Modelling Smart Meter Data

Why and How?

Kutay Bölat
TU Delft

Planning

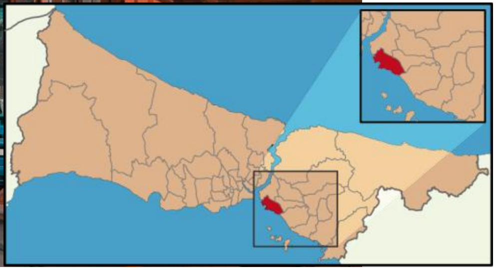
- Understand the role of smart meters in modern energy systems.
- Understand what is generative data modelling.
- Explore the challenges associated with smart meter data modelling.

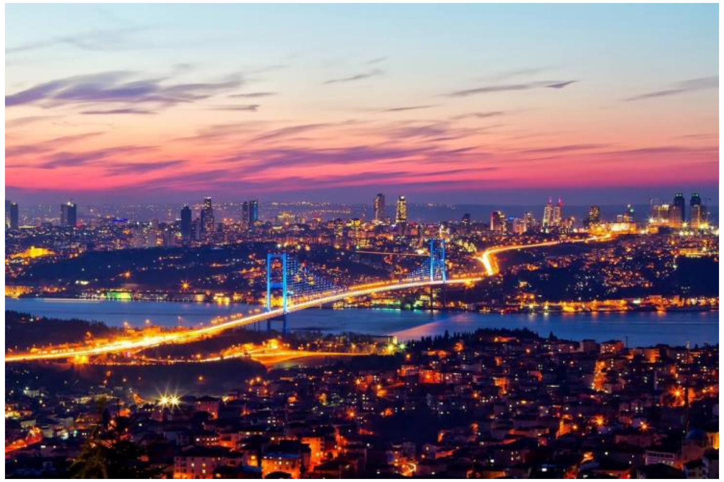
Part I – 11:15 – 12:15

Part II – 13:45 – 14:30

- Learn about data modelling techniques with a focus on Variational Autoencoders (VAEs).
- Hands-on coding session for practical understanding (afternoon session).

Who is this guy?



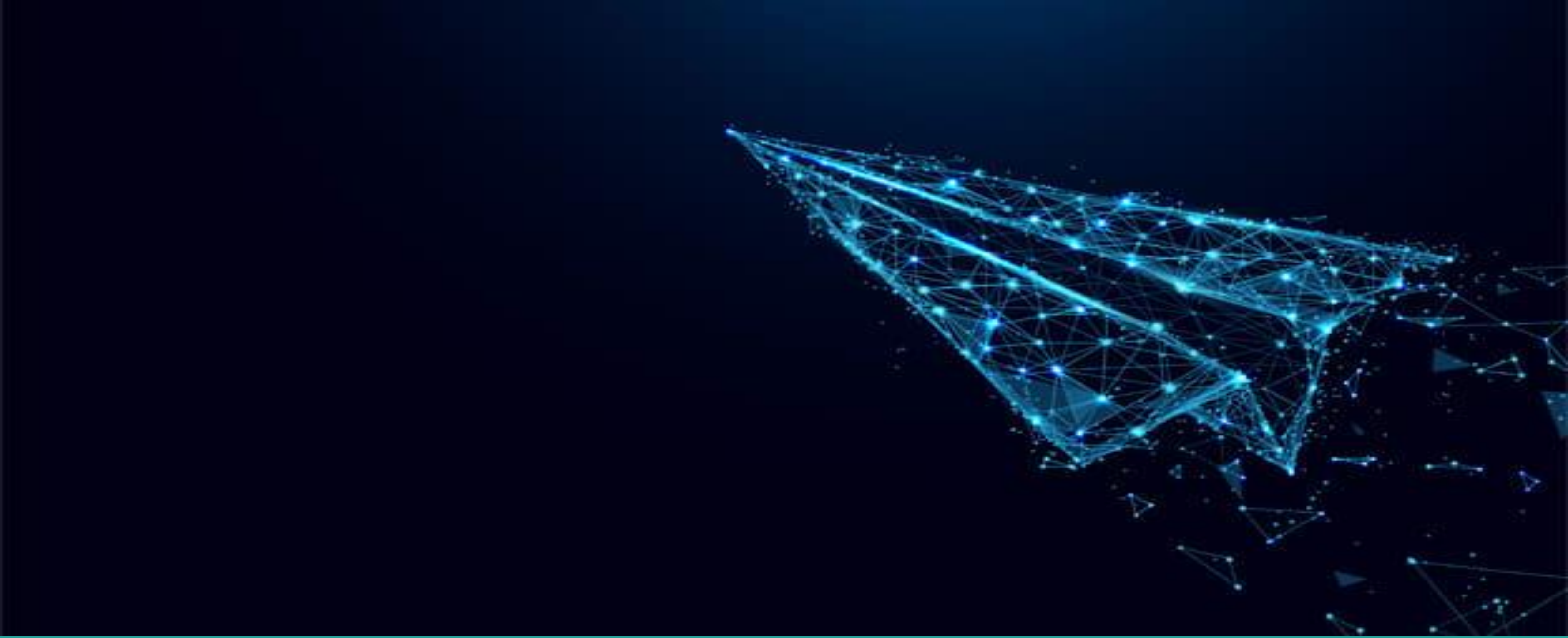


Who is this guy?

Data science domination

- (BSc) Electronics & **Communication** Engineering
  - Modulation Schemes for 6G
- (BSc) **Control** & Automation Engineering
  - Wireless Localization with Deep Learning
- (MSc) **Control** & Automation Engineering
  - Interpretable AI with autoencoders + fuzzy logic
- (PhD) Electrical Sustainable Energy
  - Synthetic smart meter data generation

Contents

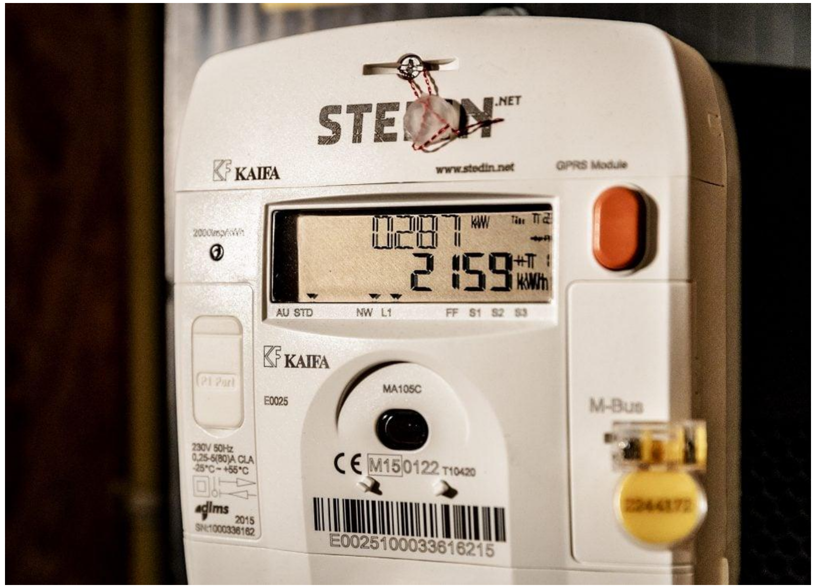


Smart meters

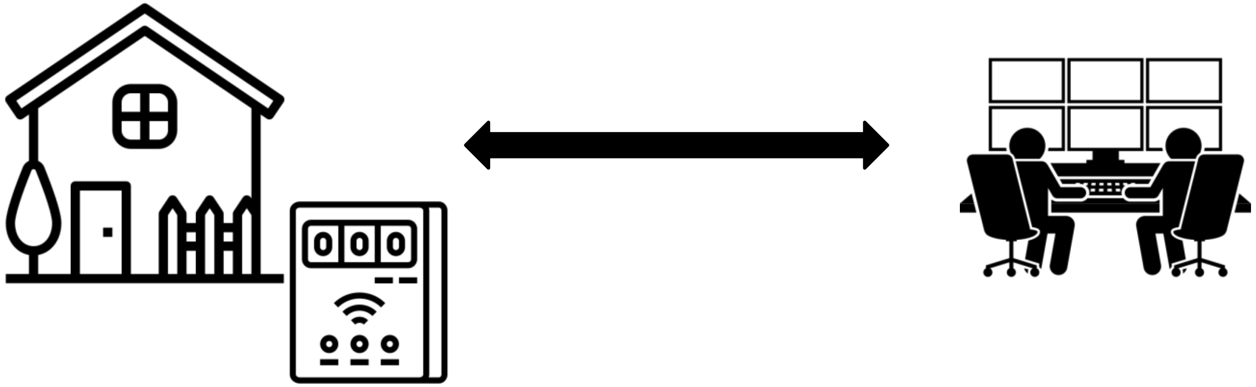
InnoCyPES

Smart Meters – What are they?

- Advanced meters that record energy prosumption* in (near) real-time and communicate the information to/from the utility company.



