Synthetic population generation using GANs and expert knowledge

Gael Lederrey,
Tim Hillel, and Michel Bierlaire

AUM2020: Online Global Workshop
28.01.2021

AUM
Symposia on Applied Urban Modelling

EPFL

Outline

- Motivation
- State-of-the-art
- GANs
- Research perspective
- DATGAN
- Conclusion

Motivation

- Agent-based simulation relies on accurate representations of a population.
- **But:** infeasible to obtain detailed socio-economic data for full population - (privacy/security/cost!)
- => Agent-based simulations typically make use of synthetic population.

State-of-the-art for population synthesis

- -> 2010’s: Iterative Proportional Fitting (IPF)
  - *Beckman et al., 1996*: First paper using IPF
  - *Auld et al., 2009*: Improvements on IPF
- 2010-2015: Monte Carlo Simulations
  - *Farooq et al., 2013*: MCMC simulation with Gibbs sampling
  - *Casati et al., 2015*: Hierarchical MCMC
- 2015-2019: Bayesian Networks
  - *Sun and Erath, 2015*: First to propose Bayesian Networks
  - *Zhang et al., 2018*: Bayesian Network as Social Network
- 2019->???: Deep Learning
  - *Borysov et al., 2019*: First use of a Variational AutoEncoder
  - *Badu-Marfo et al., 2020*: Composite Travel Generative Adversarial Network (CTGAN)

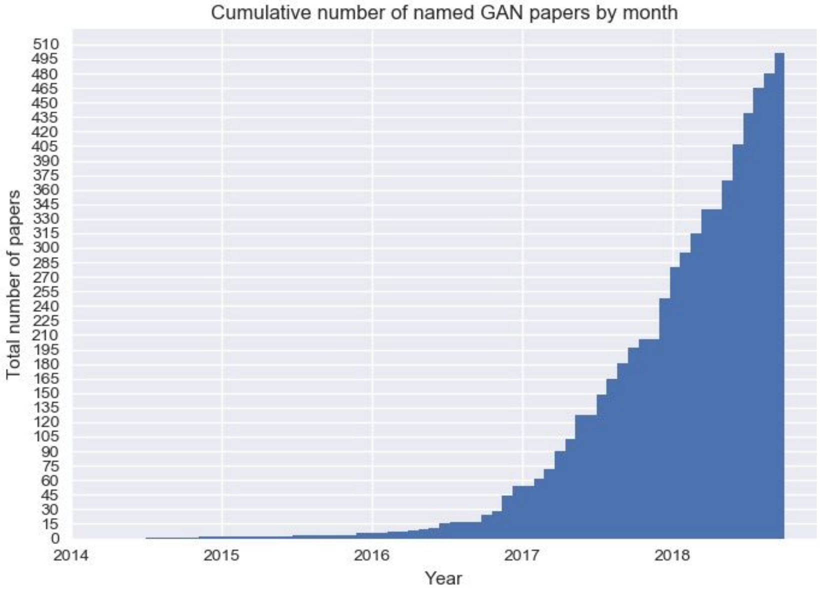
Data generation in Deep Learning

- 2014: Generative Adversarial Networks (GANs)
  - Goodfellow et al., 2014

- 2014->2021: Many iterations of GANs for images

- 2018: GANs for tabular data are proposed
  - Xu et al., 2018 & Park et al., 2018

