Introduction to Deep Learning

Lecture 6

Generative Models

Maria Vakalopoulou & Stergios Christodoulidis

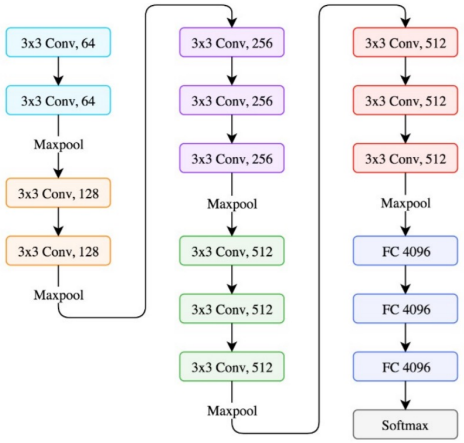
MICS Laboratory

CentraleSupélec
Université Paris-Saclay

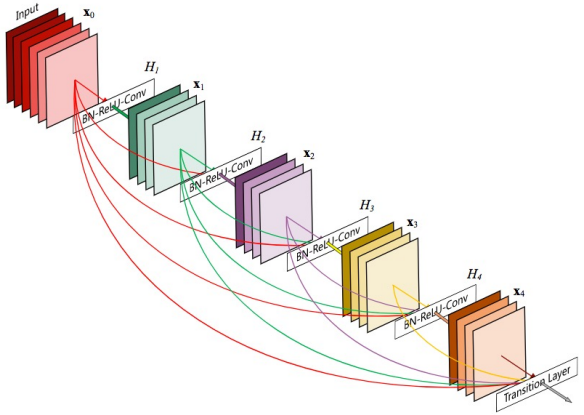
Wednesday, December 01, 2021

Last Lectures

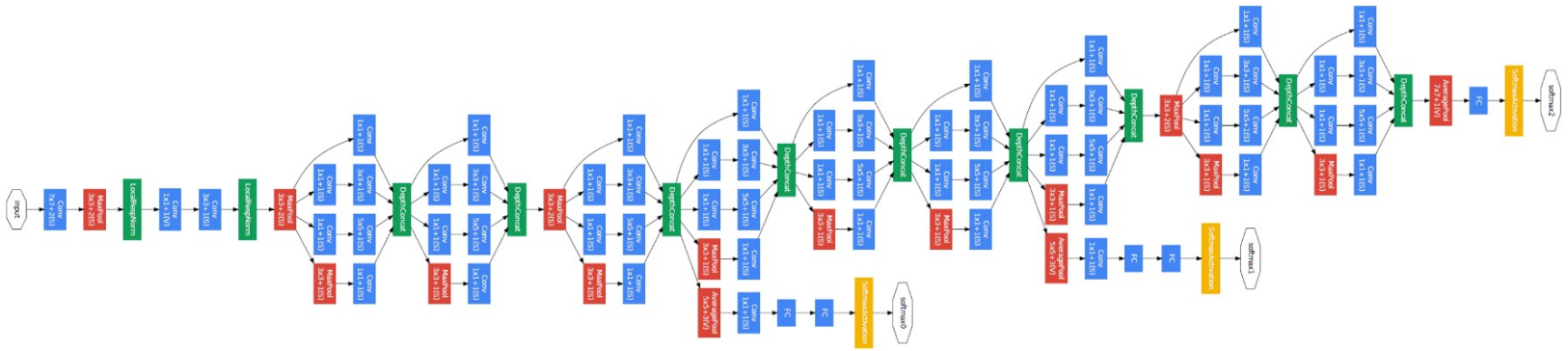
Modern Classification Convolutional Neural Networks



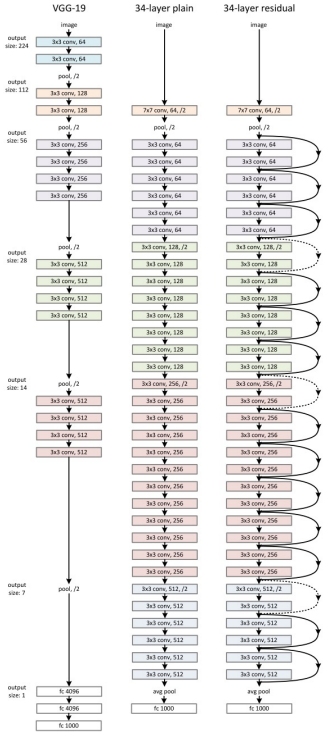
VGG [1]



DenseNet [3]



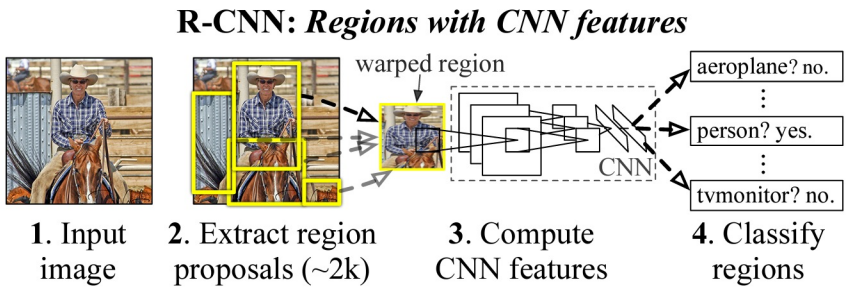
Inception Network [2]



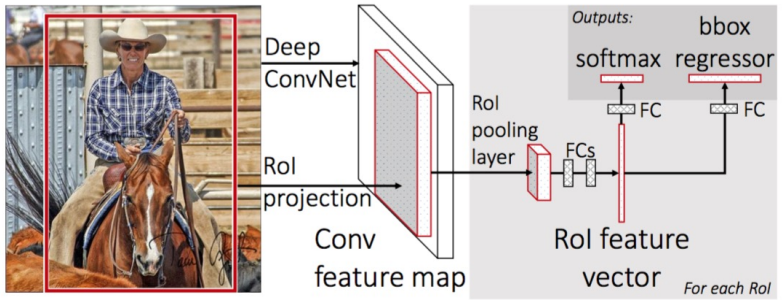
ResNet [4]

[1] https://arxiv.org/pdf/1409.1556.pdf
[2] https://arxiv.org/pdf/1409.4842.pdf
[3] https://arxiv.org/pdf/1608.06993.pdf
[4] https://arxiv.org/pdf/1512.03385.pdf

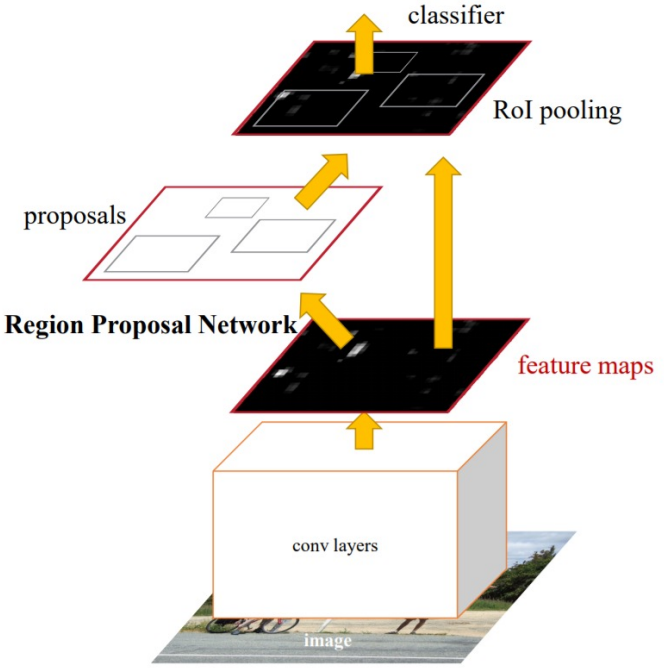
Modern Localization Convolutional Neural Networks



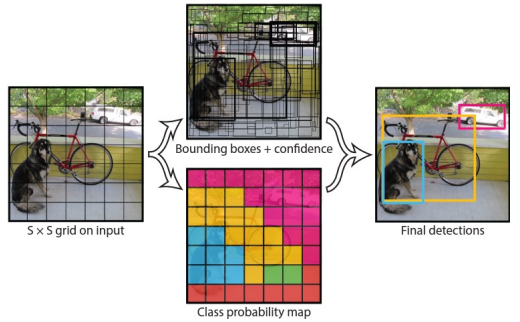
R-CNN [1]



Fast R-CNN [2]