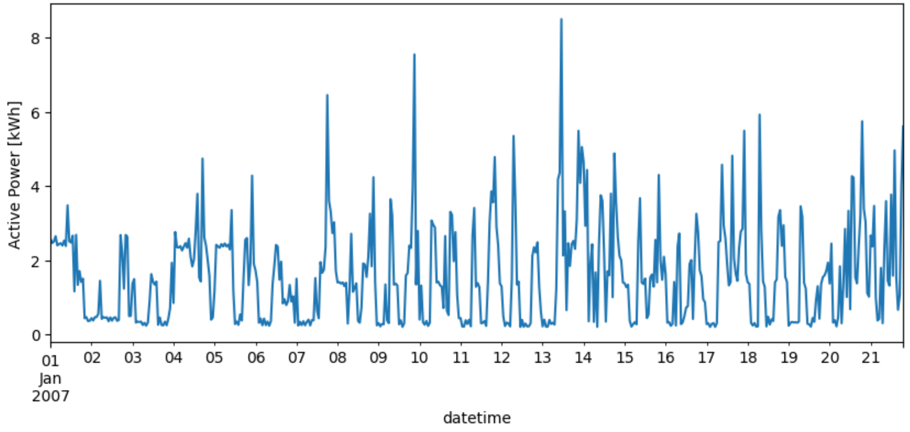
www.businessinsider.nl/slimme-meter-liander-stedin-cdma/

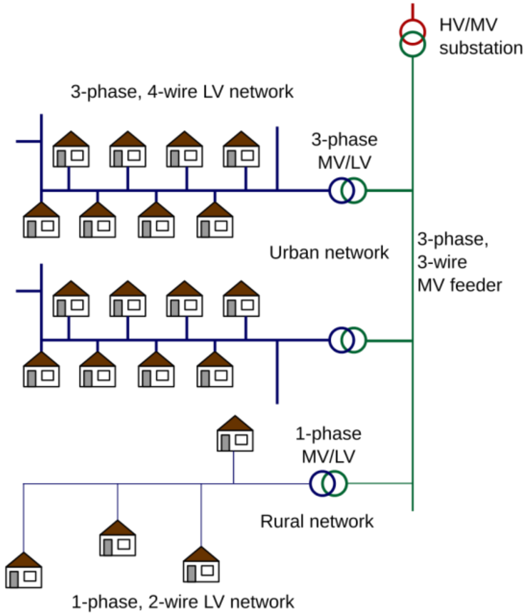


*production and consumption

Smart Meters – Why are they important?

- High resolution in time
- High resolution in space
- Already installed



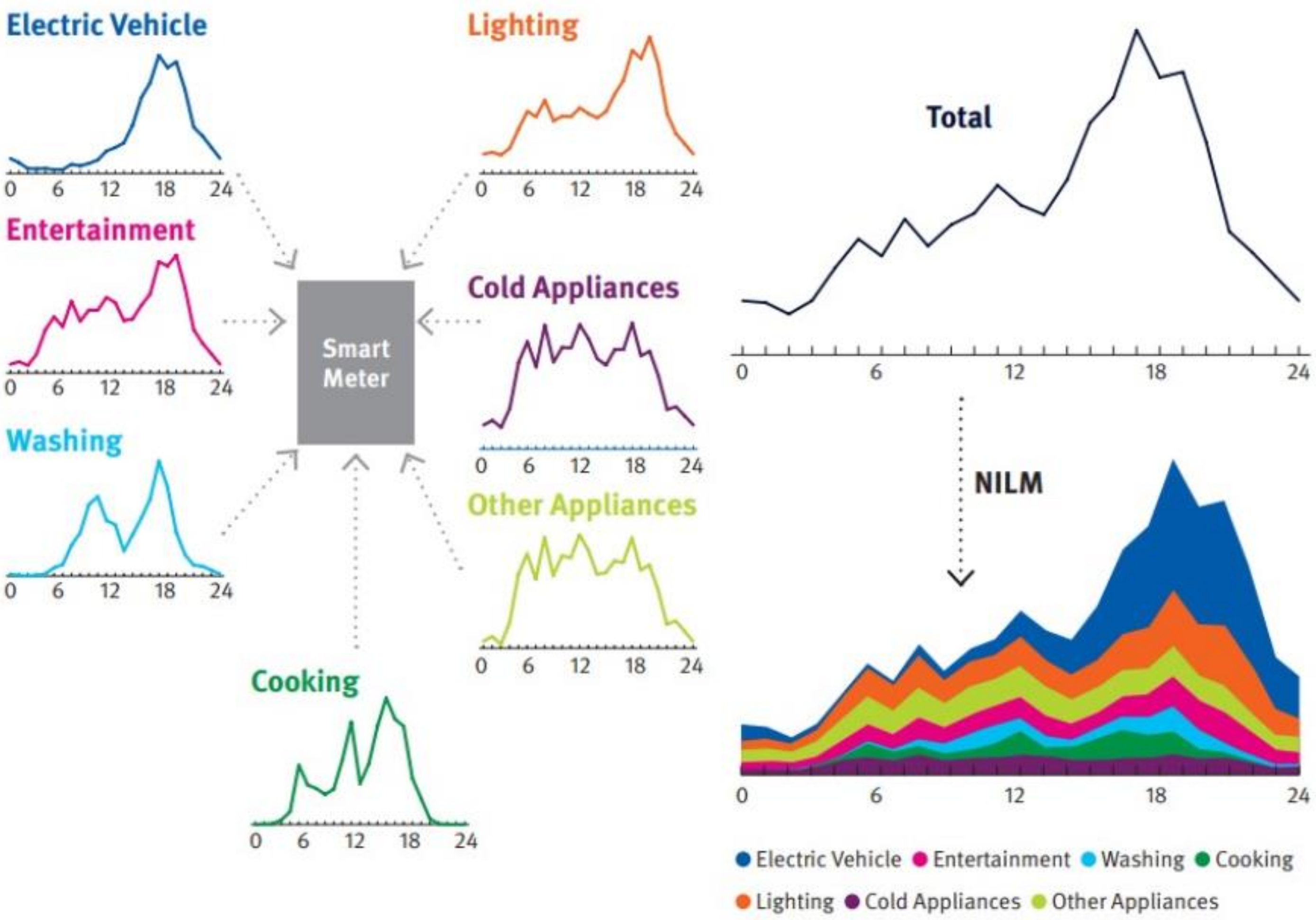


Smart Meters - Applications

- Grid Management
  - Enhances load balancing and stability of the power grid
  - Facilitates real-time monitoring and fault detection
- Tariff Design
  - Allows for dynamic pricing models based on usage patterns
  - Encourages energy saving during peak hours
- Demand Side Management
  - Helps in designing energy conservation programs
  - Provides consumers with feedback on energy usage
- Other Applications
  - Renewable energy integration
  - Electric vehicle (EV) charging optimization
  - …

Smart Meters - Privacy

- Non-Intrusive Load Monitoring (NILM)
  - NILM techniques can infer appliance-level usage from aggregate smart meter data
  - Example: Identifying when a person is at home or what applian[c]e they are using



Smart Meters - Privacy

- Privacy Risks
  - Potential to reveal personal habits and routines
  - Risk of unauthorized access to sensitive household information

- Data Accessibility Issues
  - Data protection laws restrict the availability of detailed smart meter data.
  - Utility companies face challenges in sharing data for research and development.

- Impact on Research
  - Limited access slows down innovation in smart energy solutions.

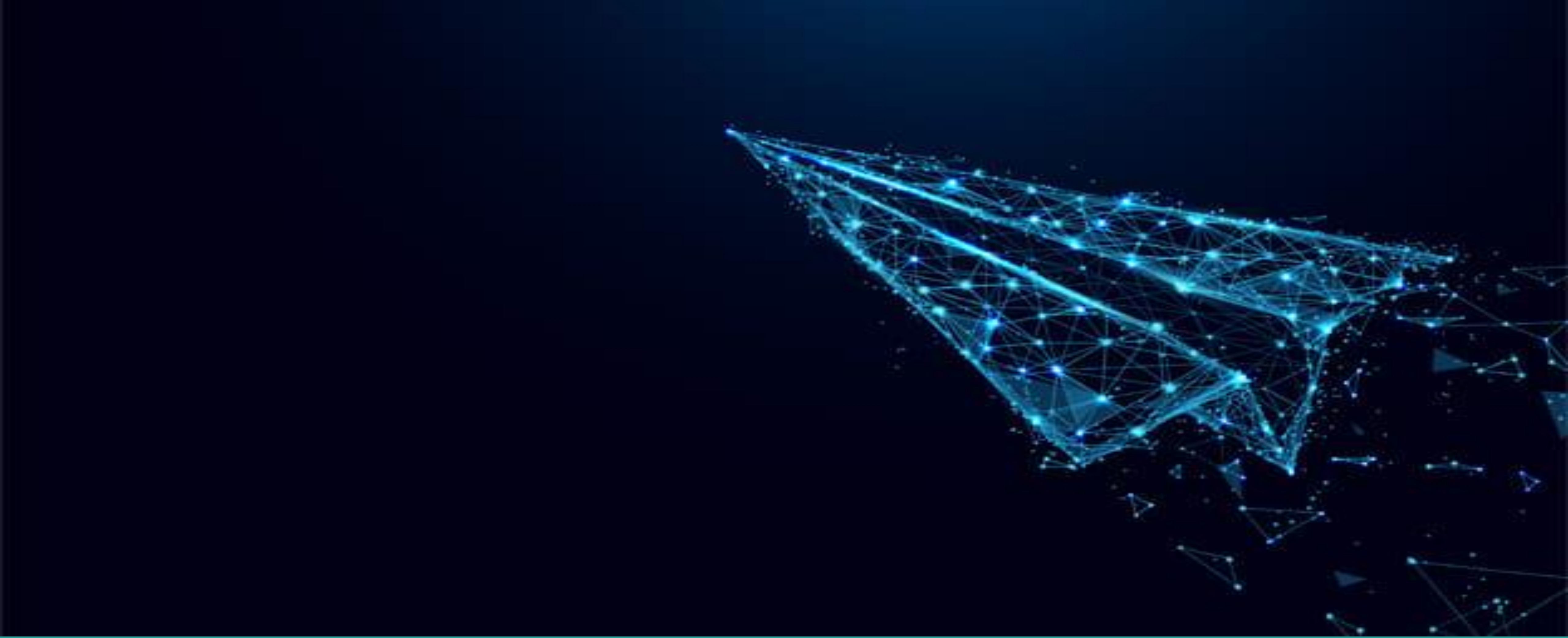
Smart Meters - What to do?