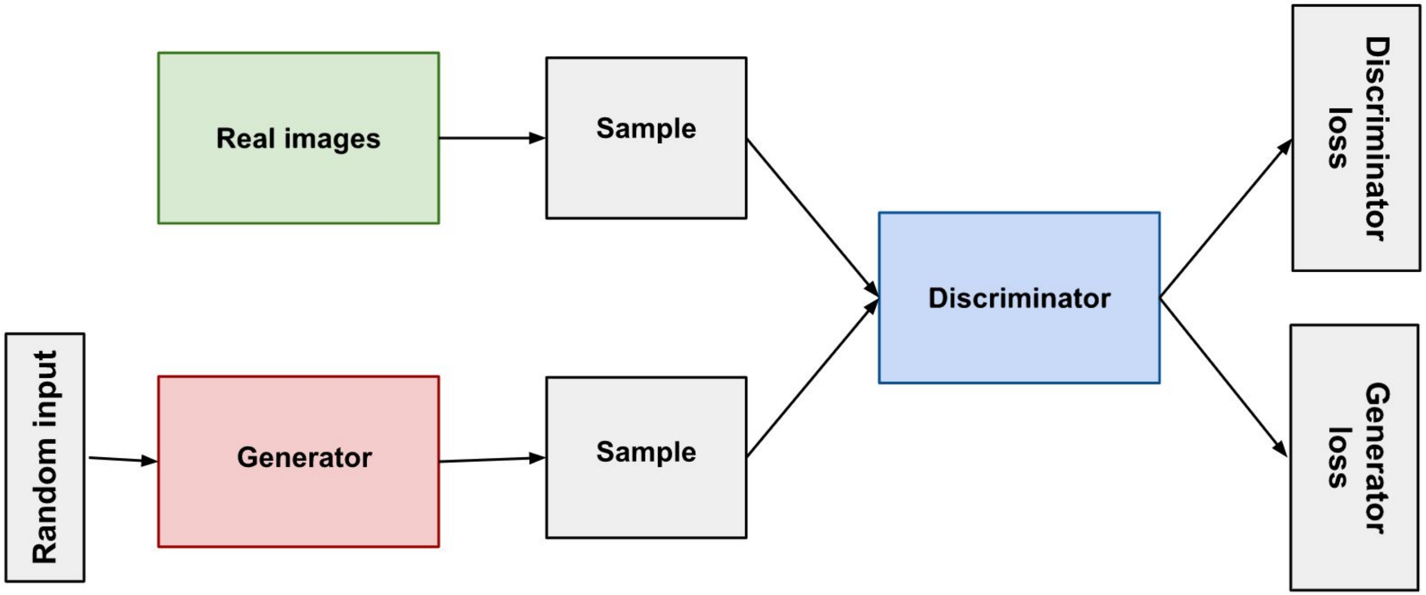
- Limited work on data representativity ar
  - Arora et al., 2017 & Liang, 2018



Generative Adversarial Networks (GANs)

- Idea: Train 2 NNs “simultaneously”, one to generate images data and one to discriminate between fake and real.

- Basic architecture:



Generative Adversarial Networks (GANs)

STAGE1 Generator

Change of architecture 30

The Generator can change its architecture to be faster and more efficient.

Fool 60

Put the opponent in a state of confusion and gains point for the loss function.

weakness resistance

retreat

VS

STAGE2 Discriminator

Train with real data 30

This technique gives a bonus of +10 to the next “Make a correct guess” attack

Make a correct guess 50

Loose its confusion status and gains points for the loss function.