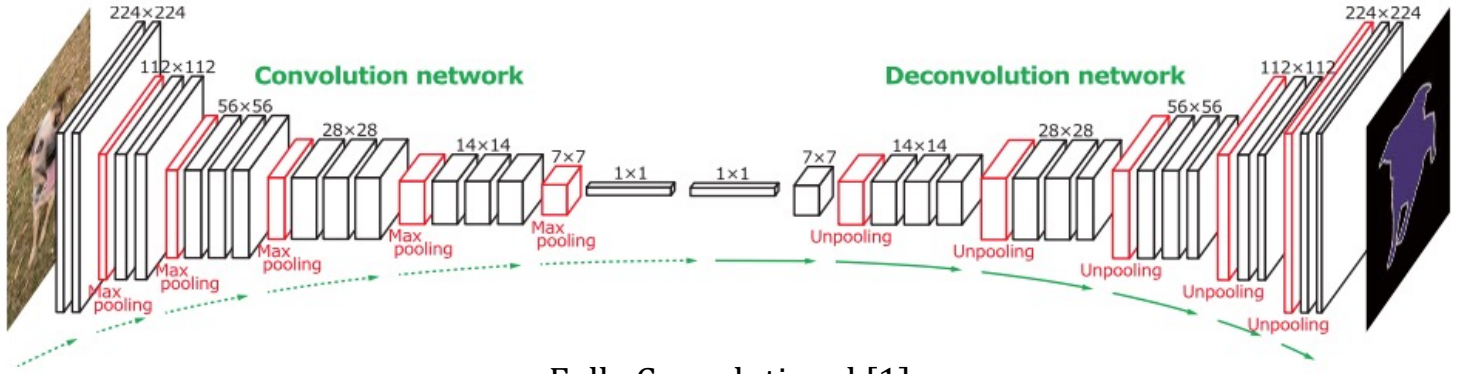
Faster R-CNN [3]



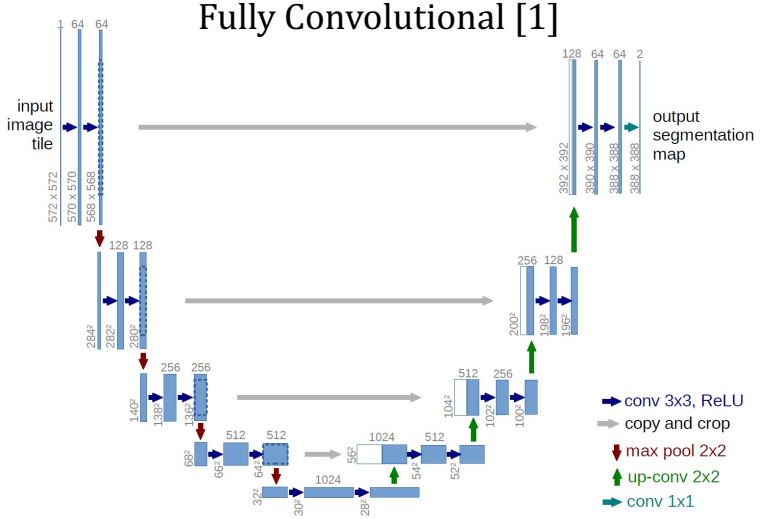
YOLO [4]

[1] https://arxiv.org/pdf/1311.2524.pdf
[2] https://arxiv.org/pdf/1504.08083.pdf
[3] https://arxiv.org/pdf/1506.01497.pdf
[4] https://arxiv.org/pdf/1506.02640v5.pdf

Fully Convolutional Neural Networks



Fully Convolutional [1]



UNet [2]

[1] https://arxiv.org/pdf/1411.4038.pdf
[2] https://arxiv.org/pdf/1505.04597.pdf

Today's Lecture

Today's Lecture

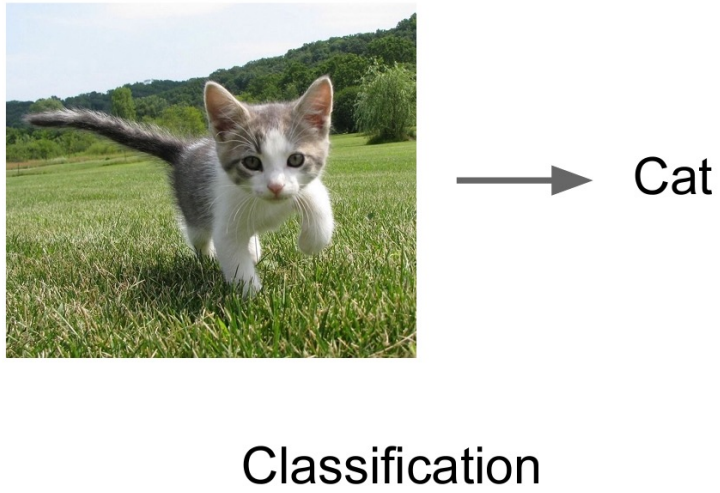
- Supervised vs Unsupervised Learning
- Discriminative vs Generative Models
- Latent Variable Models
  - Autoencoders (AE)
  - Variational Autoencoders (VAE)
  - Reparameterization Trick

Supervised vs Unsupervised Learning

Supervised Learning

- So far, we mainly discussed about <u>supervised learning.</u>

- **Data:** (X, y)
  - X is the input data
  - y is the label
- **Task:** Learn a function f_θ (or a model $\mathcal{M}$) to map X -> y
- **Examples:**
  - Classification
  - Object detection
  - Semantic Segmentation etc.

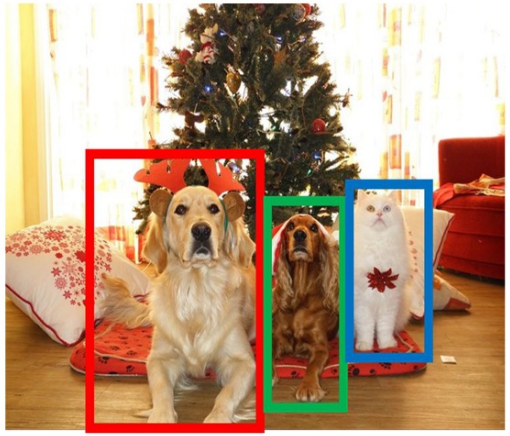


[picture ref] http://cs231n.stanford.edu/schedule.html

Supervised Learning

- So far, we mainly discussed about <u>supervised learning.</u>

- **Data:** (X, y)
  - X is the input data
  - y is the label
- **Task:** Learn a function f_θ (or a model $\mathcal{M}$) to map X -> y
- **Examples:**
  - Classification
  - Object detection
  - Semantic Segmentation etc.



DOG, DOG, CAT

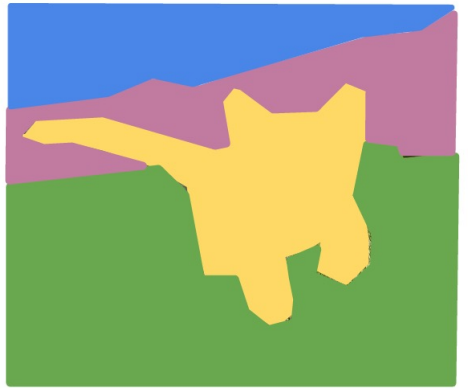
Object Detection

[picture ref] http://cs231n.stanford.edu/schedule.html

Supervised Learning

- So far, we mainly discussed about <u>supervised learning.</u>

- **Data:** (X, y)
  - X is the input data
  - y is the label
- **Task:** Learn a function f_θ (or a model $\mathcal{M}$) to map X -> y
- **Examples:**
  - Classification
  - Object detection
  - Semantic Segmentation etc.



Semantic Segmentation

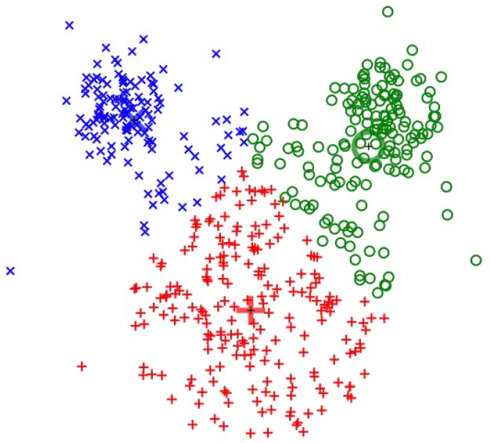
[picture ref] http://cs231n.stanford.edu/schedule.html

Unsupervised Learning

- There are settings where we are interested in investigating data in an unsupervised way.

- **Data:** X
  - X is the input data
  - Not labels available!
- **Task:** Learn the underlying structure of the data.

- **Examples:**
  - Clustering
  - Dimensionality reduction
  - Representation learning



K-means clustering

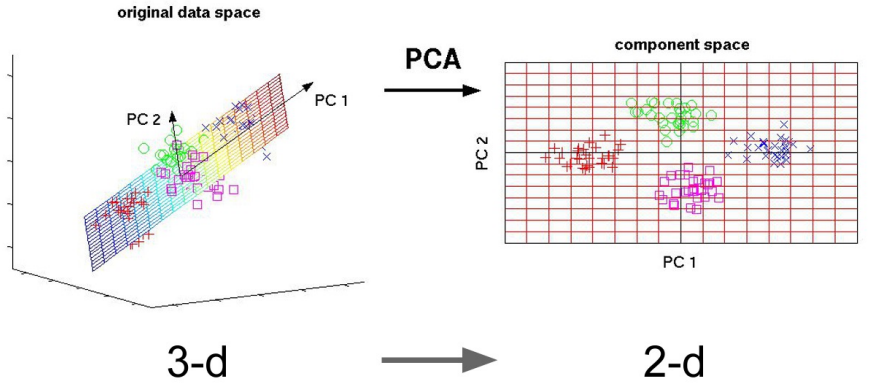
[picture ref] http://cs231n.stanford.edu/schedule.html

Unsupervised Learning

- There are settings where we are interested in investigating data in an unsupervised way.

- **Data:** X
  - X is the input data
  - Not labels available!
- **Task:** Learn the underlying structure of the data.

