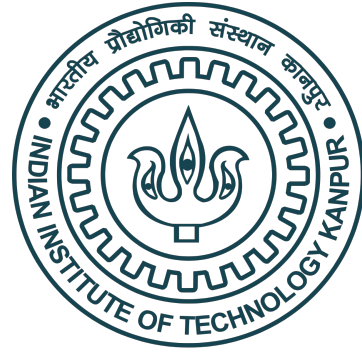



PHY654

Machine learning (ML) in particle physics



Swagata Mukherjee • IIT Kanpur
30th September 2024

Data science competition platform


 [Competitions](#) [Datasets](#) [Models](#) [Code](#) [Discussions](#) [Courses](#) [...](#)

[Sign In](#) [Register](#)

Level up with the largest AI & ML community


Join over 20M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

[Register with Google](#) [Register with Email](#)




Who's on Kaggle?


Learners
Dive into Kaggle courses, competitions & forums.



Developers
Leverage Kaggle's models, notebooks & datasets.



Researchers
Advance ML with our pre-trained model hub & competitions.



Tackle your next project with Kaggle

On Kaggle you'll find all the resources and knowledge needed for your next real-world ML project.

382K
DATASETS

1.2M
NOTEBOOKS

8,300
MODELS

Datasets in kaggle can be useful for practice

<https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results>

Kepler Exoplanet Search Results

10000 exoplanet candidates examined by the Kepler Space Observatory

<https://www.kaggle.com/datasets/nasa/meteorite-landings>

Meteorite Landings

Data on over 45k meteorites that have struck Earth

<https://www.kaggle.com/datasets/roche-data-science-coalition/uncover>

UNCOVER COVID-19 Challenge

United Network for COVID Data Exploration and Research

<https://www.kaggle.com/datasets/fedesoriano/multijet-primary-dataset>

CERN Proton Collision Dataset

Particle collision events from the MultiJet primary dataset from CMS open data.

If you sign up, you will get notified about new competitions.
Not all of them will be aligned to your interest, but some will.

Hi

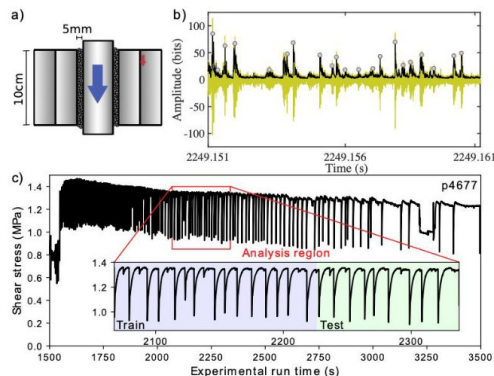
Excessive internet use among children is an increasing concern in today's digital world. Screen time often takes a big 'byte' out of the day, leaving little time for physical activity and affecting overall health and development.

In this competition sponsored by Dell Technologies and NVIDIA, you'll develop a model that leverages physical activity and fitness data to identify early indicators of problematic technology use in children and adolescents.

Earthquake Prediction

LANL / Penn State / Purdue

Model laboratory earthquake data



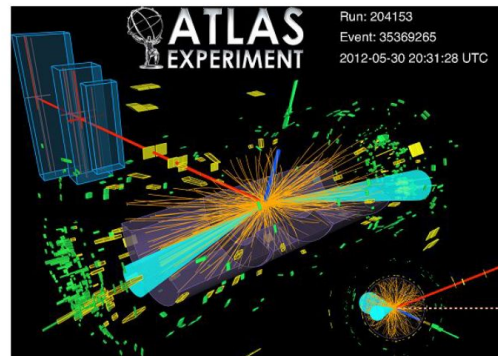
<https://arxiv.org/pdf/1810.11539.pdf>

Higgs Boson Machine Learning Challenge

CERN

Use the ATLAS experiment to identify the Higgs boson, predicting tau decay of a Higgs boson vs background