weakness resistance

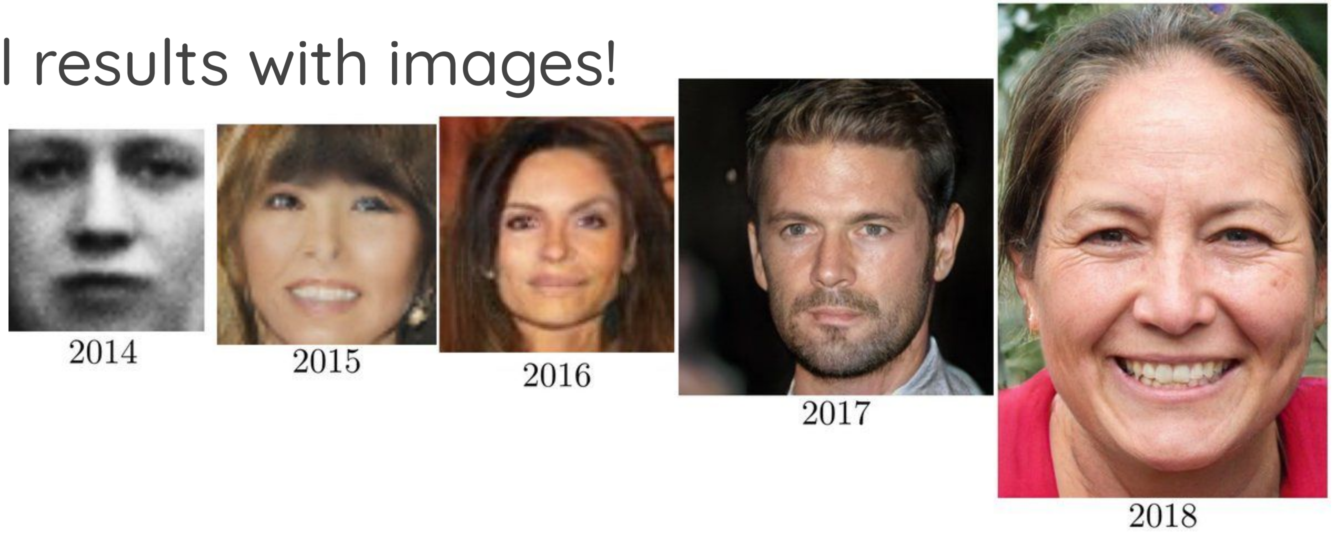
retreat

GANs - Early models

- Standard architectures for both NNs.
  - ANN in both cases

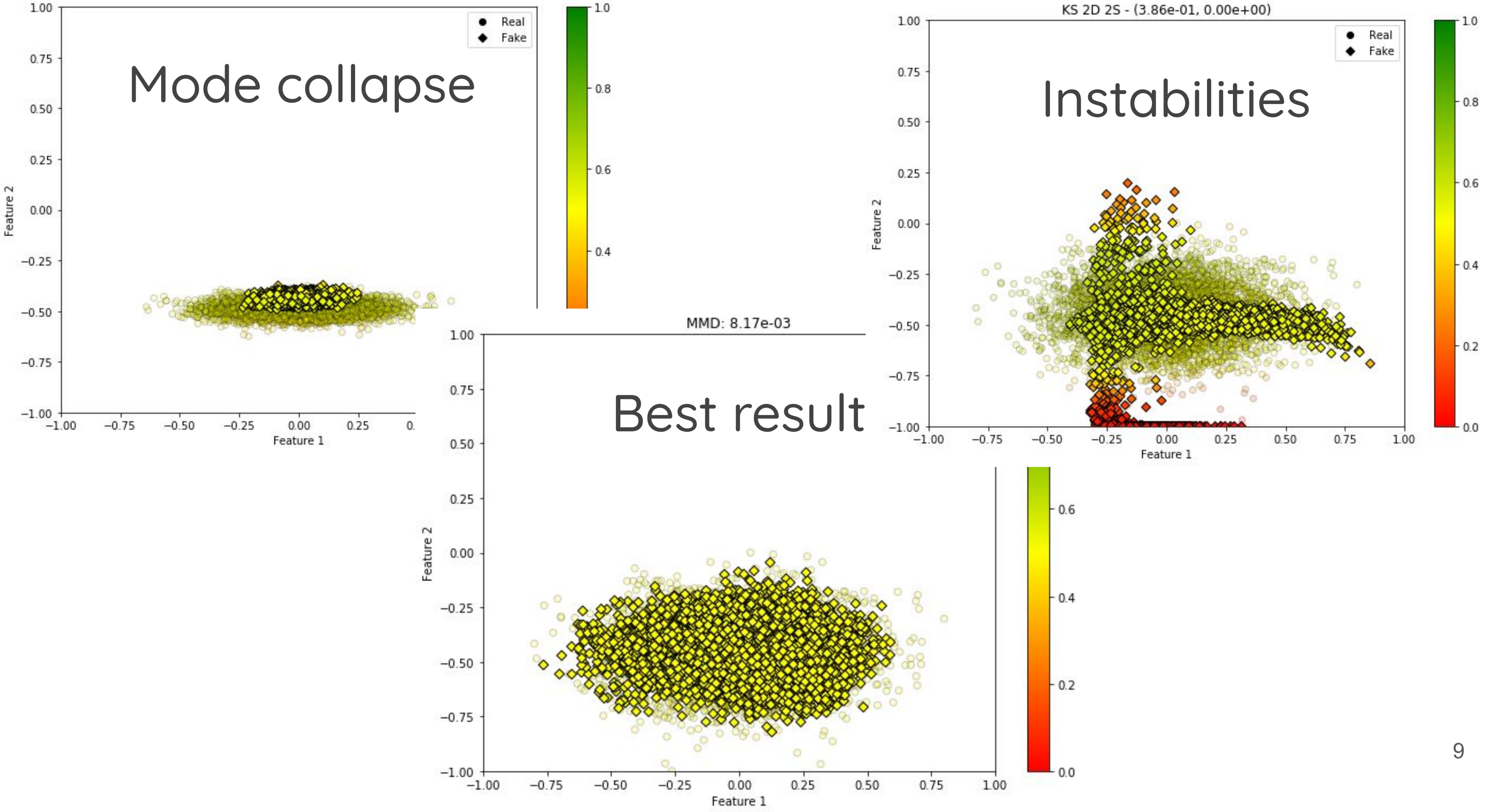
- First improvements made on
  - Loss function (Wasserstein GAN, Cramer GAN)
  - Training stability (WGAN-GP)
  - Coverage and Representativity (MMD-GAN)