- **Model the smart meter data.**

- Share the model (outputs), instead of the original data.

Disclaimer:

There are more than one method of modelling smart meter data. We will refer only the generative (probabilistic) modelling.



Generative models

Generative models - Introduction

thispersondoesnotexist.com

Generative models - Introduction

"Teddy bears working on new AI research…"

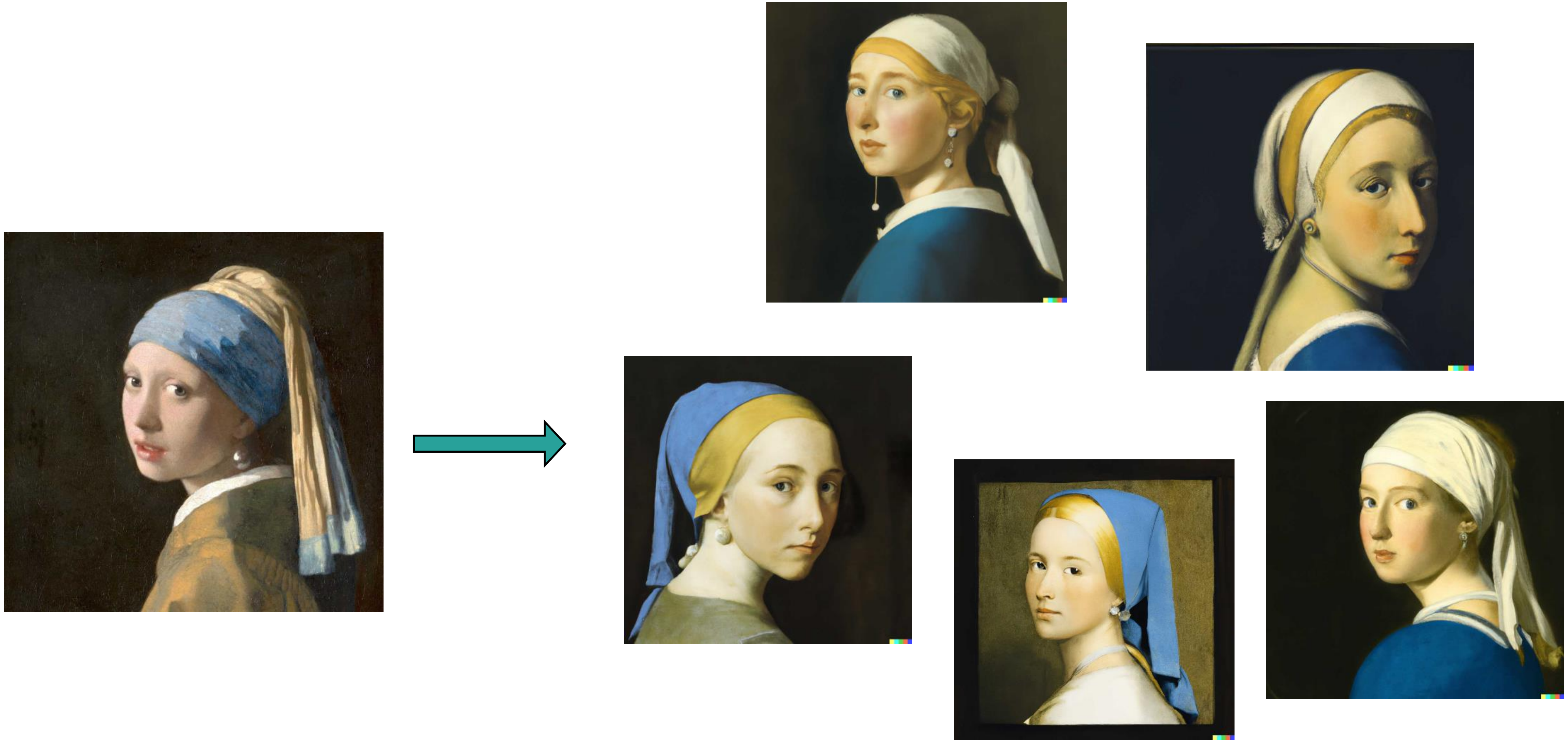
"… as kids' crayon art."

"… on the moon 1980s."

"… underwater with 1990s technology."

https://openai.com/dall-e-2/#demos

Generative models - Introduction

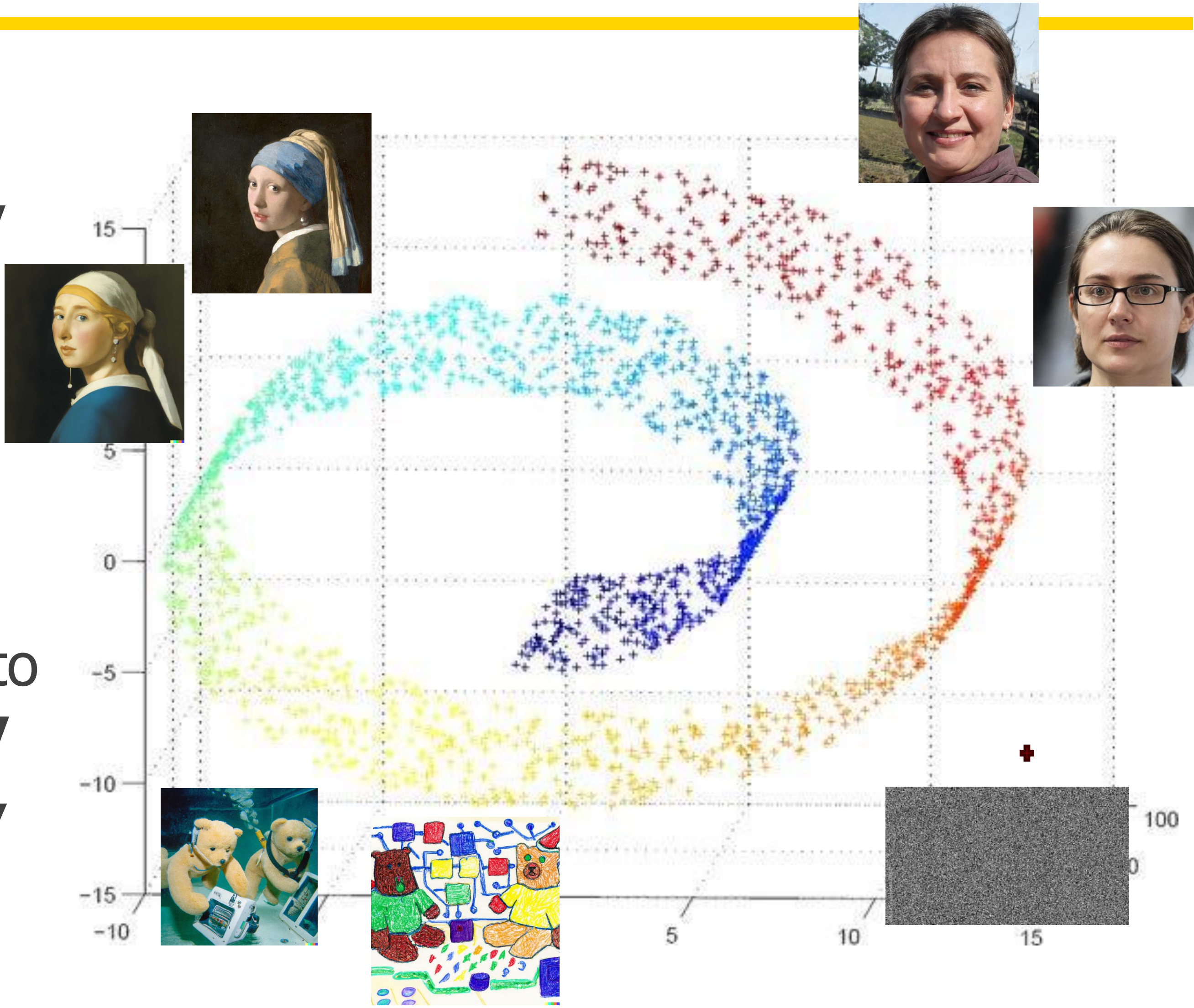


https://openai.com/dall-e-2/#demos

Generative models - Introduction

- Why is this so impressive?
  - Image and text data reside in **very high dimensional** spaces.
    - +1M dimensions for a 1028*1028 image.
    - Most of this space is empty (meaningless).

  - It is nearly impossible for humans to comprehend and describe the **very complex** relations in these spaces, mathematically or algorithmically.