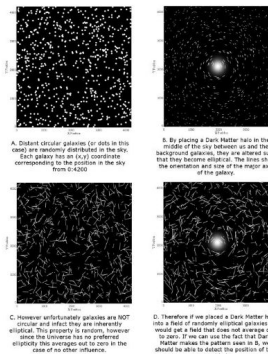
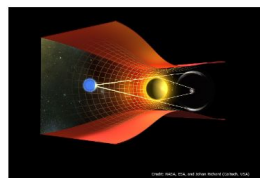
<https://www.kaggle.com/c/higgs-boson/overview>



Dark Worlds - Gravitational Lensing

Winton Capital

Predicting dark matter
Top solutions used Bayesian methods



<https://www.kaggle.com/competitions/DarkWorlds>

What do you see in this image?

Do you see anything unusual?



Generative model (unsupervised ML)

<https://thispersondoesnotexist.com/>

thispersondoesnotexist.com



Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)

x is data, y is label

Goal: Learn a *function* to map $x \rightarrow y$

Unsupervised Learning

Data: x

Just data, no labels!

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

What is Generative model?

Given training data, generate new samples from same distribution



Training data $\sim p_{\text{data}}(x)$



Generated samples $\sim p_{\text{model}}(x)$

Want to learn $p_{\text{model}}(x)$ similar to $p_{\text{data}}(x)$

What is Generative model?

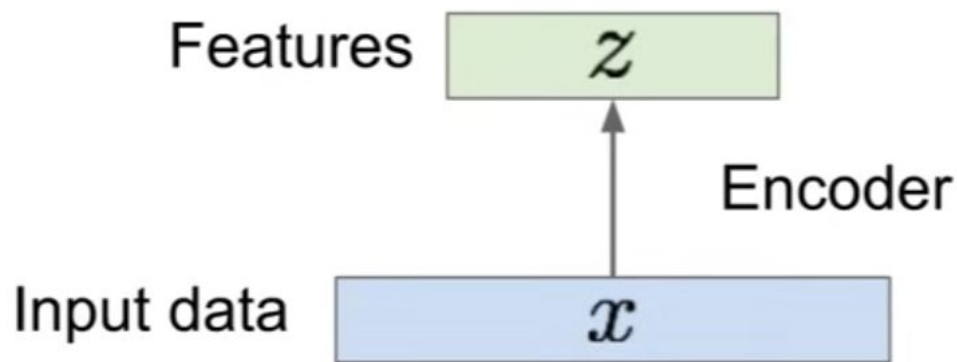
A model that can produce high quality content (example text, images and audio).

Example text-generators: ChatGPT/OpenAI, Bard/Google, Bing Chat/Microsoft

Use-case in particle physics: Simulation

Generative model 1 : Autoencoder

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



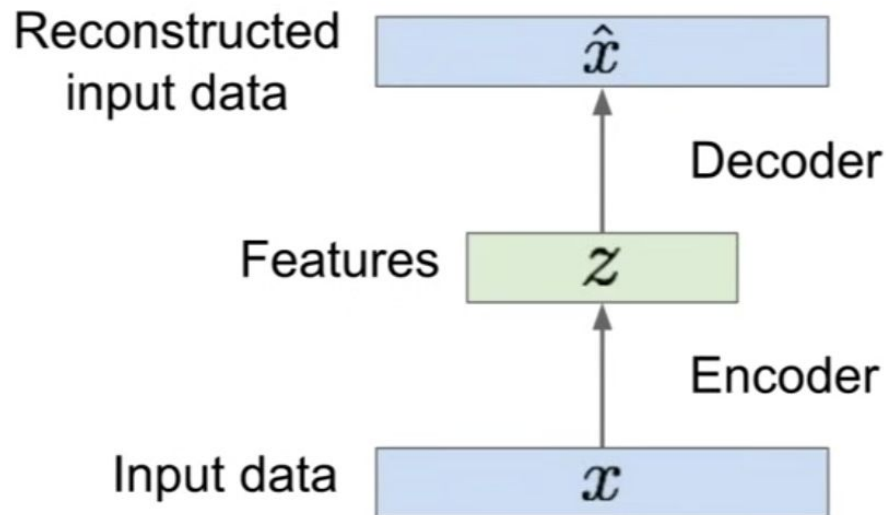
z is usually smaller than x
(dimensionality reduction)

z should represent the most important features in x .