- **Examples:**
  - Clustering
  - Dimensionality reduction
  - Representation learning



Principal Component Analysis
(Dimensionality reduction)

[picture ref] http://cs231n.stanford.edu/schedule.html

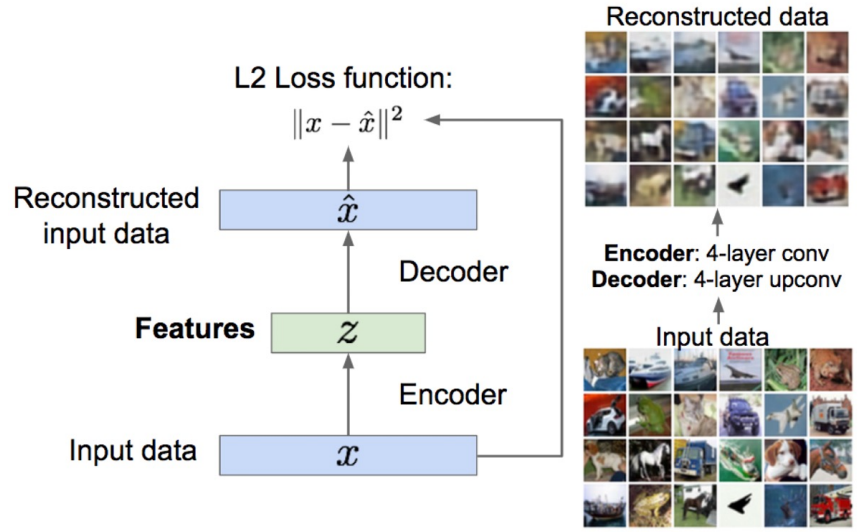
Unsupervised Learning

- There are settings where we are interested in investigating data in an unsupervised way.

- **Data:** X
  - X is the input data
  - Not labels available!
- **Task:** Learn the underlying structure of the data.

- **Examples:**
  - Clustering
  - Dimensionality reduction
  - Representation learning

Feature Learning
(e.g. autoencoders)



[picture ref] https://web.eecs.umich.edu/~justincj/teaching/eecs498/FA2020/

Discriminative vs Generative Models

Probabilistic View

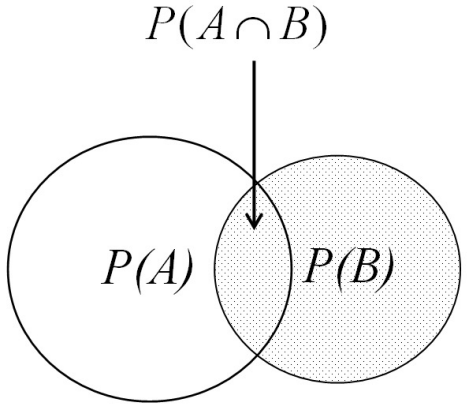
- Until now we have only described our models using functions.
  - E.g., $y = f_\theta(X)$ where y is our prediction, x is our input and f_θ is our model parametrized by θ.
  - Facilitates a clear understanding on how these models are implemented.

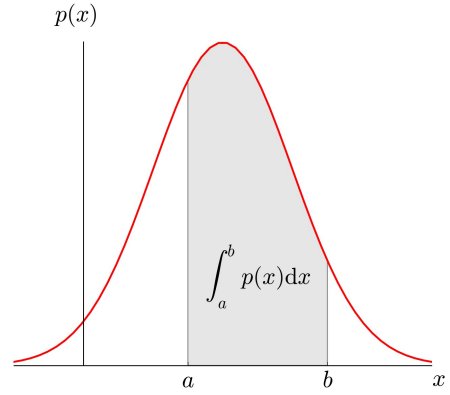
- We can also study our models under a probabilistic view.
  - We are not changing anything in terms of implementation.
  - We are just using another tool to better understand their characteristics.

- On this ground, we can:
  - Gain more insights on how our model works.
  - Help identifying different methods and schemes.

Some terminology

- Marginal Probability $p(A)$
  - Probability of event A.
- Conditional Probability $p(A|B)$ or $p(A \cap B)$
  - Probability of event A given B.
- Joint Probability $p(A, B)$ or $p(A \cup B)$
  - Probability of event A and B simultaneously
- Probability density function (PDF)
  - A function that specifies the probability of a variable to fall in a value or range of values.
- Bayes Rule:
  - $p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$



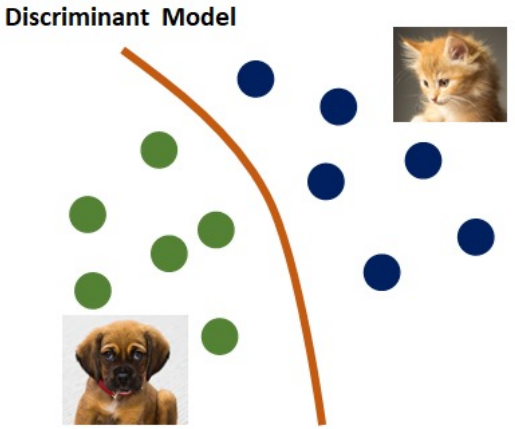


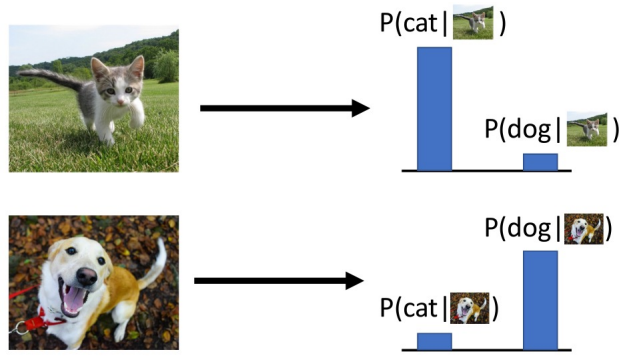
Discriminative Models

- Discriminative models are mostly used under a supervised learning scheme.

- Given an input $X \rightarrow$ define a model f_θ in order to predict y:
  - The correct label (classification)
  - The correct value (regression)

- <u>Probabilistic View:</u> Our model captures the probability of the observation X (image) to belong to class y.
  - $p(y|X;\theta)$, meaning: our model seeks maximize the conditional probability of y given an input X and model f_θ.



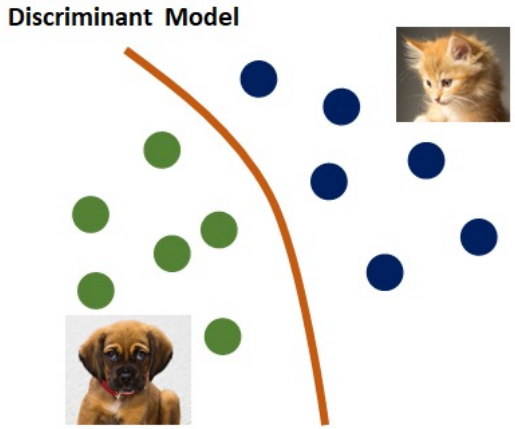


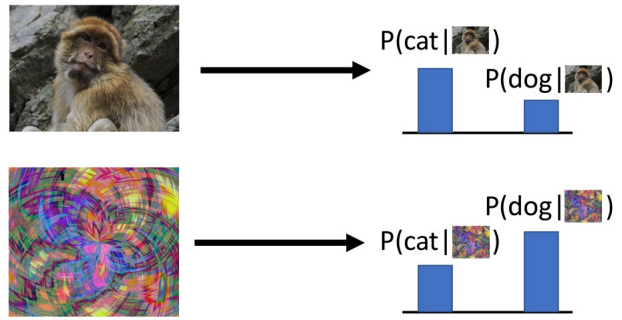
Discriminative Models

- Discriminative models are mostly used under a supervised learning scheme.

- Given an input $X \rightarrow$ define a model f_θ in order to predict y:
  - The correct label (classification)
  - The correct value (regression)

- <u>Probabilistic View:</u> Our model captures the probability of the observation X (image) to belong to class y.
  - $p(y|X;\theta)$, meaning: our model seeks maximize the conditional probability of y given an input X and model f_θ.





*No way to handle unreasonable inputs

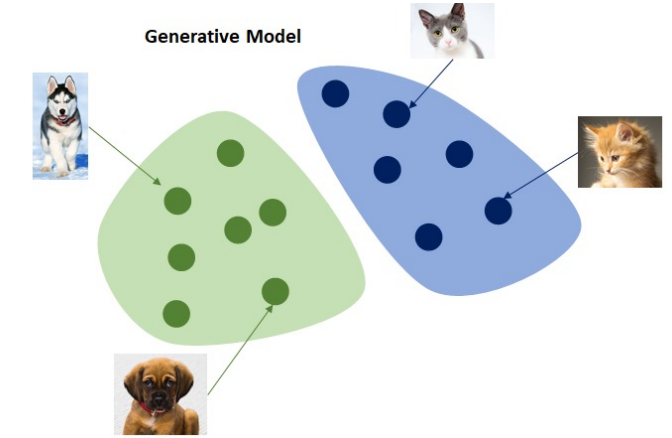
Can we Solve the Prediction Task in Another Way?