Generative models - Benefits

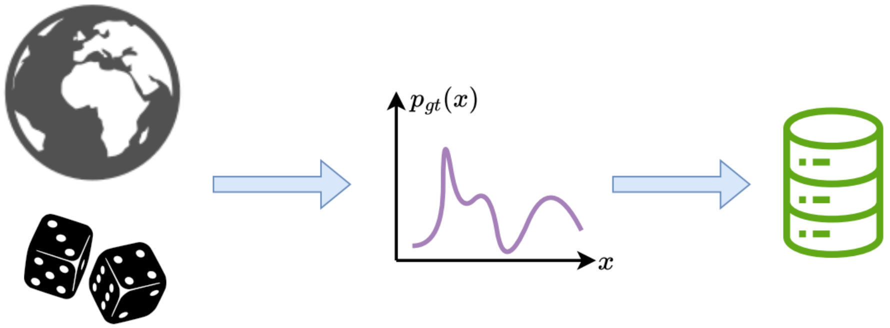
- Generate novel ‘test data’: *how does my system/process perform with unseen scenarios?*
- Generate large amounts of training data for other machine learning models
  - Train models that are prone to overfitting (incl. adversarial models)
  - **Warning**: there is no free lunch – you don’t generate more information
- Embed bias in generated data:
  - Bias during model training, e.g. physical constraints on outputs
  - Bias during data generation, e.g. generate extreme weather scenarios
- …

Generative models – (Soft) Objectives

1. **Individually**, samples should be ‘realistic’
2. **Collectively**, samples should look like the population
3. The model should **generalise** from the training data
4. There may be **privacy/ownership concerns** over individual data points

Generative models - Objectives

- **Assumption:** Every dataset is a collection of samples from an <u>unknown</u> real-life probability distribution

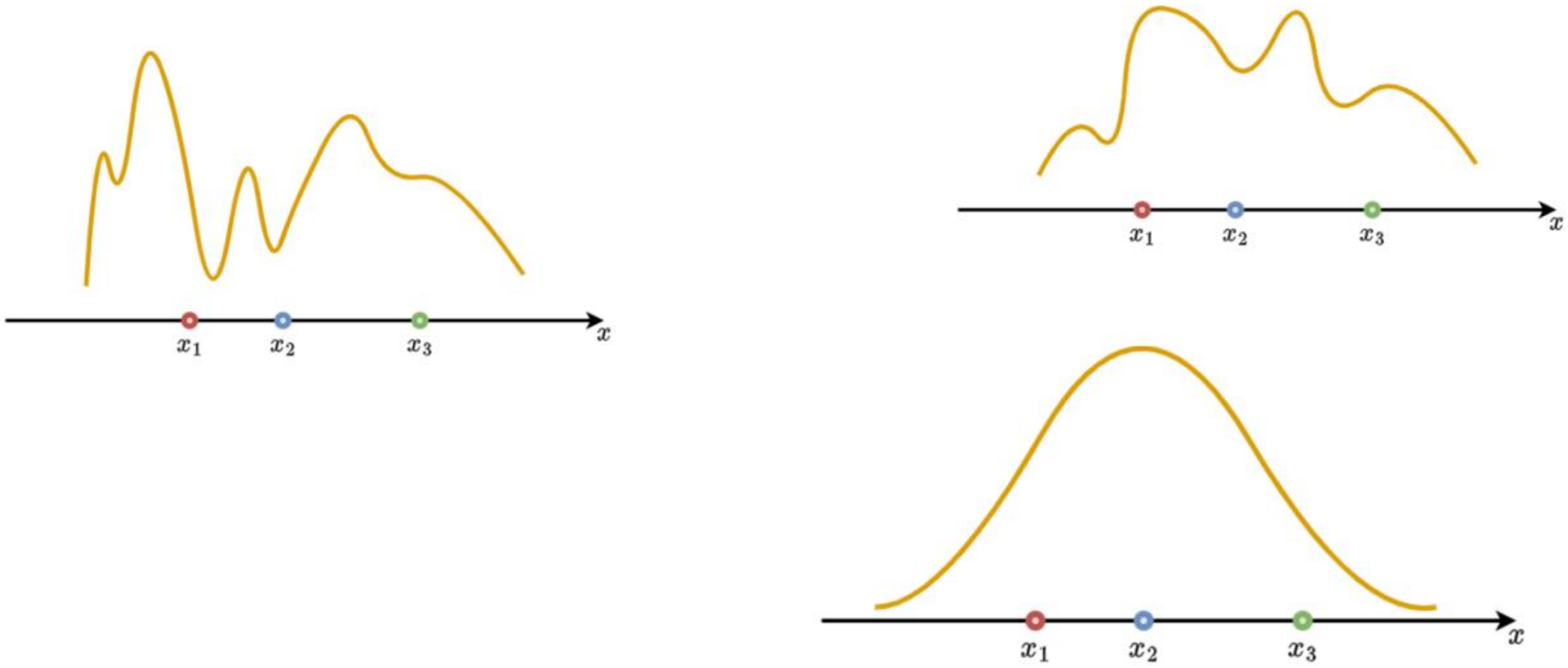


$$\mathcal{X} = \{\mathbf{x}_i | \mathbf{x}_i \sim p_{\text{real-life}}(\mathbf{x})\}_{i=1}^{N}$$

- **Motivation:** If we can estimate $p_{real-life}(\boldsymbol{x})$ as $p_\theta(\boldsymbol{x})$, we can sample more data from $p_\theta(\boldsymbol{x})$ (**synthetic data generation**).

Generative models - Objectives

- How are we going to choose $p_\theta(\boldsymbol{x})$?



Generative models - Objectives

- **Objective:** Minimize the dissimilarity between $p_{real-life}(\boldsymbol{x})$ and $p_\theta(\boldsymbol{x})$.

$$\min_\theta \; D_{KL}(p_{\text{real-life}}(\mathbf{x}) \| p_\theta(\mathbf{x}))$$

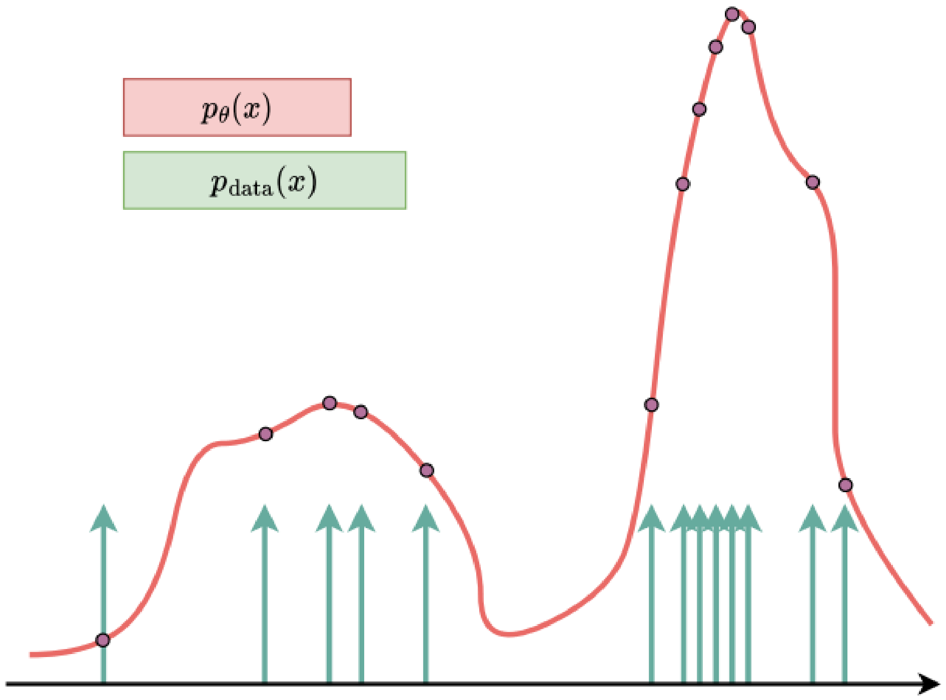
$$= \min_\theta \int_{\mathbb{X}^D} p_{\text{real-life}}(\mathbf{x}) \log \left(\frac{p_{\text{real-life}}(\mathbf{x})}{p_\theta(\mathbf{x})} \right) d\mathbf{x}$$

https://gnarlyware.com/blog/kl-divergence-online-demo/

Generative models - Objectives

$$p_{\text{real-life}}(\mathbf{x}) \rightarrow p_{\text{data}}(\mathbf{x}) = \frac{1}{N}\sum_{\mathcal{X}} \delta(\mathbf{x} - \mathbf{x}_i)$$

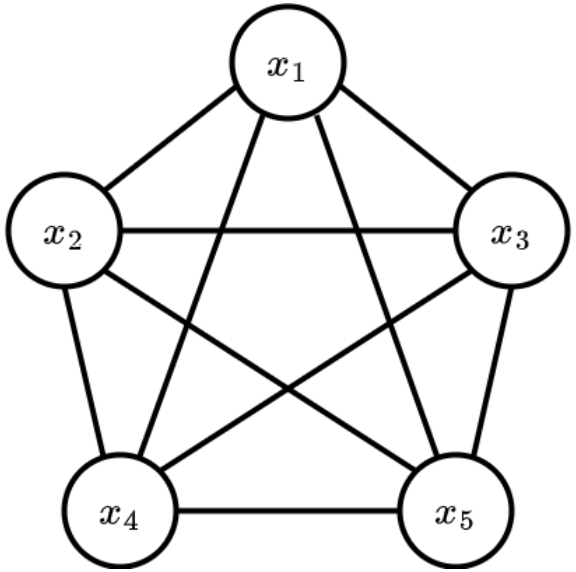
$$\underset{\theta}{\text{argmin}}\ D_{KL}(p_{\text{data}}(\mathbf{x})||p_\theta(\mathbf{x})) = \underset{\theta}{\text{argmax}}\ \frac{1}{N}\sum_{\mathcal{X}} \log p_\theta(\mathbf{x}_i)$$



Generative models – High dimensions

Goal: Find a (parameterised) probabilistic model $p(\boldsymbol{x})$, where x is high-dimensional.