Data compression

Autoencoder

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

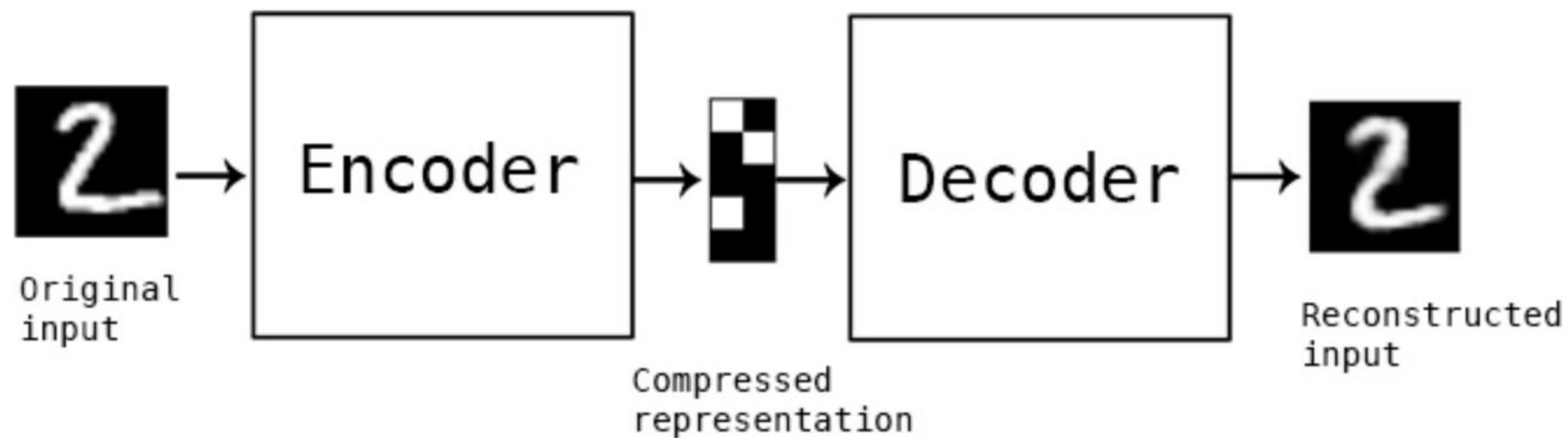


For encoder and decoder

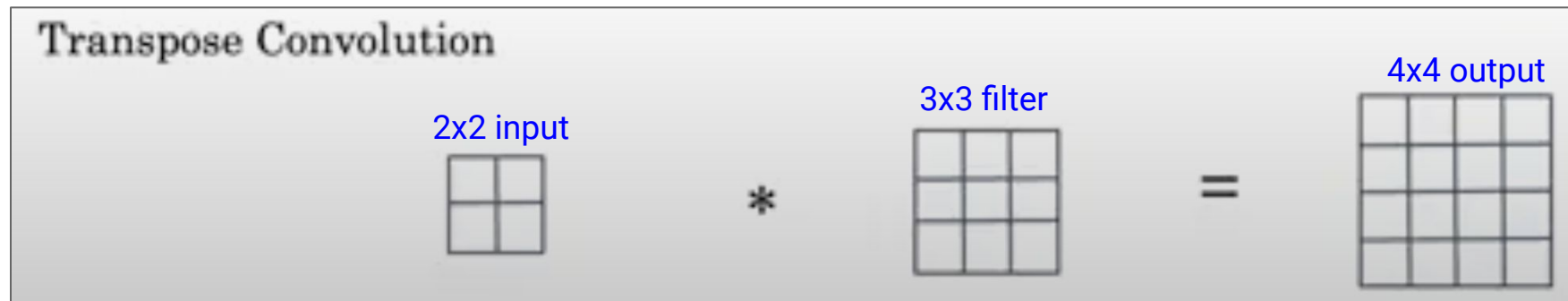
Originally: Linear + nonlinearity (sigmoid)

