Gen Hack 2022

Generative Modelling Challenge

1st session tutorial

2022, November 10th

MERCATOR OCEAN INTERNATIONAL

BNP PARIBAS
The bank for a changing world

FONDATION ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE
IP PARIS

Objectives of generative modelling

Data augmentation, data generation, data completion:

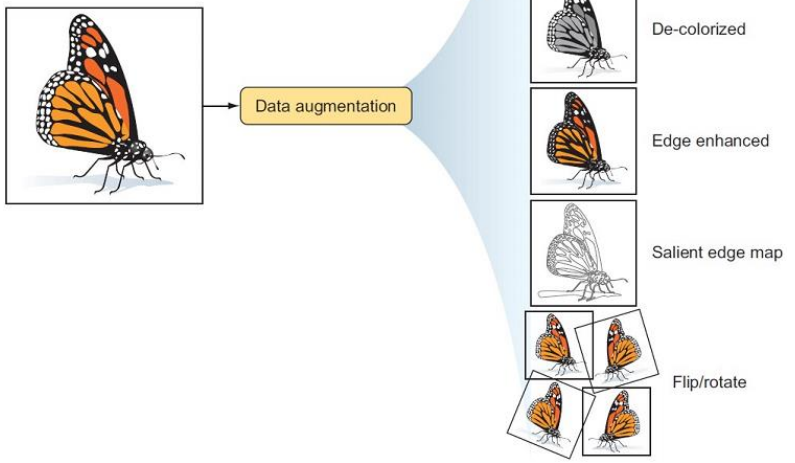
- simulating new **(statistically realistic)** data from a data set
- for risk analysis, decision making, data sharing, …

Unsupervised learning:
discovering the hidden patterns in order to get
a meaningful representation of the data