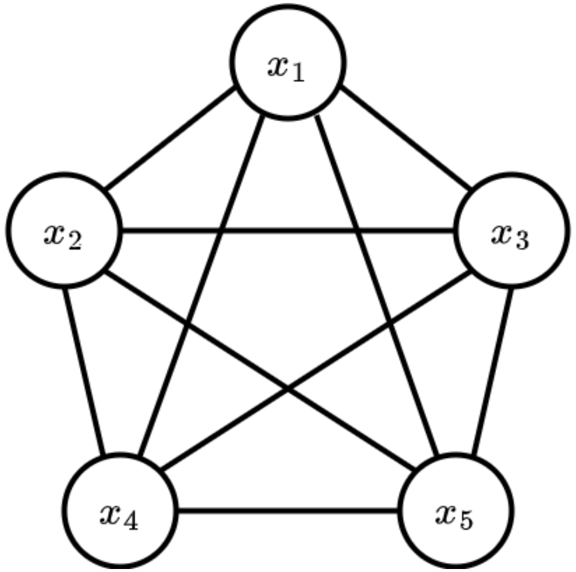
Problem: Finding/learning relations between **many** features is exceedingly hard (even for very deep and wide neural networks).



$$p(\boldsymbol{x}) = p(x_1, x_2, x_3, \dots x_d)$$
$$= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_d|x_1, \dots, x_{d-1})$$

Generative models – High dimensions

- An example – Parameter efficiency
  - $\boldsymbol{x} = [x_1, x_2, x_3, x_4, x_5, \dots, x_d]' \in \{0,1,\dots,K\}^d$
  - Total number of probability values to learn/estimate:
    - $(K-1) + K(K-1) + K^2(K-1) + \ \dots + K^{d-1}(K-1) = (K-1)\sum_{i=0}^{d-1} K^i$
  - For a 16x16 image ($d = 256$)
    - Black and white $(K = 2)$: $\sim \mathbf{10^{77}}$
    - Grey-scale $(K = 256)$: $\sim \mathbf{10^{616}}$
    - RGB $(K = 768)$: $\sim \mathbf{10^{738}}$

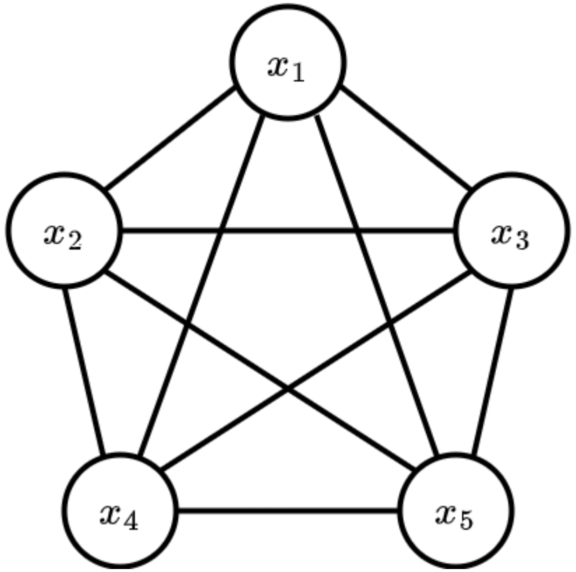


$$p(\boldsymbol{x}) = p(x_1, x_2, x_3, \dots x_d)$$
$$= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_d|x_1, \dots, x_{d-1})$$

Generative models – High dimensions

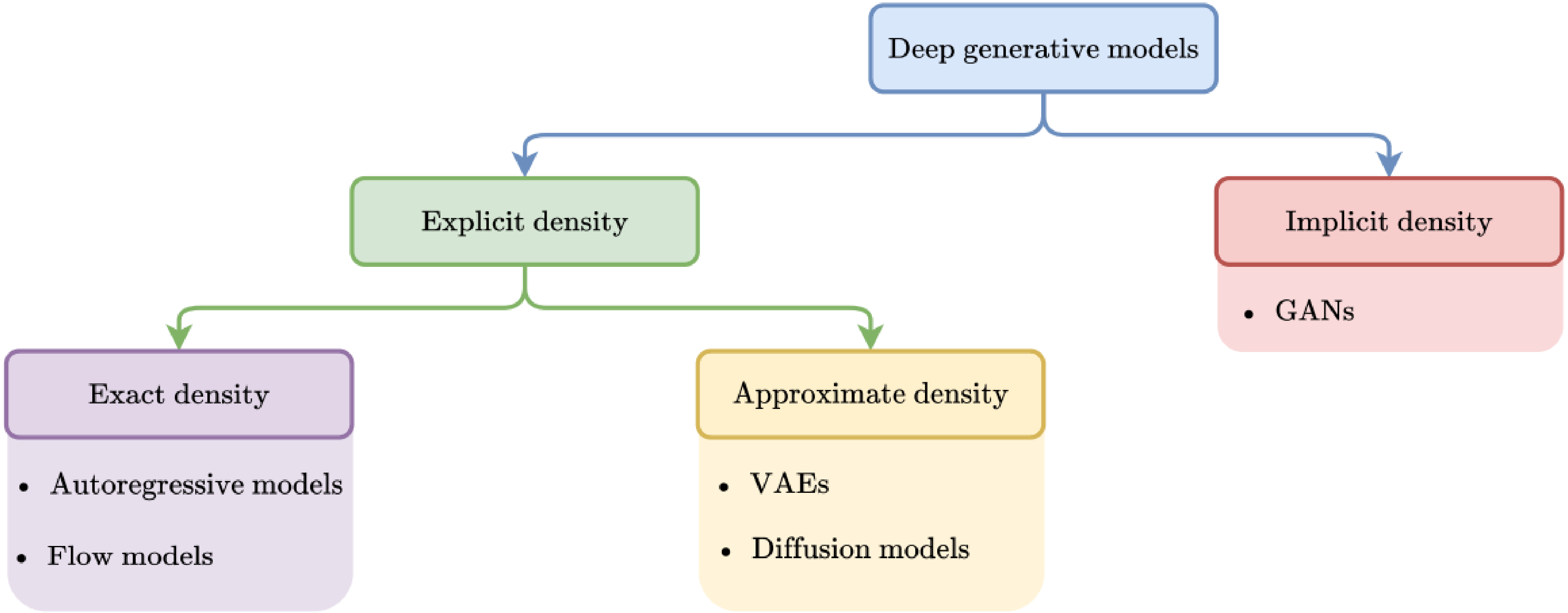
Solution?

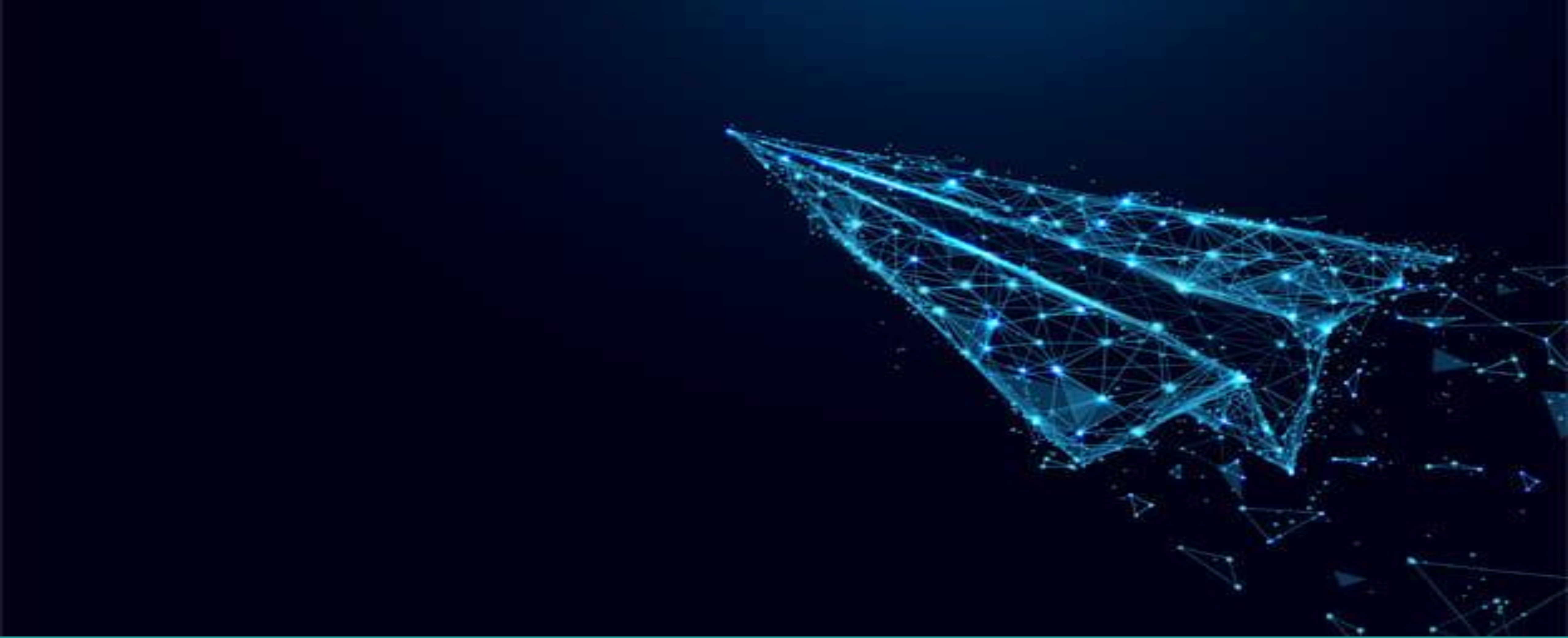
Deep Learning



$$p(\boldsymbol{x}) = p(x_1, x_2, x_3, \dots x_d)$$
$$= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_d|x_1, \dots, x_{d-1})$$