Later: Deep, fully-connected

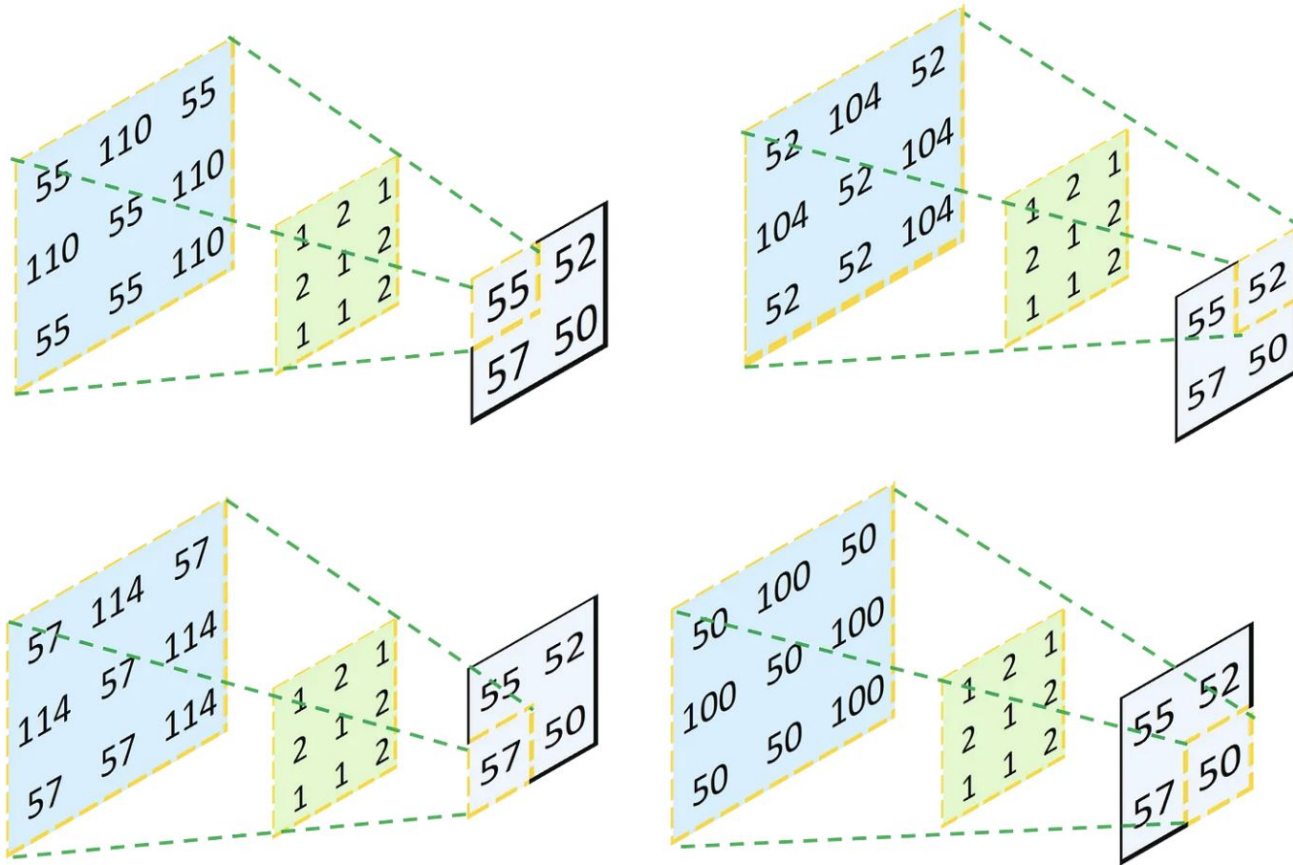
Later: ReLU CNN (upconv)