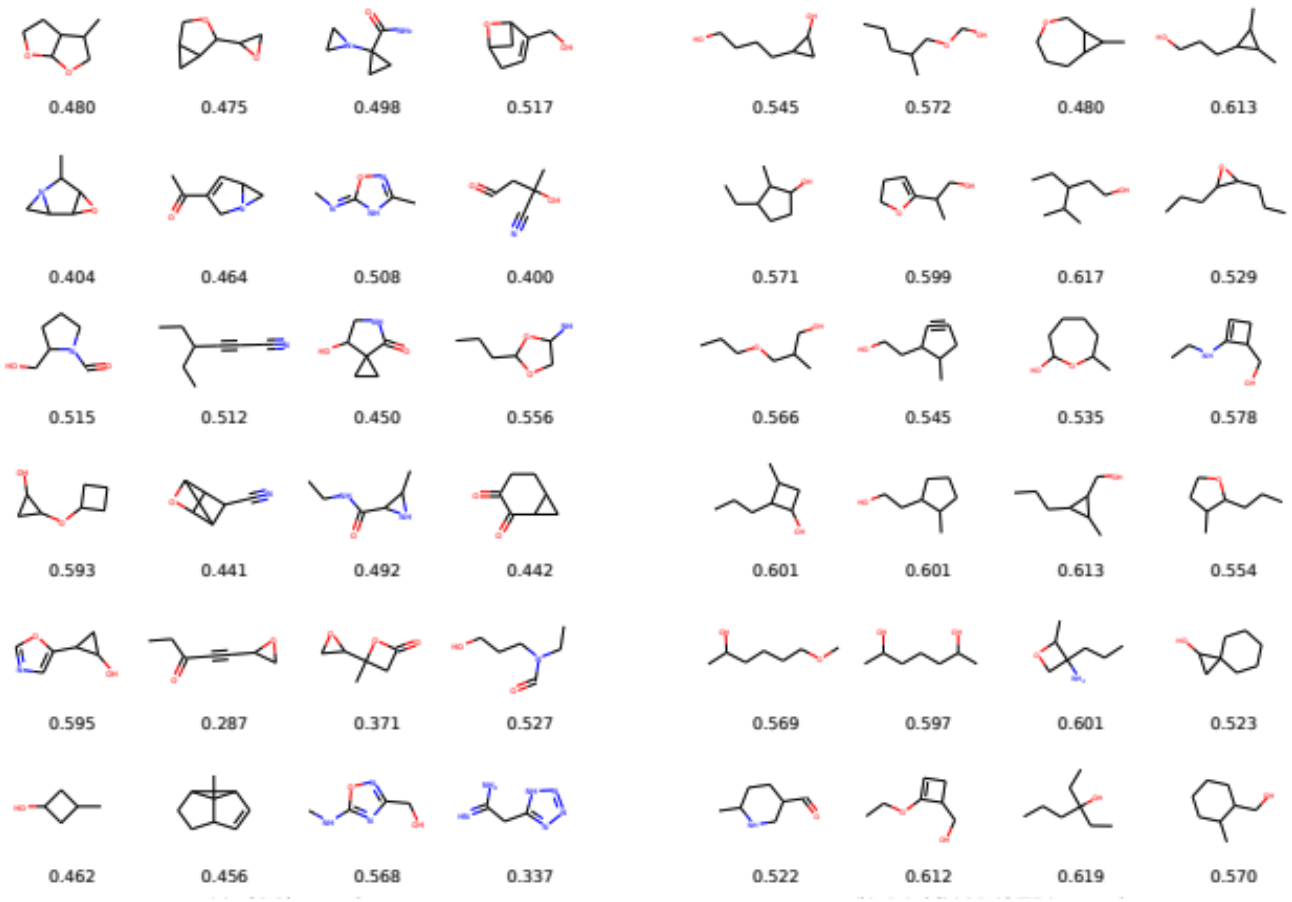
Density estimation is not required

- > see later

More than classical data augmentation

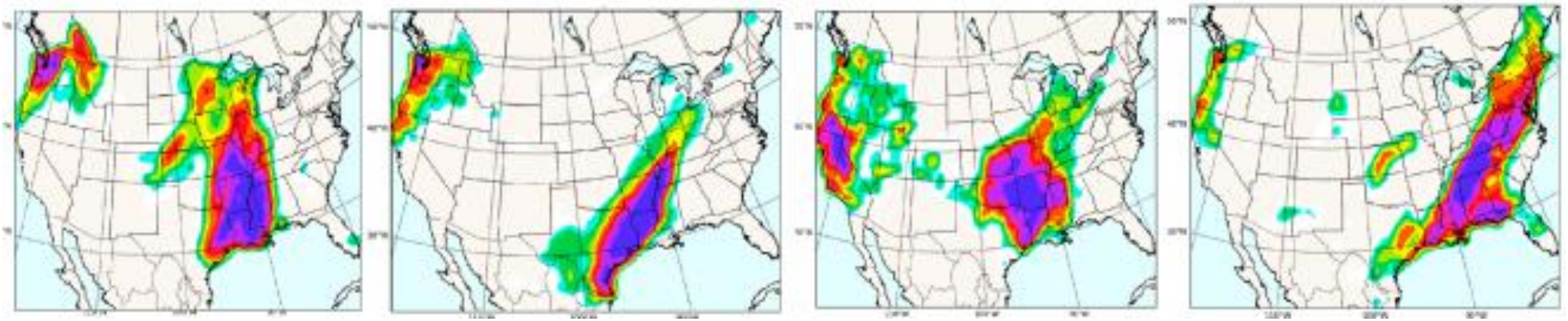


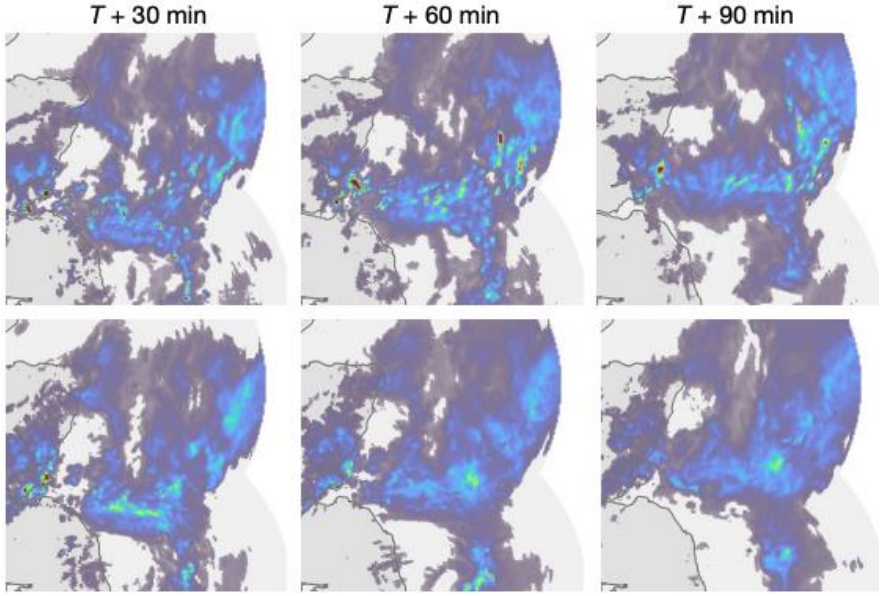
Drug design