(Deep) Generative models - Taxonomy

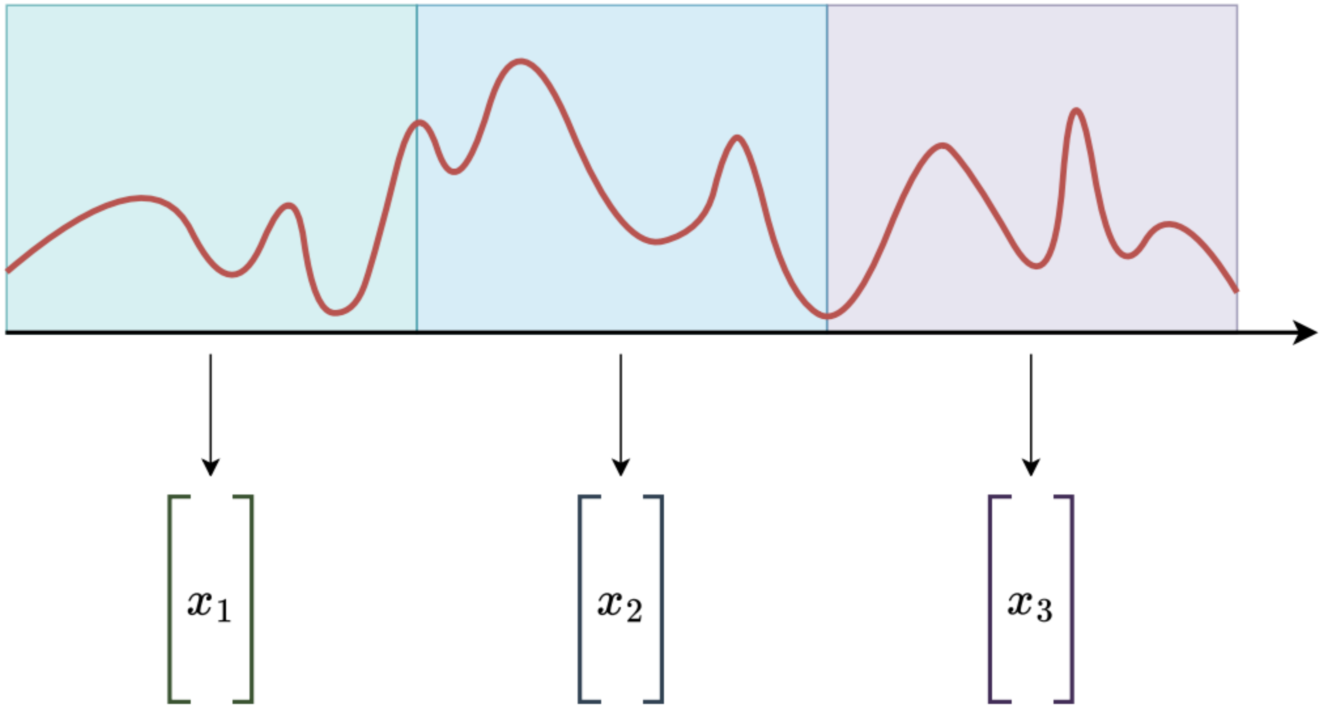




Modelling smart meter data

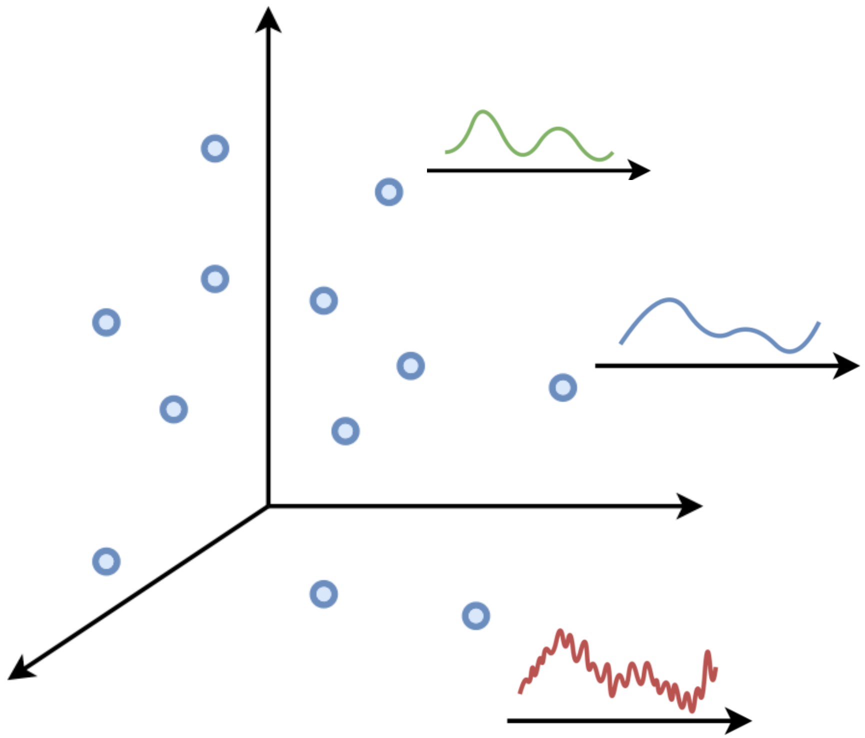
Modelling smart meter data

- We can assume that the smart meter data consists of “snapshots” of load profiles.



Modelling smart meter data

- After “profiling”, these snapshots are only some numeric points in some sort of metric space.
- Now, you can apply your favourite generative model to your snapshot dataset…
- … with some challenges.



Modelling smart meter data - Challenges

- Modern "$p_\theta(\boldsymbol{x})$ options" are
  - powerful
  - versatile, and
  - (almost) ready-to-use

- What can go wrong?

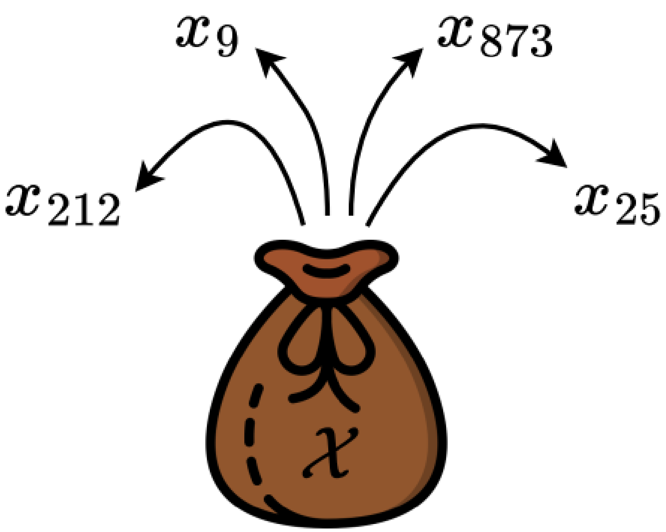
Modelling smart meter data - Challenges

Challenge #1: Too much flexibility!

- **Intuition:** $p_{data}(\boldsymbol{x})$ is a probability distribution too!

- **Recall:** $\underset{\theta}{\text{argmin}}\ D_{KL}(p_{\text{data}}(\mathbf{x}) || p_\theta(\mathbf{x}))$

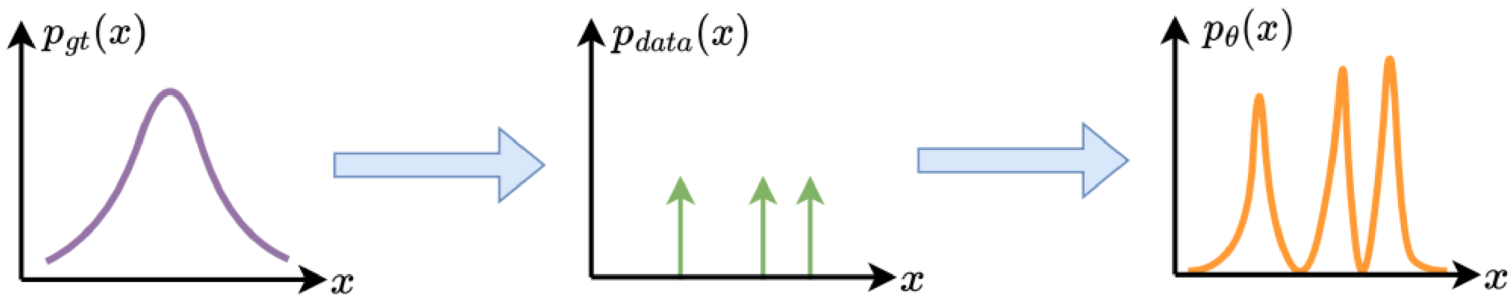
- Our objective forces $p_\theta(\boldsymbol{x}) \approx p_{data}(\boldsymbol{x})$.



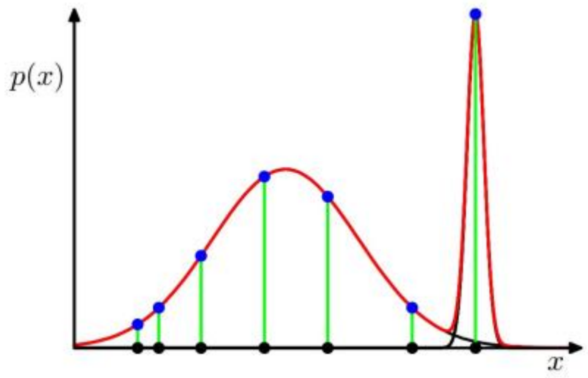
Modelling smart meter data - Challenges

Challenge #1: Too much flexibility!