- Successful results with images!



Early models and tabular data

- Standard GAN trained on 2D data => bad results!



Mode collapse

KS 2D 2S - (3.86e-01, 0.00e+00)

Instabilities

MMD: 8.17e-03

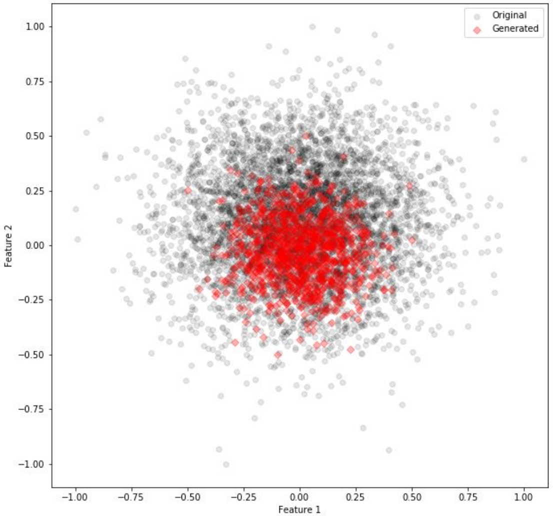
Best result

What is “representativity” in data?

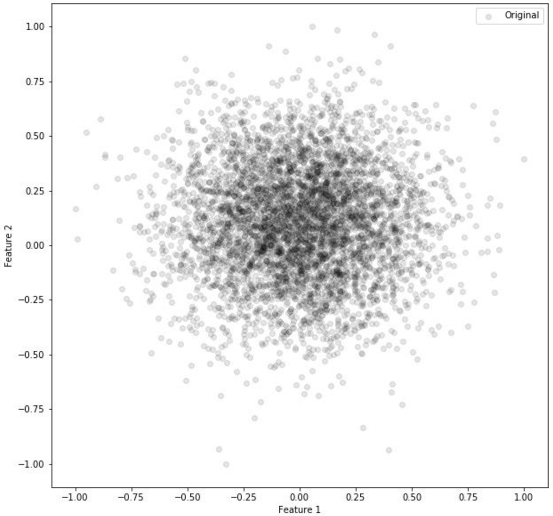
- Concept of representativity = generate new data that reflect the original distribution.

- ⚠️ different from generating data that fool a discriminator!