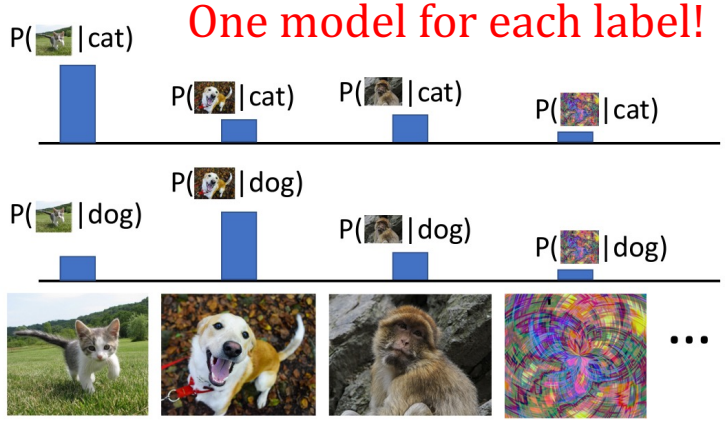
Generative Models

Probabilistic View:

- Generative models capture the underlying distribution that describe the data!

- In other words: Our model learns the probability of the image to come from to the distribution of our dataset.
  - $p(X \mid y)$, meaning: our model seeks to maximize the conditional probability of X given y (Supervised learning, one model for each label)





One model for each label!

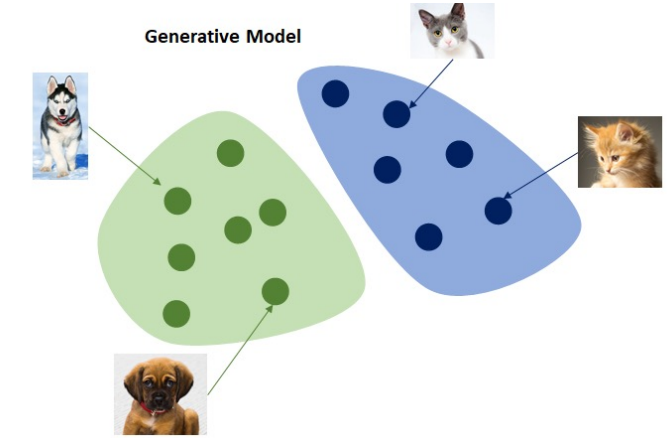
Much more difficult task!

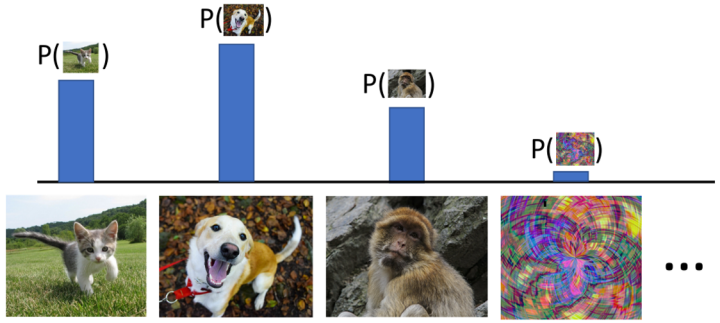
Generative Models

Probabilistic View:

- Generative models capture the underlying distribution that describe the data!

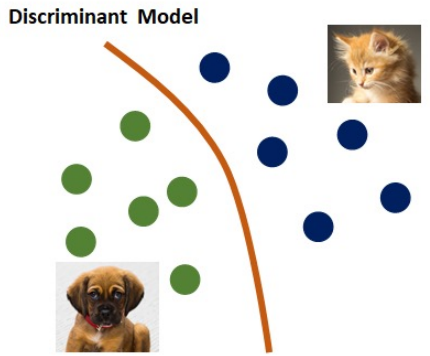
- In other words: Our model learns the probability of the image to come from to the distribution of our dataset.
  - $p(X \mid y)$, meaning: our model seeks to maximize the conditional probability of X given y (Supervised learning, one model for each label)
  - $p(X)$, meaning: our model seeks only the marginal probability of the observations if there are no labels (Unsupervised learning).



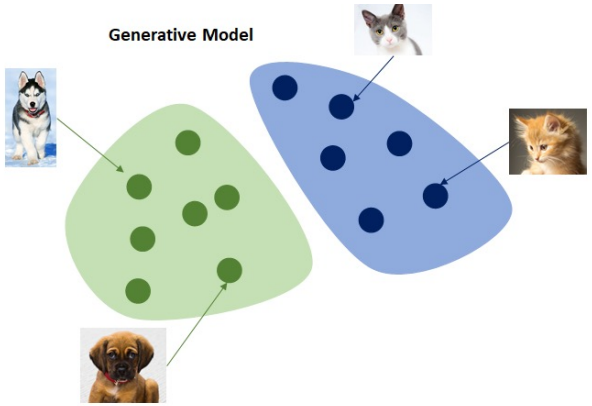


Discriminative vs Generative Learning

- Discriminative Models:
  - Discriminate between different kinds of data instances.
  - Captures the conditional probability: $p(y \mid X)$
  - Task-oriented
  - E.g., Logistic Regression, SVM, Classification Deep Nets



- Generative Models:
  - Models the underlying distribution that describe the data.
  - Captures the joint probability: $p(X \mid y)$ or just $p(X)$
  - Model the world → Perform tasks, e.g., use Bayes rule to classify, i.e., $p(y \mid X)$
  - Can be used to generate data
  - Autoregressive models, Variational Autoencoders, GANs

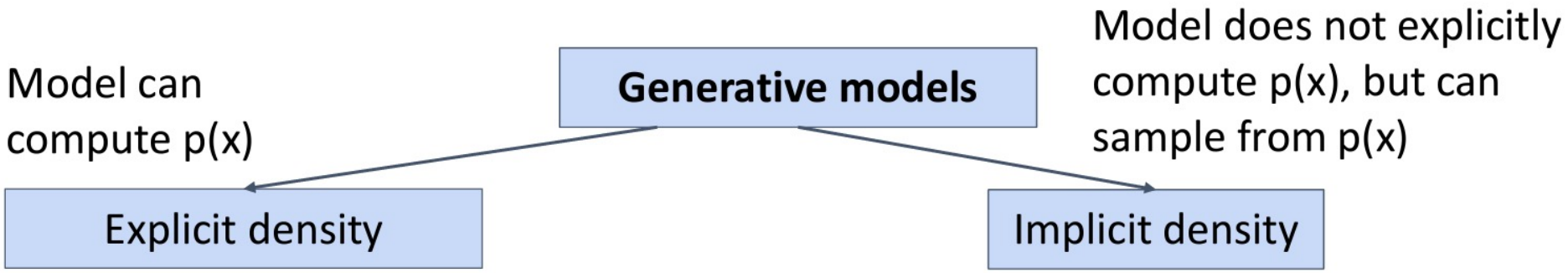


Joint probability $p(X, y)$ can also be captured.