- Overfitting



- Data-copying



Modelling smart meter data - Challenges

Challenge #1: Too much flexibility!

- **Take-away messages:**
  - Individual data points have **no uncertainty**.
  - Only information we have is the dataset itself. It is not possible to produce more information out of a **limited information**.
  - The uncertainty coming from the model (**epistemic uncertainty**) is not the same as the uncertainty of the real-life distribution (**aleatoric uncertainty**).

Modelling smart meter data - Challenges

Challenge #2: Curse of unsupervision

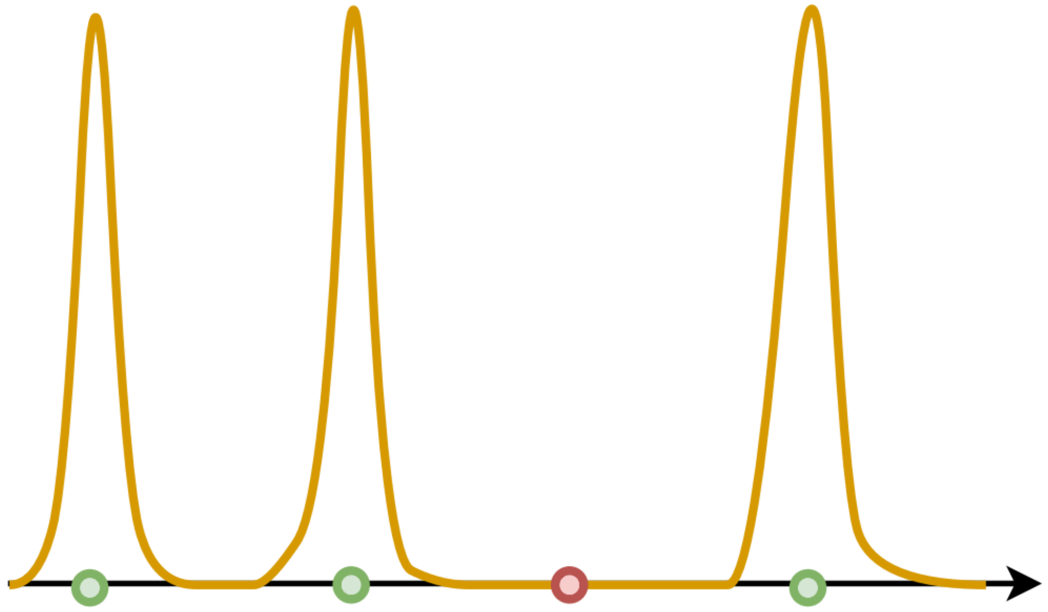
- Evaluation of the "**generative performance**" is not straightforward.

- Two possible vectors of evaluation:
  - Checking **log-likelihood of test data**
  - Checking the generated **samples**

Modelling smart meter data - Challenges

Challenge #2: Curse of unsupervision

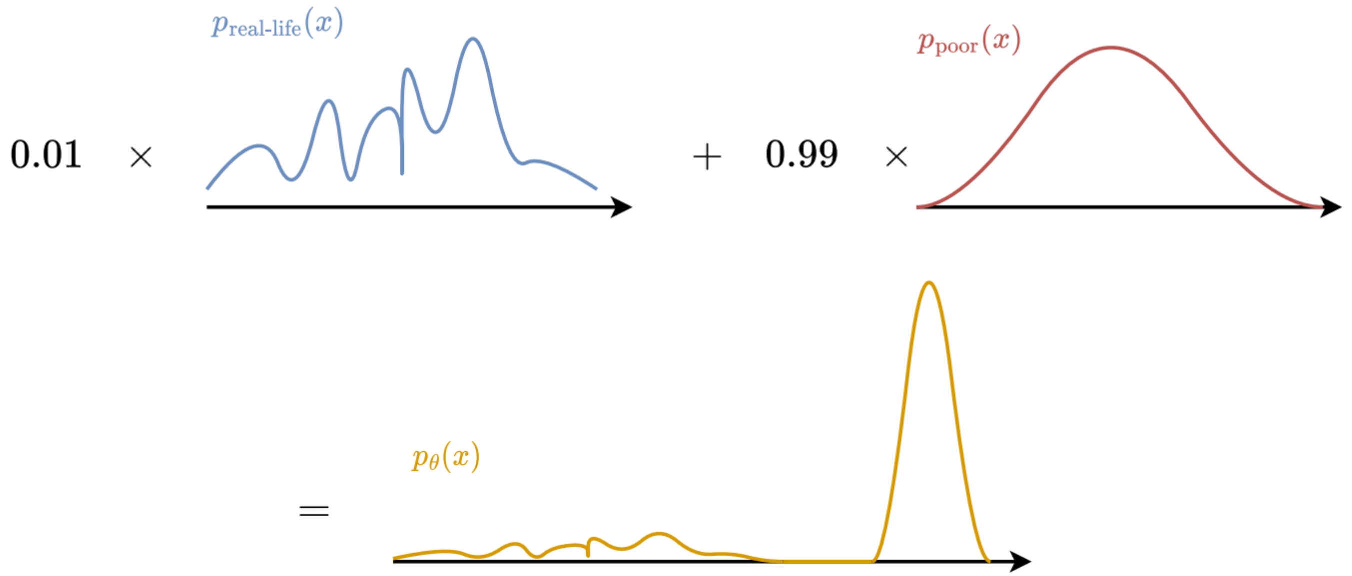
- **Poor likelihood & Great samples**



Modelling smart meter data - Challenges

Challenge #2: Curse of unsupervision

- **Great likelihood & Poor samples**

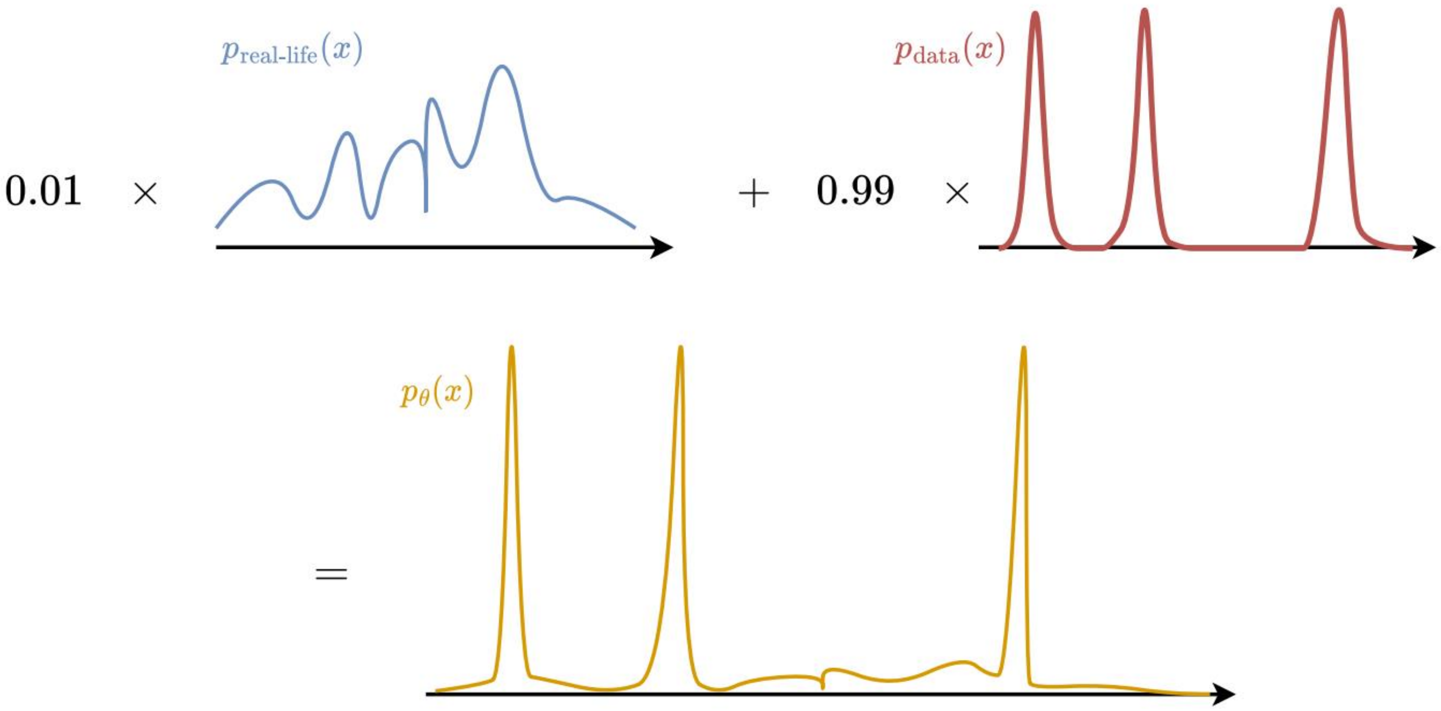


$$\log(0.01 p_{\text{real-life}}(x) + 0.99 p_{\text{poor}}(x)) > \log p_{\text{real-life}}(x) - \log 100$$

Modelling smart meter data - Challenges

Challenge #2: Curse of unsupervision

- **Great likelihood & Great samples**



Modelling smart meter data - Challenges

Challenge #2: Curse of unsupervision

- **Take-away messages:**
  - Evaluation of generative models is still an **open research topic**.
  - **Don’t trust** your log-likelihood values and generated samples.
  - Validation/test set strategies are not sufficient for a **good generative performance**.
  - What does “good generative performance” mean anyways?