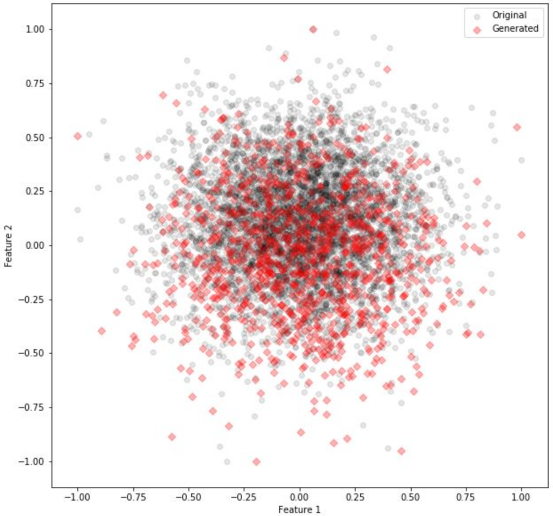
Bad repr.



Original data



Better repr.



Research perspective

- Two parallel research directions:

Improvement of population synthesis

Develop new robust ML models for synthetic population generation

=> Current SoR: TGAN

Representativity assessment

Develop new statistical method to better assess the model performance

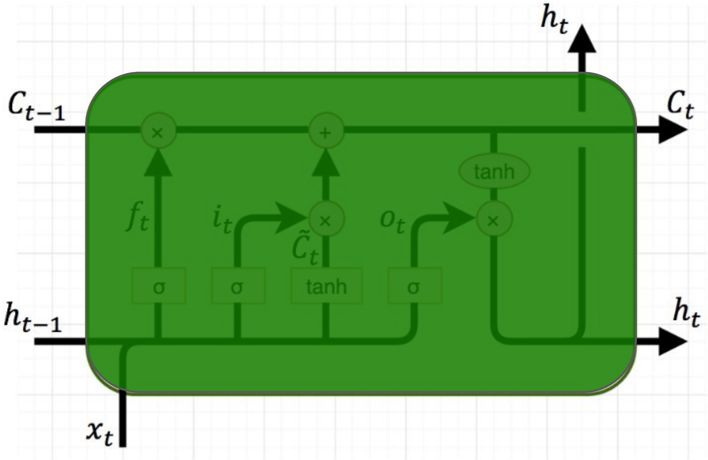
Current SoR: SRMSE

Mueller and Axhausen, 2011

TGAN and flaws

- TGAN stands for Table GAN
  - Xu et al., 2018
- Main idea:
  - Architecture for Generator = sequence of LSTM cells