Taxonomy of Generative Models

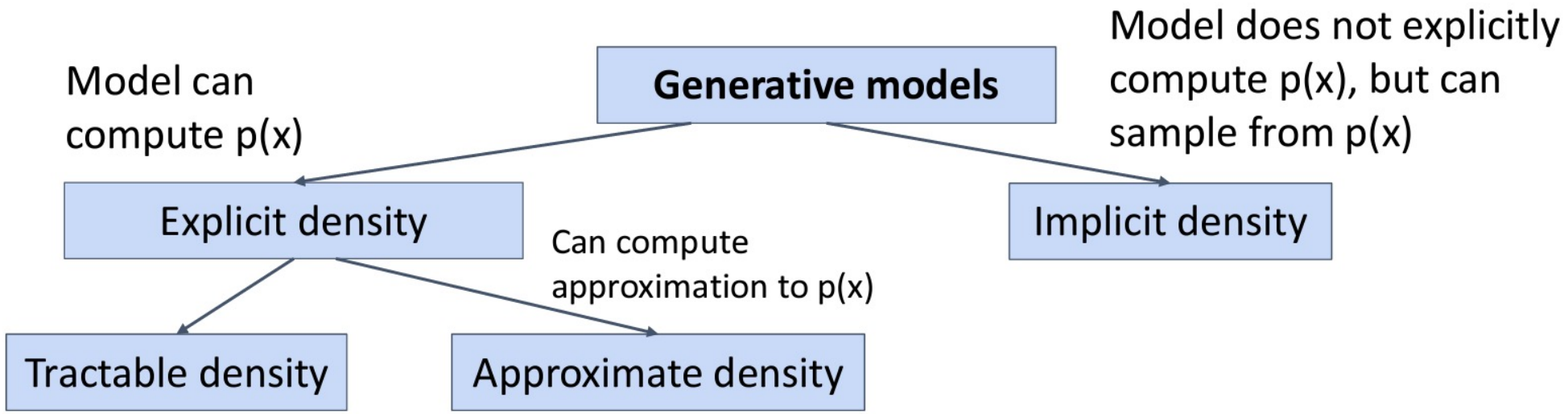
Generative models

Taxonomy of Generative Models



[picture ref] https://web.eecs.umich.edu/~justincj/teaching/eecs498/FA2020/

Taxonomy of Generative Models

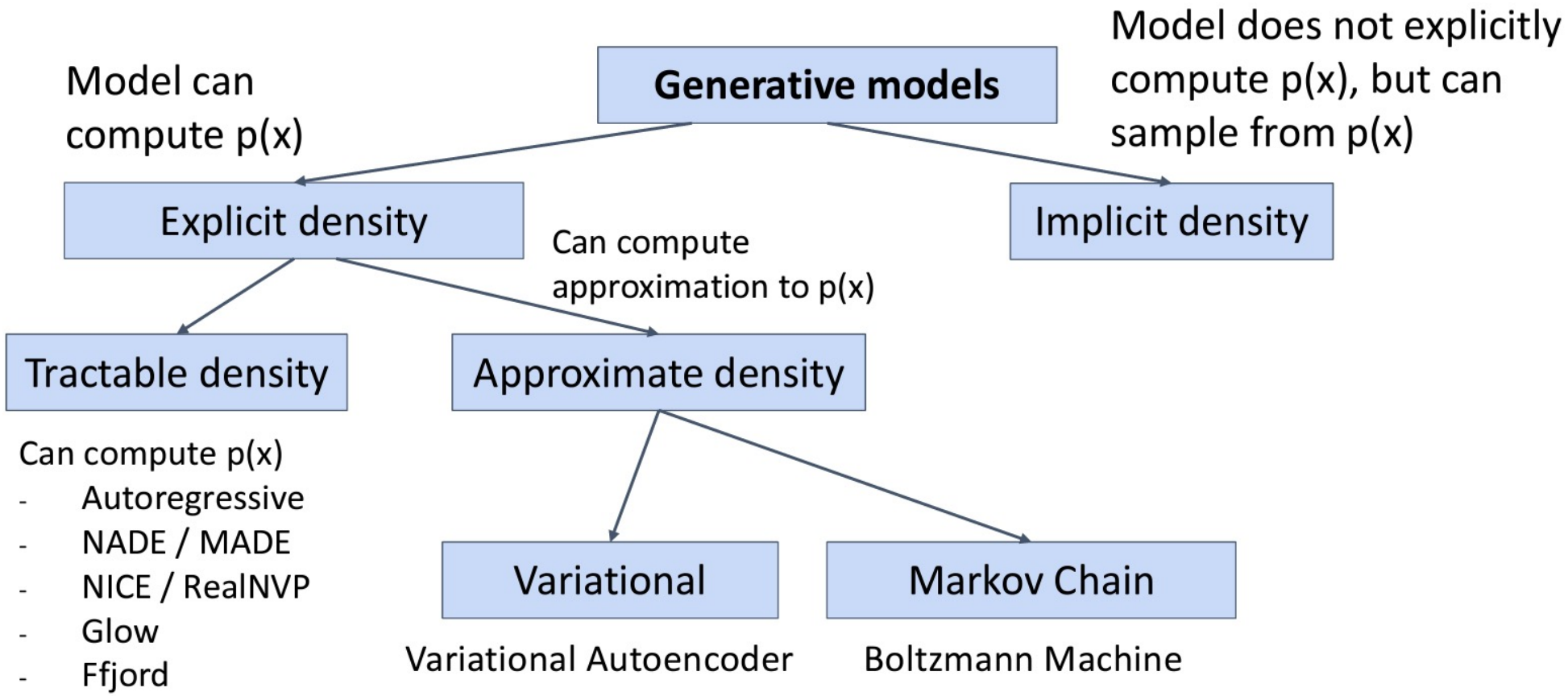


Can compute p(x)

- Autoregressive
- NADE / MADE
- NICE / RealNVP
- Glow
- Ffjord

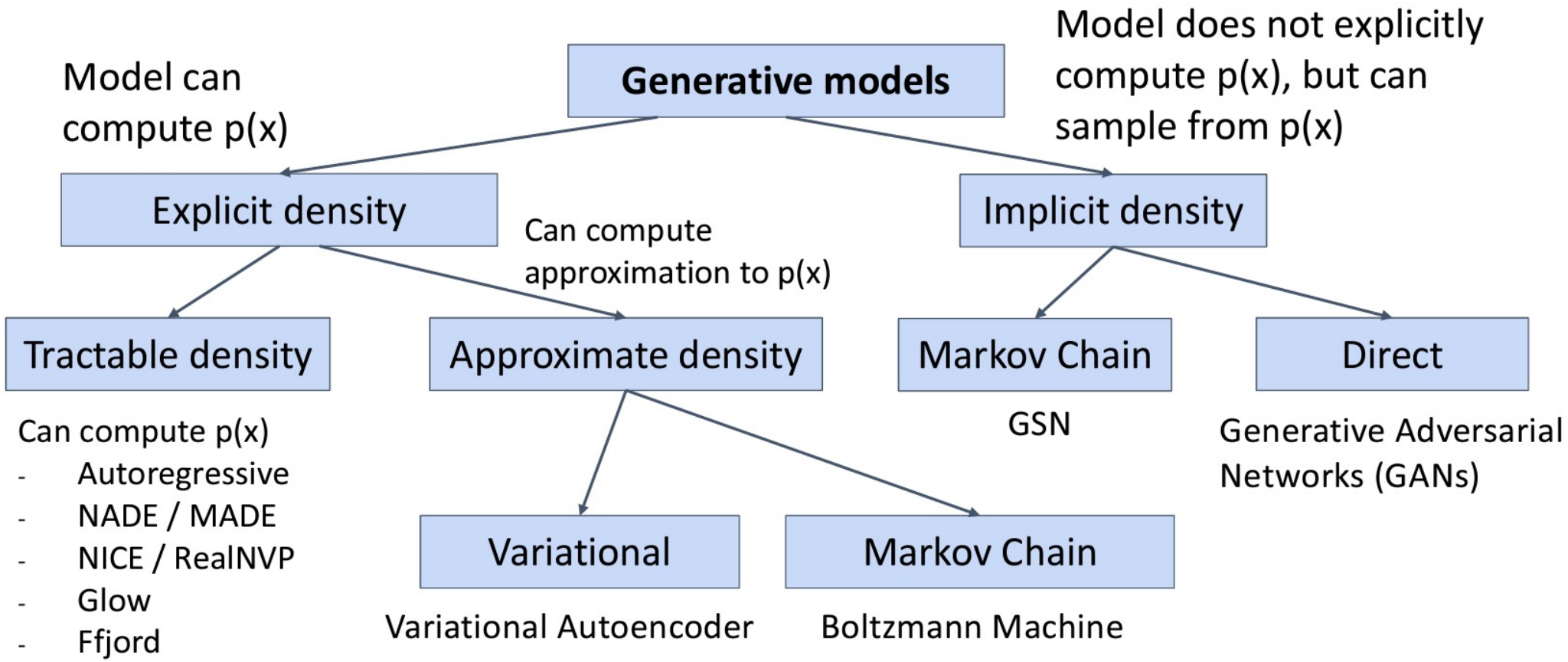
[picture ref] https://web.eecs.umich.edu/~justincj/teaching/eecs498/FA2020/

Taxonomy of Generative Models



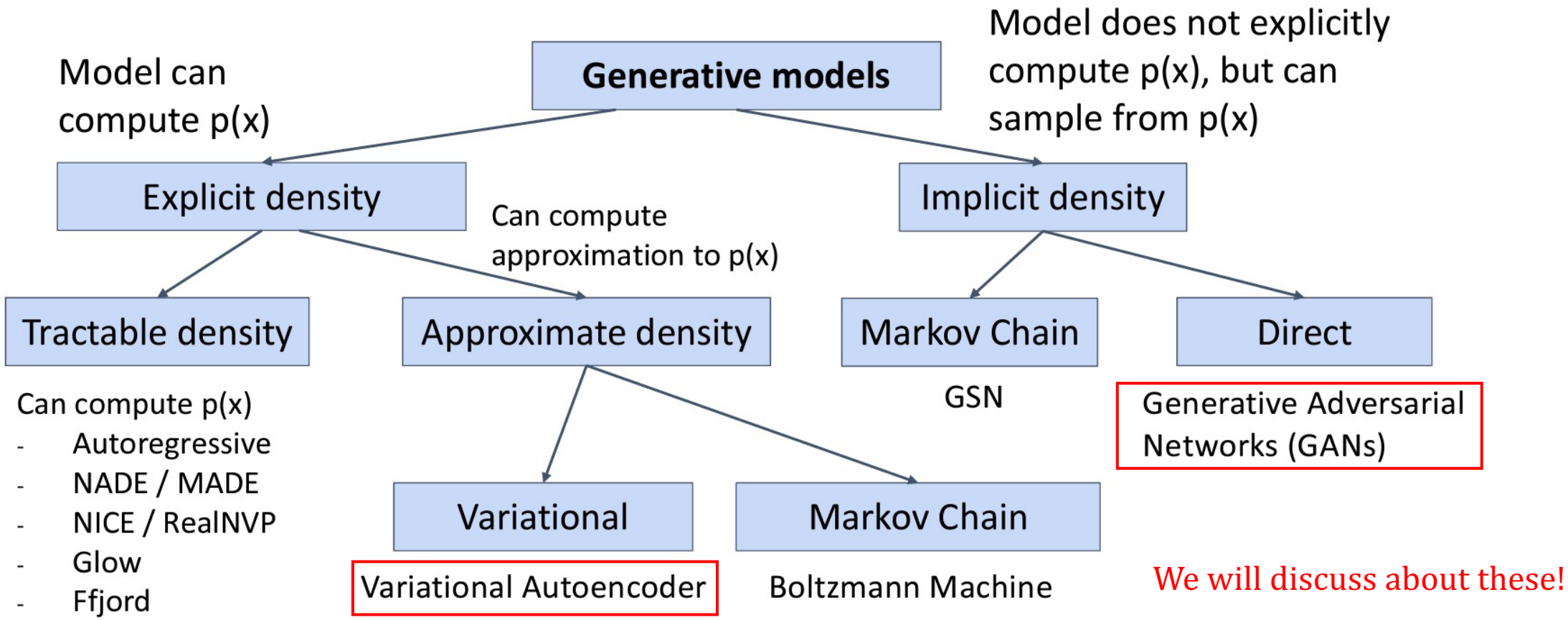
[picture ref] https://web.eecs.umich.edu/~justincj/teaching/eecs498/FA2020/