Modelling smart meter data - Challenges

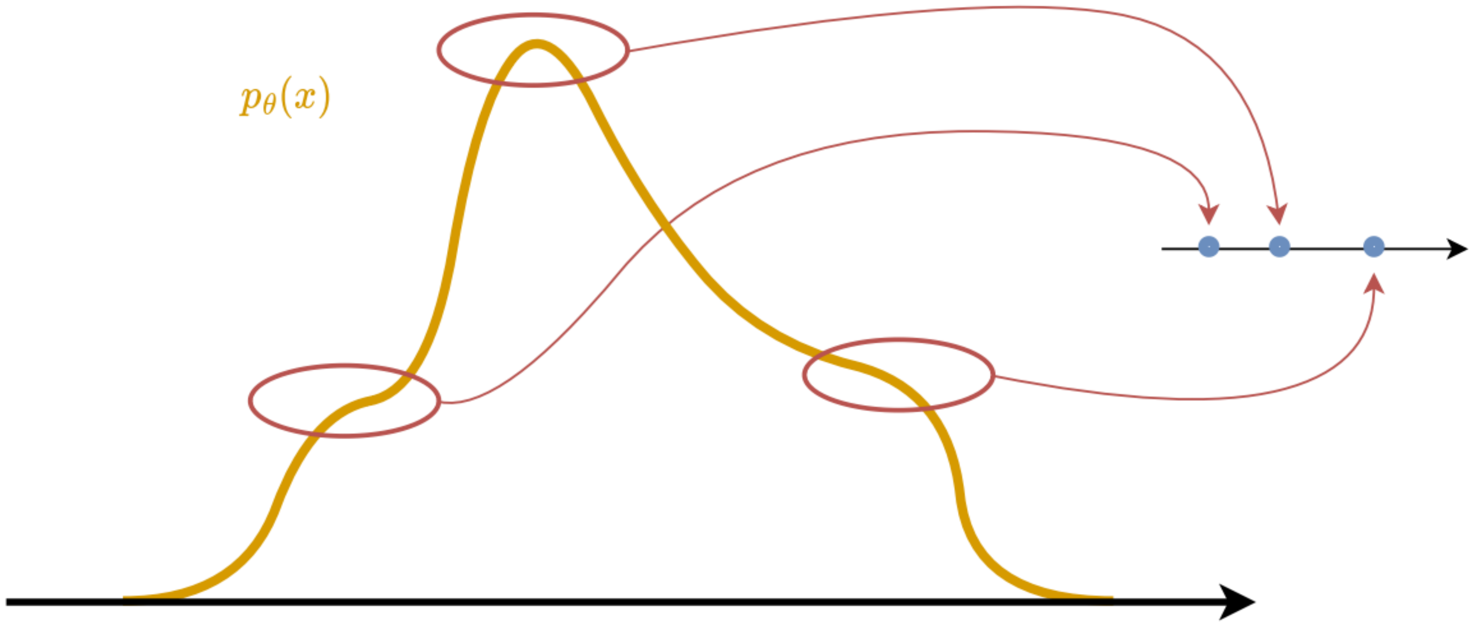
Challenge #3: Privacy. Privacy? Privacy!

- IF overfitting or data-copying, THEN privacy violation
  - Other way around is not necessarily true!
- A generative model can violate privacy even if it does not copy any data!

Modelling smart meter data - Challenges

Challenge #3: Privacy. Privacy? Privacy!

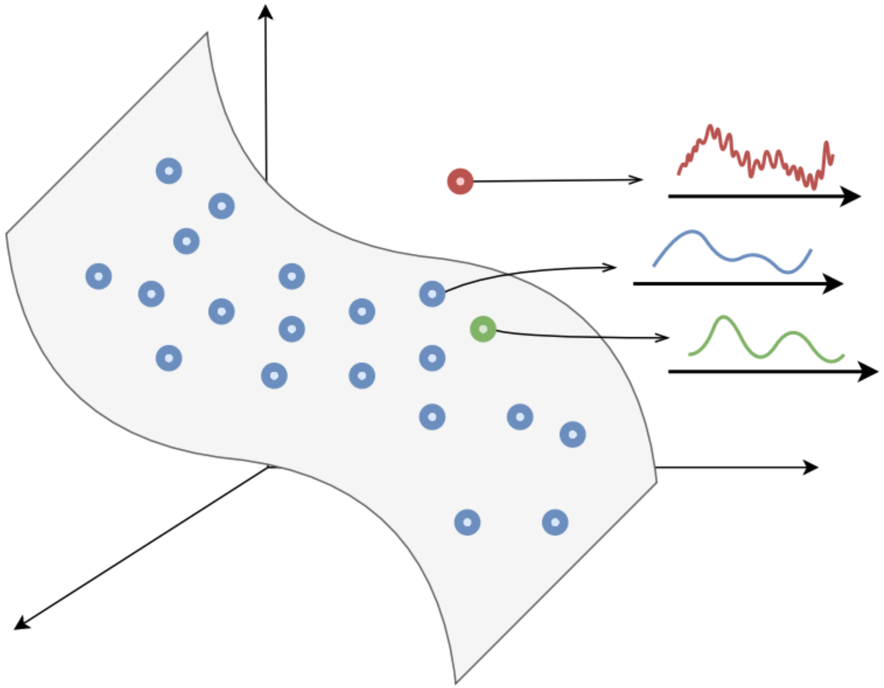
- **Membership inference attacks:**



Modelling smart meter data - Challenges

Challenge #3: Privacy. Privacy? Privacy!

- Euclidean distance from the original data points is not indicative for privacy preservation.



Modelling smart meter data - Challenges

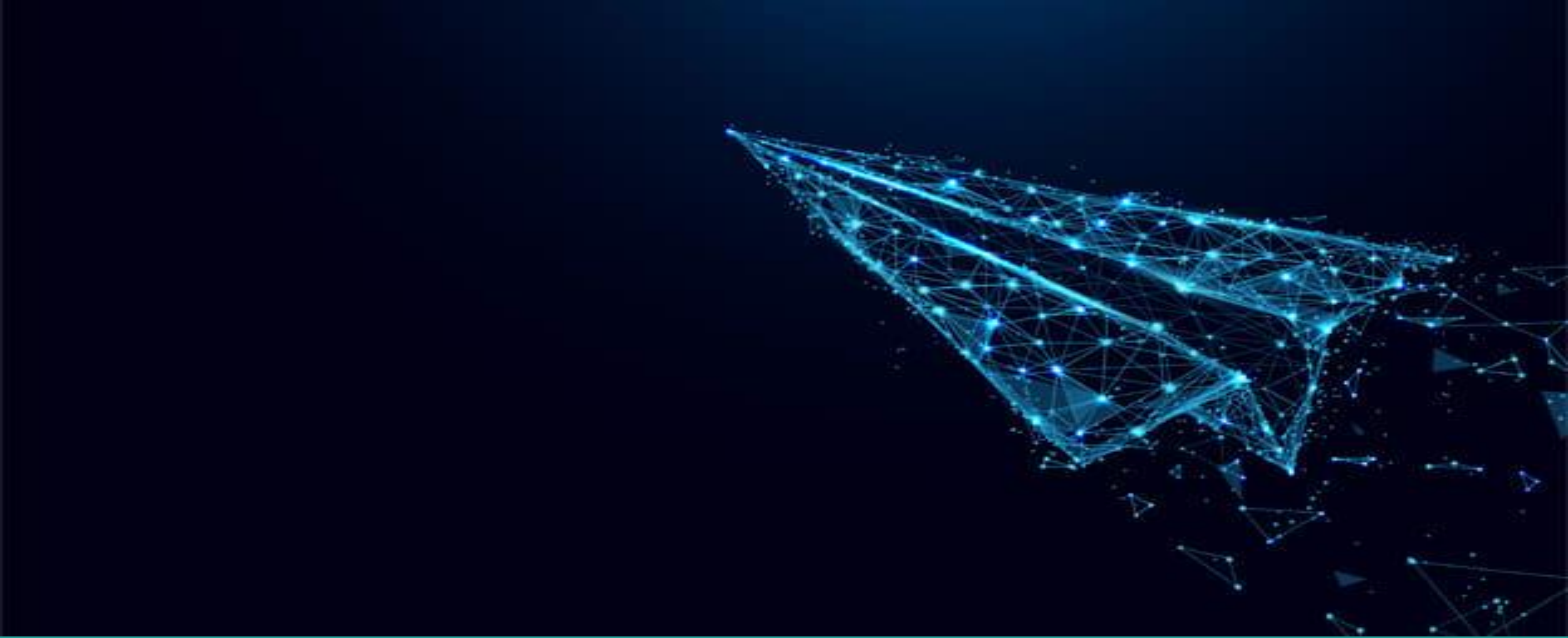
Challenge #3: Privacy. Privacy? Privacy!

- **Take-away messages**:
  - Synthetic data is **not a silver bullet** for privacy.
  - Assessing privacy preservation is not straightforward for **raw data**.
  - There is no consensus on the mathematical definition of **smart meter privacy**.
  - We do not have **attack models** for smart meter data to test our generative model or synthetic dataset.

Modelling smart meter data - Challenges

Honourable challenges

- How do we include user-level statistics in the model?
- What about the spatio-temporal constraints?
- Privacy-by-design or dataset curation?
- How to convince privacy officers and lawyers?



Conclusion

Conclusion

- Smart meters are crucial for modern energy systems.
- Modeling smart meter data helps in generating synthetic data and enhancing grid management.
- Deep generative models are great for this task, but they come with challenges.

Thanks for your attention!

Any questions?

Kutay Bölat

K.Bolat@tudelft.nl
github.com/kabolat

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956433.