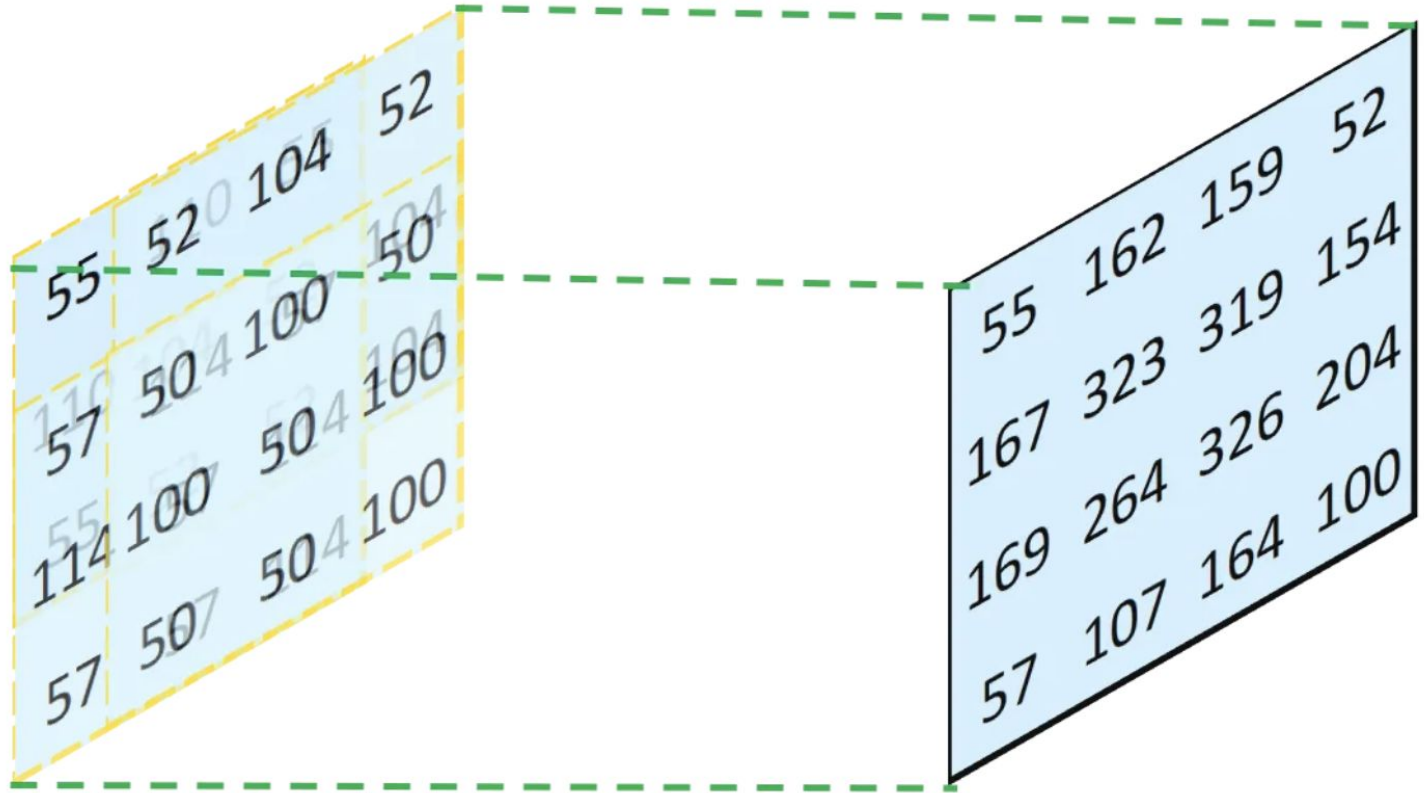
Transpose convolution



Output size depends on choice of stride, and padding



Multiply Each Element in the Input Layer by Each Value in the Kernel



Combine All Four Resulting Layers Together And Sum the Overlapped Values (Image by Author)