Taxonomy of Generative Models



[picture ref] https://web.eecs.umich.edu/~justincj/teaching/eecs498/FA2020/

Taxonomy of Generative Models

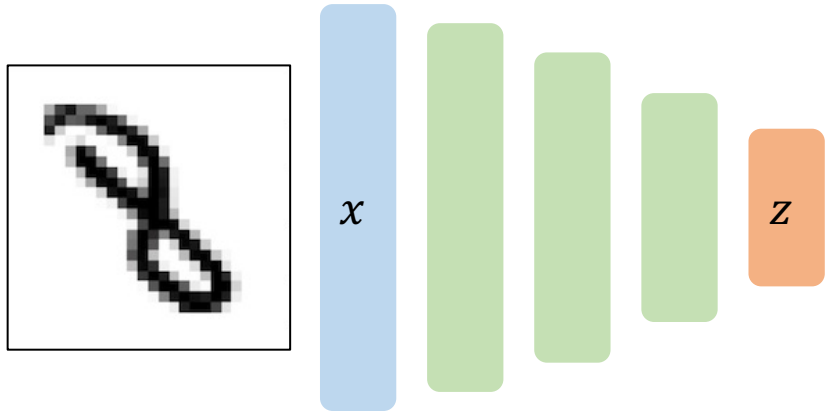


[picture ref] https://web.eecs.umich.edu/~justincj/teaching/eecs498/FA2020/

Variational Autoencoders

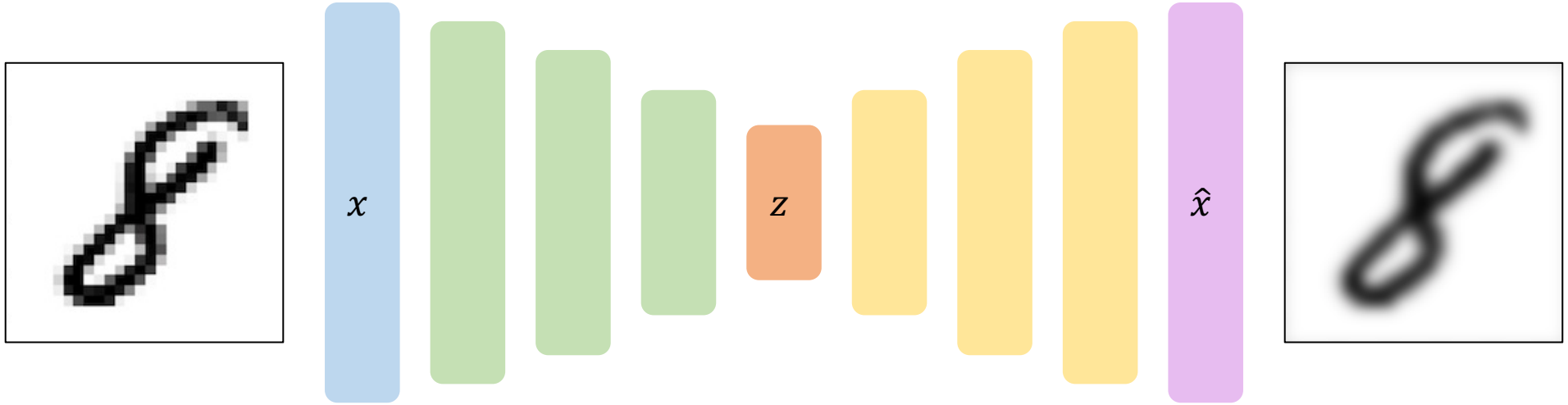
Autoencoders (AE)

- Consider an unsupervised task.
  - A number of images are available but without labels.
- Our task is to extract useful information from our data in an unsupervised manner.
  - Typically referred as latent representation learning.
- How can we learn this transformation (encoding) from our data?
  - Can we learn these in a supervised manner?



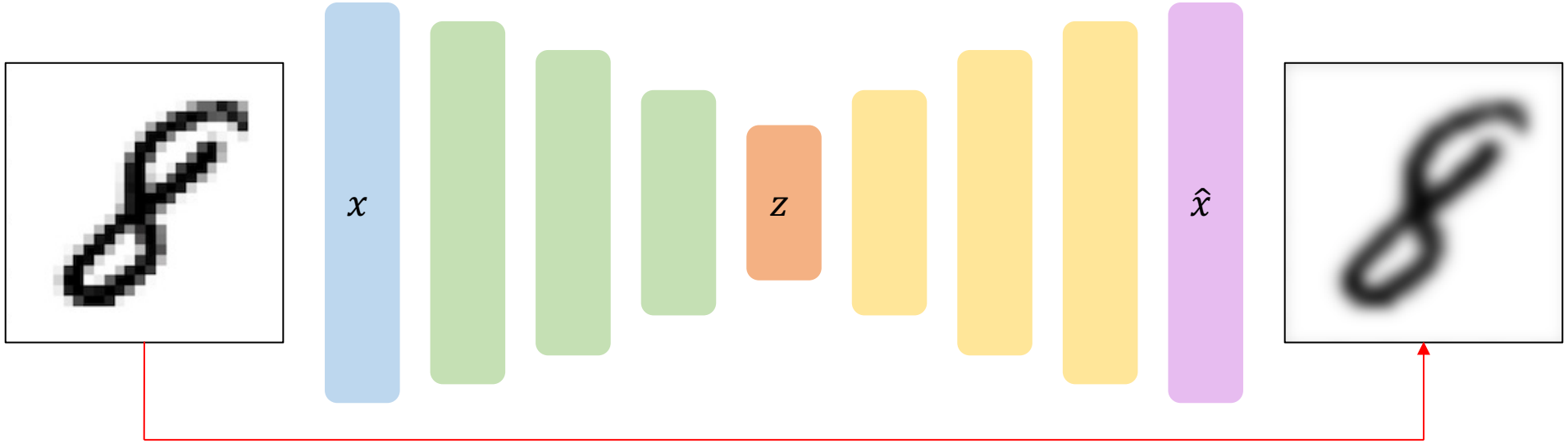
Autoencoders (AE)

- Idea: Use the latent representation to reconstruct the input data
  - Autoencoding = encoding itself.



Autoencoders (AE)

- Encoder: maps the observed data x to a latent representation z
- Decoder: reconstructs the observed data $\hat{x}$ using the latent representation z



$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

Dimensionality of Latent Space

- The smaller latent space it is the more details will be lost
  - High compression
- By increasing the latent space we will be able to keep more and more details.

2D Latent Space

5D Latent Space

Ground Truth

AE Summary

- Autoencoder architectures could be used to encode the observed data to some kind of compressed representation (latent vector)
- They utilize an encoder and a decoder part.
- The reconstruction loss is forcing the model to keep only the relevant information from the data
- The reconstruction loss generates the input using this compressed latent representation.
- This paradigm is somewhat close the the generative idea!

- Limitation: There is no constrain on the representation *z*:
  - Can we generate new data using the decoder?
  - Can we be sure that the different latent variables are disentangled and unique?
  - Is the representation z a probability density function?

Probability Density Function

- Properties:
  1. Non-negativity: $p_\theta(x) \geq 1, \forall x$
  2. Probabilities of all events must sum up to $1 : \sum_x p_\theta(x) = 1$

- Summing up to 1 means that:
  - Predictions improve relative to each other
  - Model cannot trivially get better scores by predicting higher values
  - Model forces to make non-trivial improvements

Building generative models for PDF estimation

- Normalize by the total volume of the function (partition):