Credits: https://arxiv.org/pdf/1805.11973.pdf

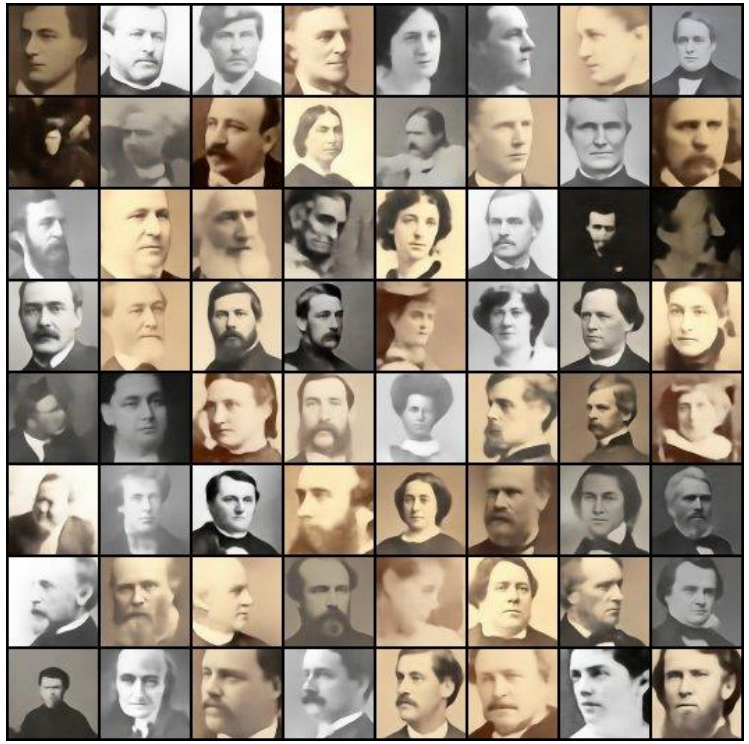
Meteorological Modeling



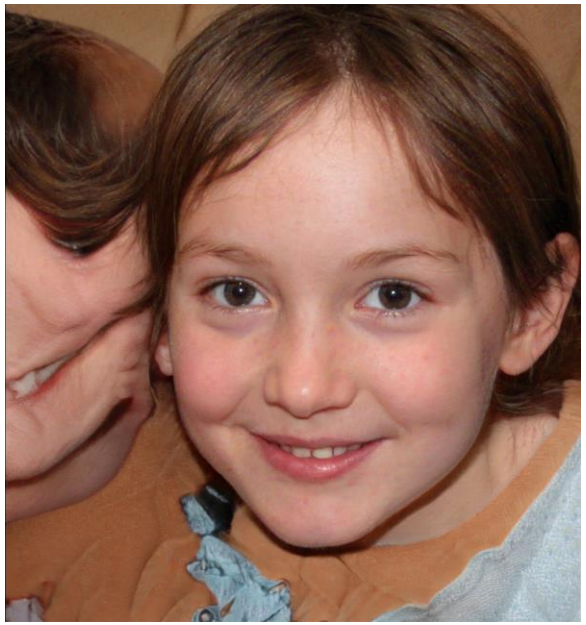


Credits: https://www.nature.com/articles/s41586-021-03854-z.pdf
https://arxiv.org/pdf/2009.08454.pdf

