progressive GAN, Karras 2018.

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Redes Adversárias Generativas

2017: Explosion of GANs

Source->Target domain transfer

CycleGAN. Zhu et al. 2017.

Text -> Image Synthesis

Reed et al. 2017.

Many GAN applications

Pix2pix. Isola 2017. Many examples at https://phillipi.github.io/pix2pix/

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Summary: GANs

Don't work with an explicit density function
Take game-theoretic approach: learn to generate from training distribution through 2-player game

Pros:
- Beautiful, state-of-the-art samples!

Cons:
- Trickier / more unstable to train
- Can't solve inference queries such as p(x), p(z|x)

Active areas of research:
- Better loss functions, more stable training (Wasserstein GAN, LSGAN, many others)
- Conditional GANs, GANs for all kinds of applications

Crédito imagem/slide: L. Fei-Fei, R. Krishna, D. Xu

Referências Bibliográficas

www.deeplearningbook.org (Capítulo 20)