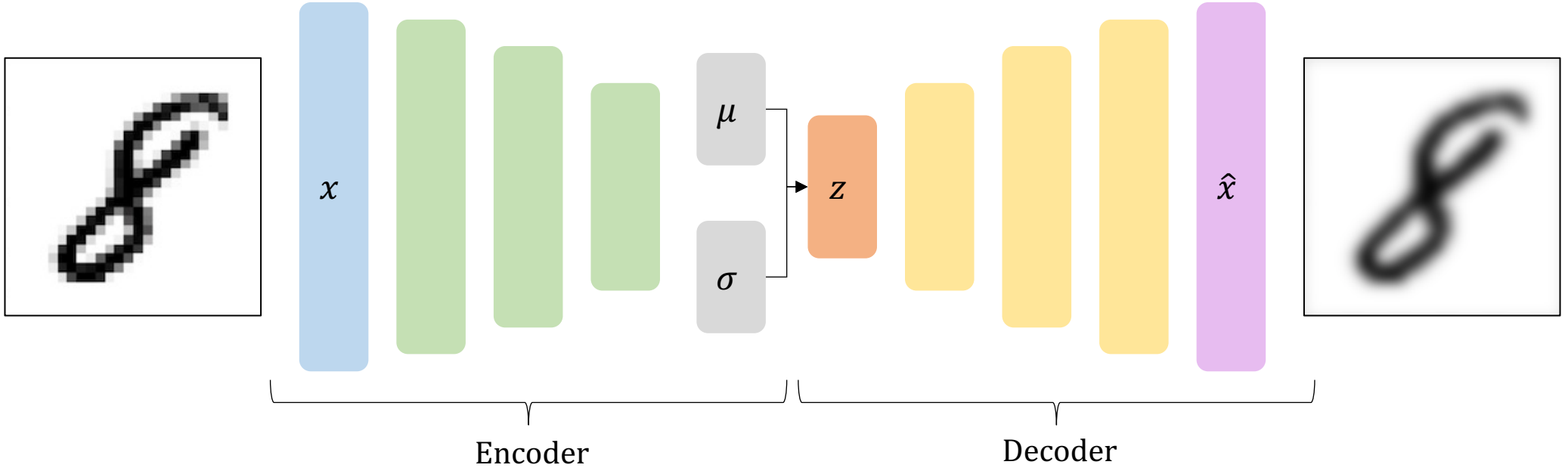
$$p_\theta(x) = \frac{1}{volume(g_\theta)} g_\theta(x) = \frac{1}{\sum_x g_\theta(x)} g_\theta(x)$$

  - In simple words, equivalent to normalizing [3,1,4] as $\frac{1}{[3+1+4]}$ [3,1,4]

- Examples of such functions:
  - Gaussian: $g_\theta(x) = g_{\{\mu,\sigma\}}(x) = e^{\frac{-(x-\mu)^2}{2\sigma^2}} \rightarrow volume(g_\theta) = \sqrt{2\pi\sigma^2}\ for\ x \in \mathbb{R}$
  - Exponential: $g_\theta(x) = g_{\{\lambda\}}(x) = e^{-\lambda x} \rightarrow volume(g_\theta) = \frac{1}{\lambda},\ for\ x \geq 0$

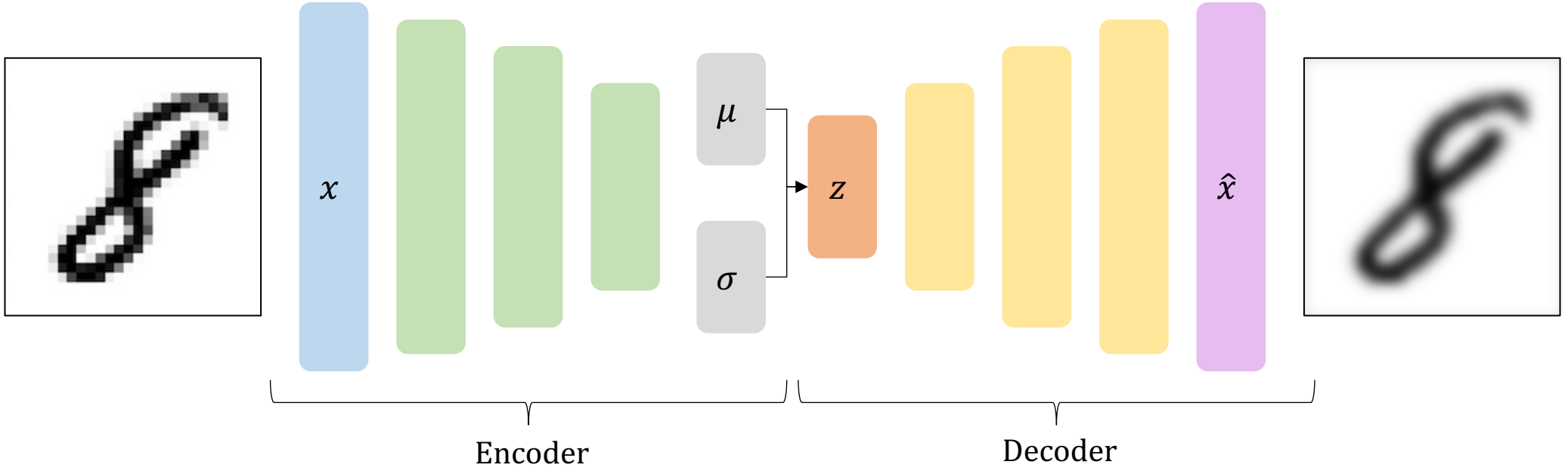
Variational Autoencoders (VAE)

- Instead of straight learning the latent representation z
- We learn the parameters of a multivariate gaussian from which we sample z
- Not deterministic any more → Stochastic sampling operation



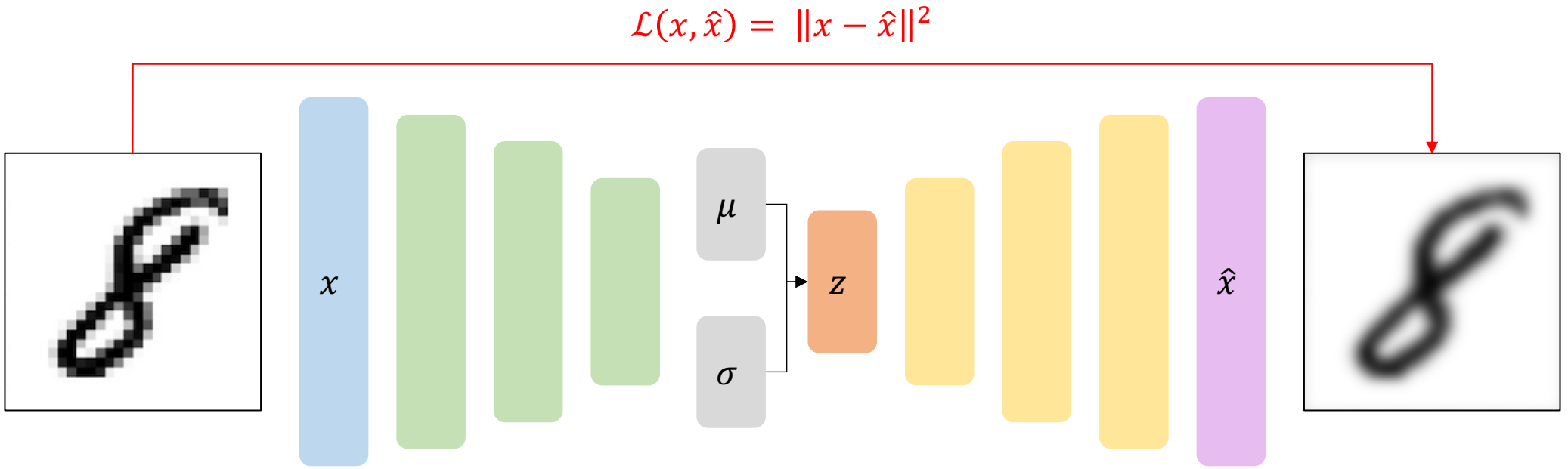
Variational Autoencoders (VAE)

- Probabilistic twist:
  - Encoder learns $p(z|x)$
  - Decoder learns $p(x|z)$ Generative Model!



Variational Autoencoders (VAE)

- Reconstruction loss
  - Mean Square Error

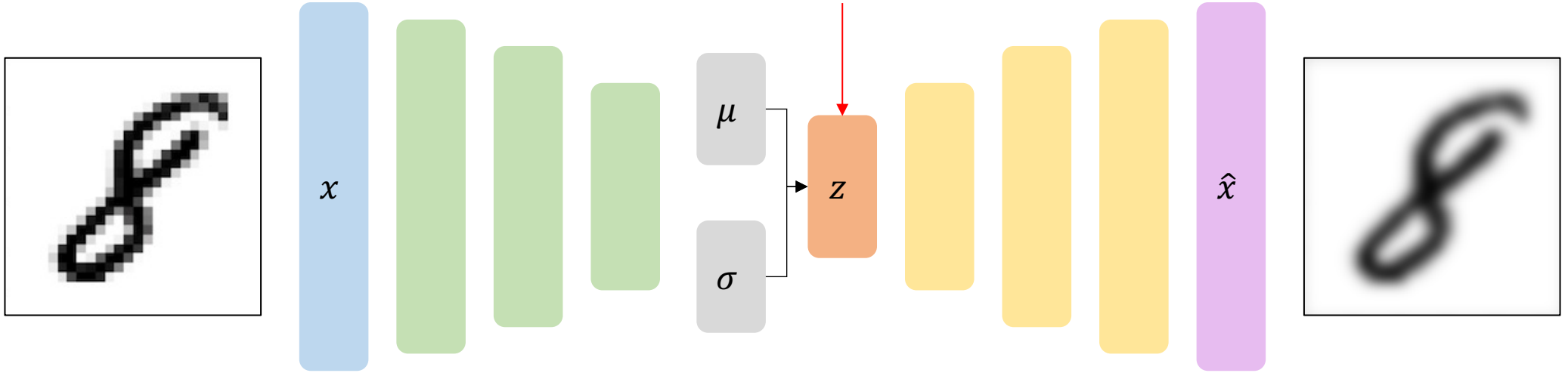


$$\mathcal{L}(x, \hat{x}) = \boxed{(reconstruction\ loss)} + (regularization\ term)$$

Variational Autoencoders (VAE)