High-resolution image synthesis

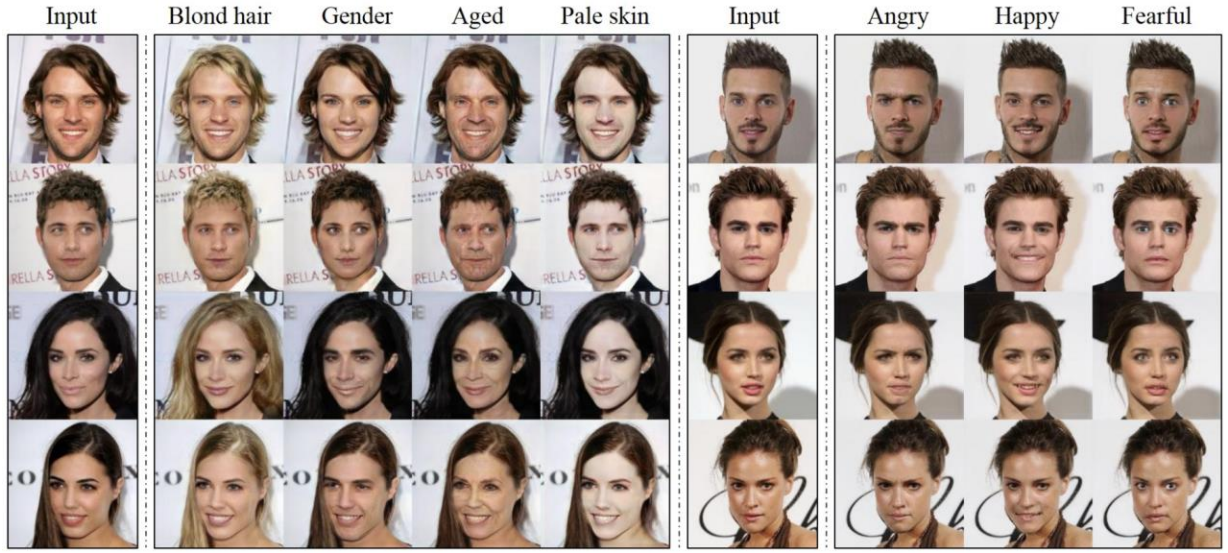


Credits: https://twitter.com/pleonard



Credits: www.thispersondoesnotexist.com

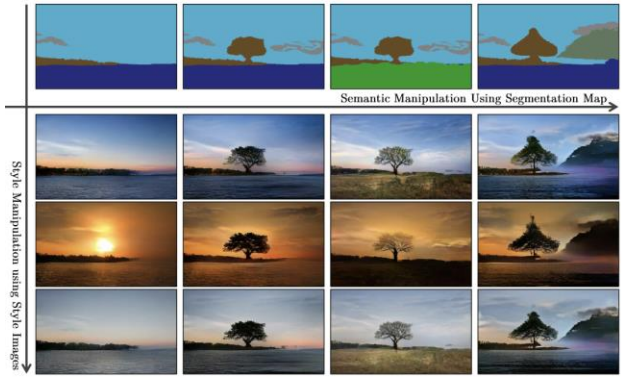
Conditional generation



Credits: https://arxiv.org/pdf/2003.12267.pdf

https://arxiv.org/pdf/1711.09020.pdf

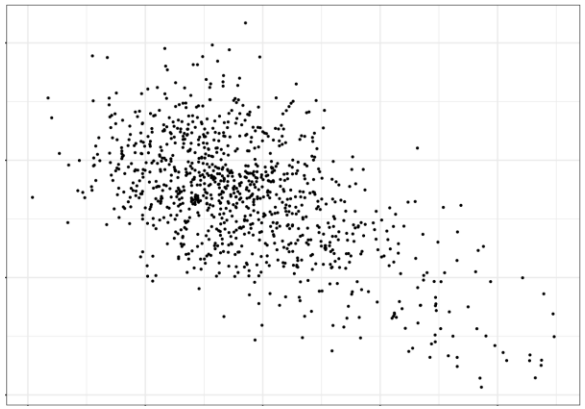
https://arxiv.org/pdf/1903.07291.pdf



Deep fashion

Naive approach and issues

Data set for training



Two Density-based generation:

- parametric model
- non-parametric model

Parametric estimation:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- choice of the parametric density?
- non convex problems?

-> perfect if we know the statistical model family
-> usually too restrictive for complex data

Non-parametric estimation:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X(i)}{h}\right).$$

- choice of the kernel?
- choice of the width?

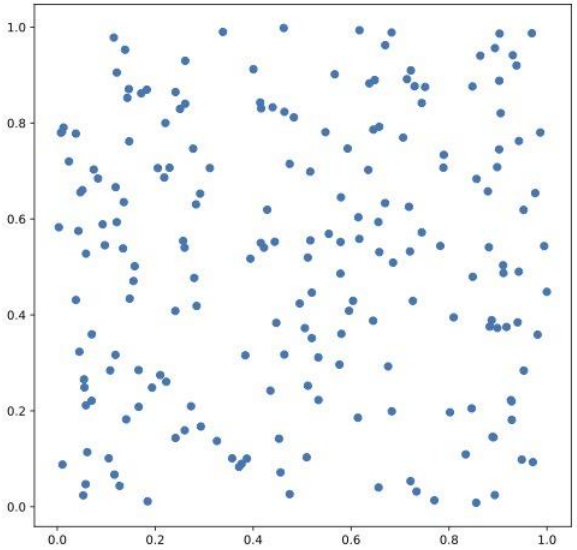
-> bad convergence properties when dimension increases (curse of dimensionality)

Monte Carlo sampling:

- acceptance/rejection : which proposal? Possibly not efficient
- MCMC approach

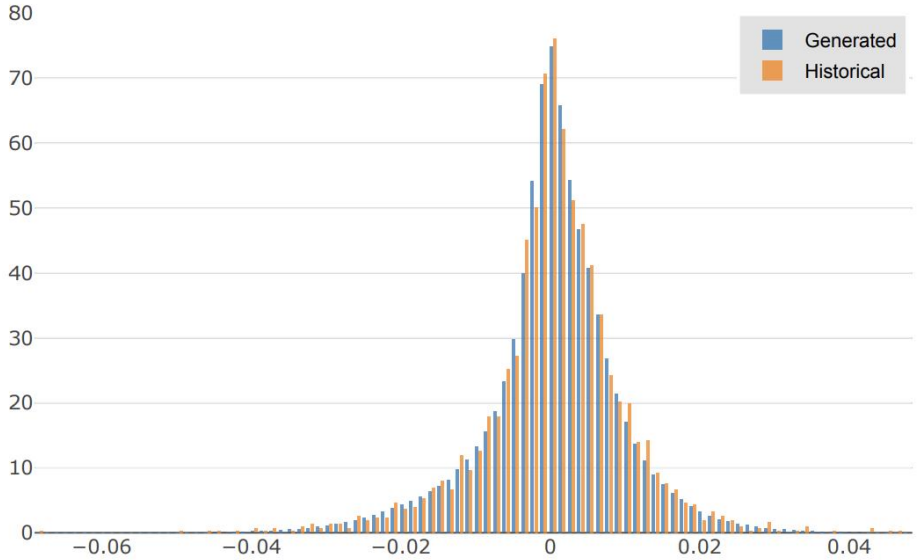
Maybe density does not exist (data in manifold)

Other approaches avoiding density estimation



Noise

Generator



Data « histograms »

Questions:

- which type of noise?
- how to build and to learn the Generator?
- which distance, or divergence, or loss function with sampled data and training data?