Reference:

<https://www.jeremyjordan.me/variational-autoencoders/>

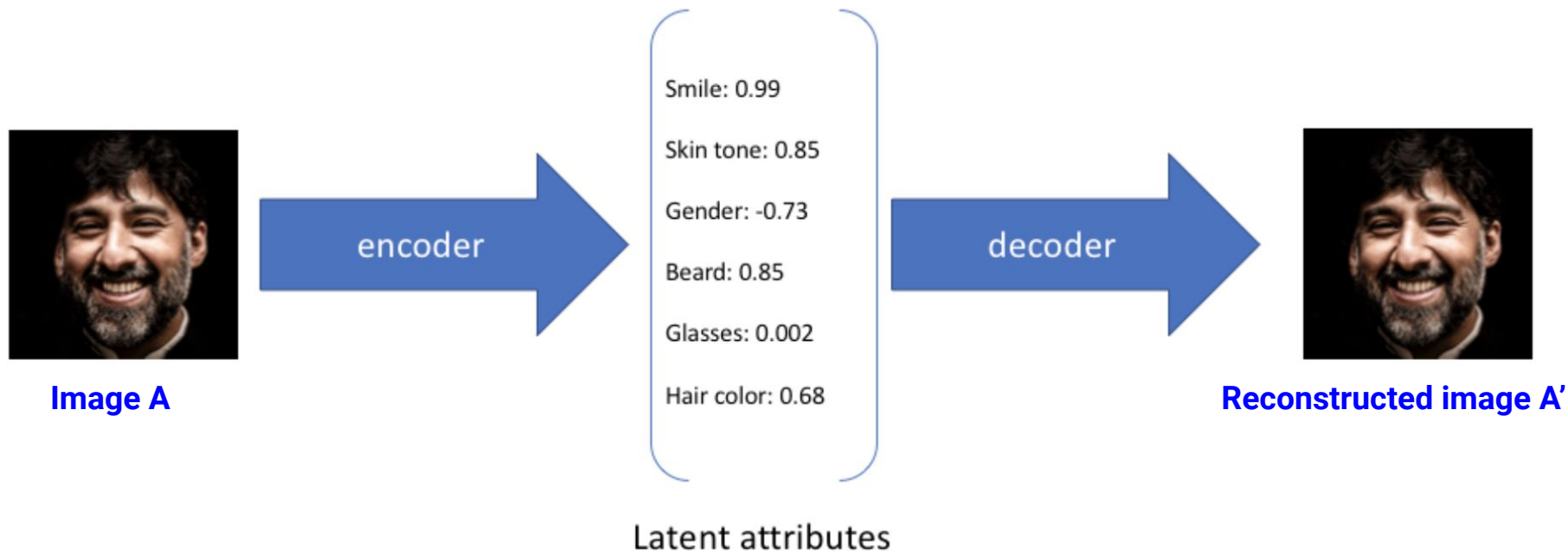
Information of the image is **encoded** in this vector (V)

This can be used in image search

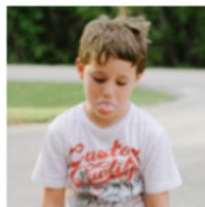
Image search

Two images of same person: $V_1 \approx V_2$

Two images of different persons: $V_1 \neq V_2$



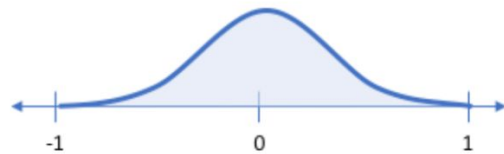
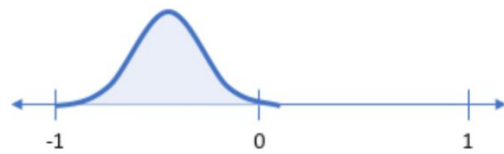
We should try to minimize the difference between A and A'



Smile (discrete value)



Smile (probability distribution)



VS.