- Regularization Term
  - KL divergence between the inferred latent distribution and a prior distribution.
  - Typically zero mean unit std normal distribution

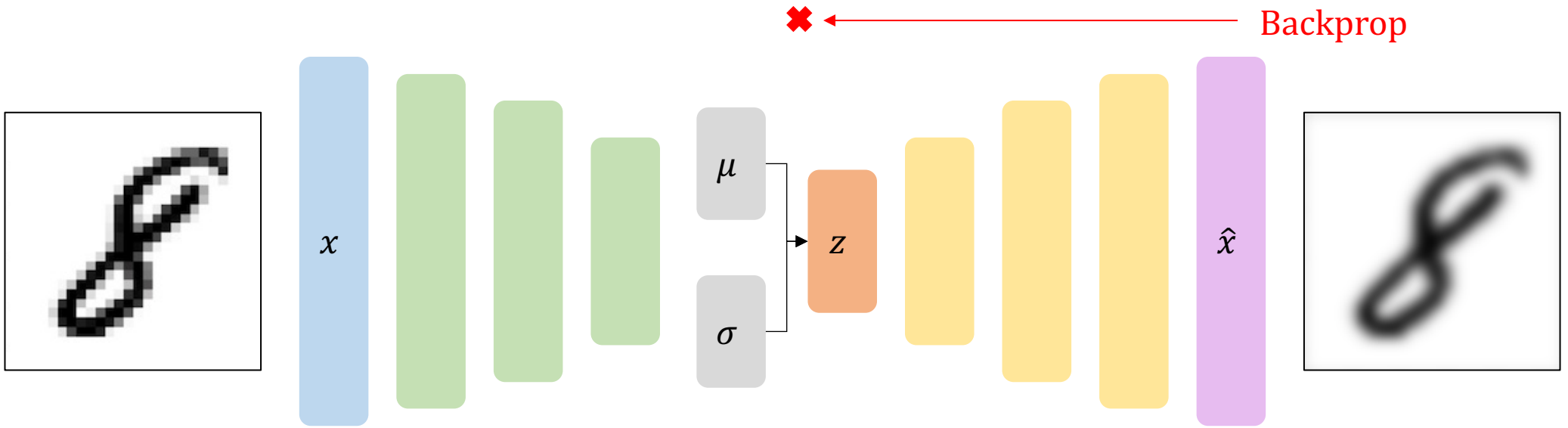
$$\mathrm{R}(z) = D_{KL}(p(z|x) \parallel \mathcal{N}(\mu = 0, \sigma = 1))$$



$$\mathcal{L}(x, \hat{x}) = (reconstruction\ loss) + \boxed{(regularization\ term)}$$

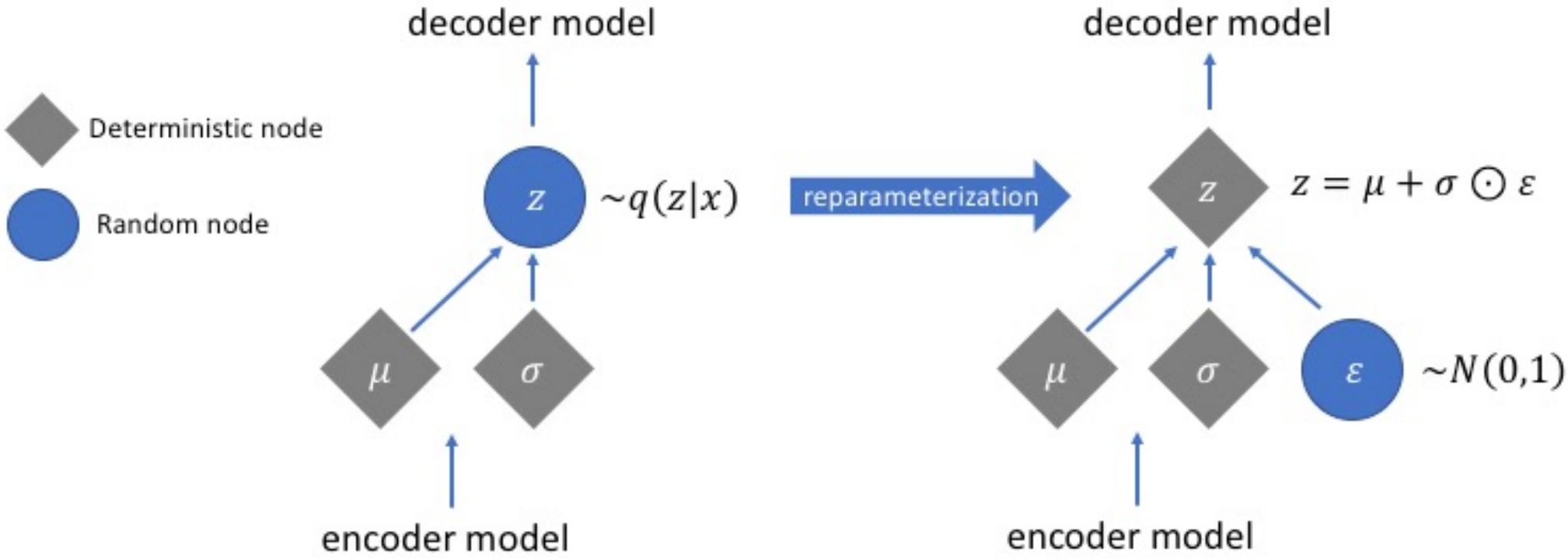
Computational Graph

- Sampling operation is not differentiable
  - How can we solve this?



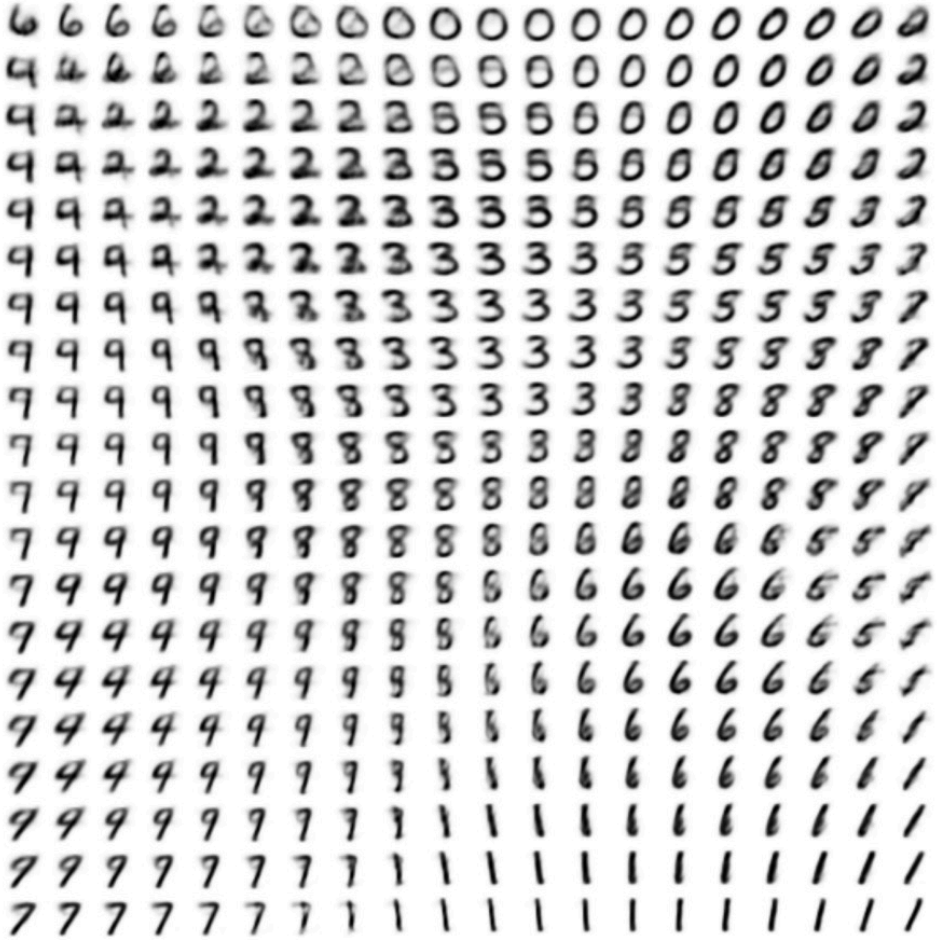
$$\mathcal{L}(x, \hat{x}) = (reconstruction\ loss) + (regularization\ term)$$

Reparameterization Trick



Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

VAE: Latent Perturbation



Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

VAE: Latent Perturbation

[http://dpkingma.com/sgvb_mnist_demo/demo.html]

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

VAE: Summary

- Compresses observed data to some smaller representation
- Unsupervised setting
- Reparameterization trick for end-to-end training
- Force latent representation to imitation a gaussian distribution (KL divergence)
- The latent variables can be interpreted by perturbating their values
- Can be used to generate new samples

Take Home Messages

- We discussed about two groups of learning based on the availability of labels (supervised, unsupervised).
- We discussed about two groups of models based on their probabilistic properties (discriminative, generative).
- Discriminative models learn the conditional probability and are used to performed pre-defined tasks (e.g. classification).
- Generative models model the underlying probability distribution of the data and can be used to generate new un-seen data.
- Variational auto-encoders (a variation of autoencoders) is an example of a generative model in the context of deep learning.