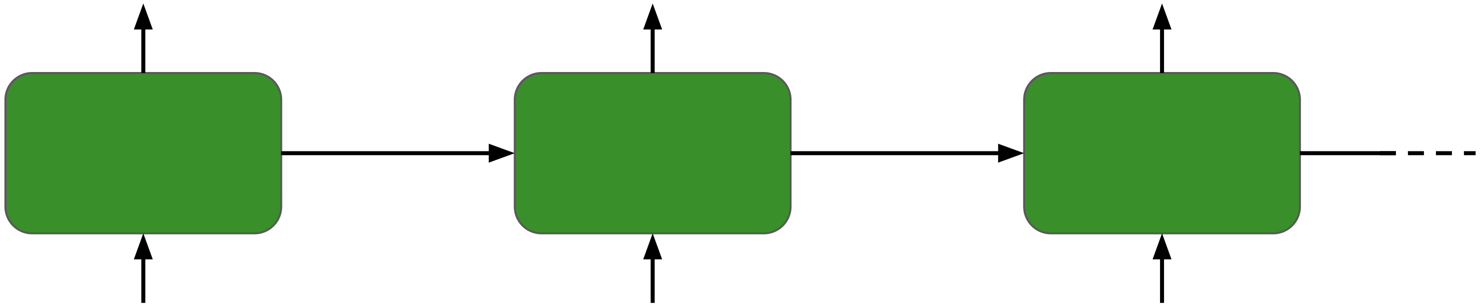
h_t: output
C_t: cell state
X_t: input



TGAN and flaws

- TGAN stands for Table GAN
  - Xu et al., 2018

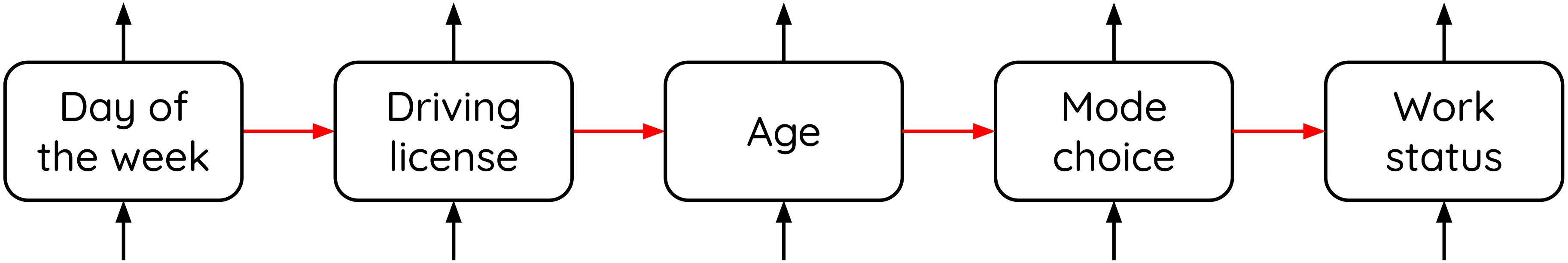
- Main idea:
  - Architecture for Generator = sequence of LSTM cells



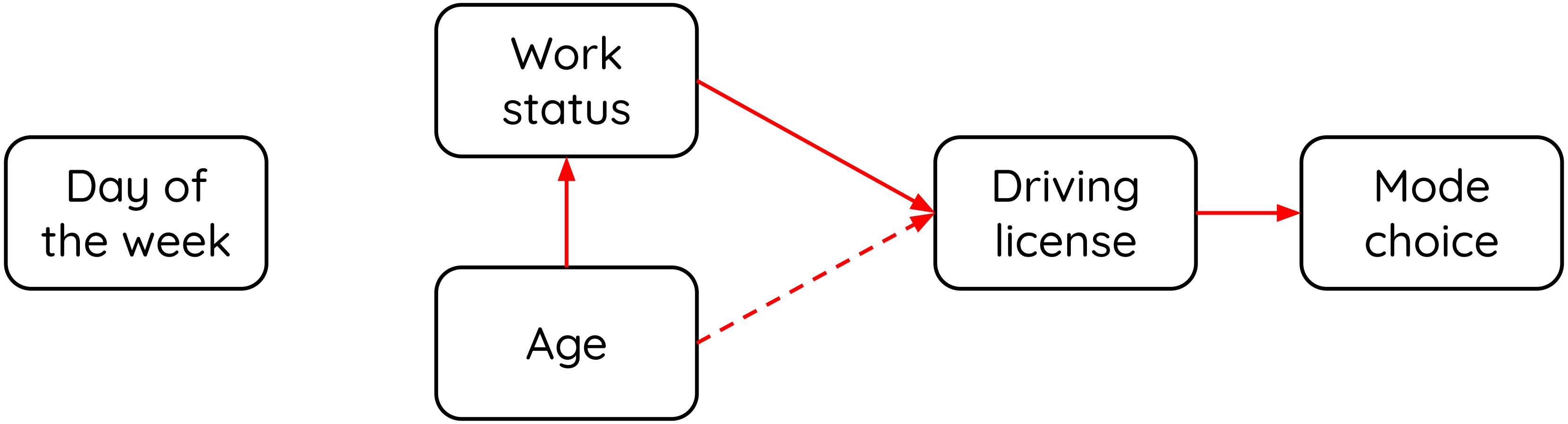
- Flaws:
  - **No “specific” relations between the variables in the dataset**
  - Selection of discrete values using arg max on predicted probabilities

DATGAN

- TGAN



- DATGAN (Directed Acyclic TGAN)



Current work-in-progress

- DATGAN is ~ trainable
- Problem: “How to add multiple inputs to an LSTM cell”
- Possible solutions?:
  - Concatenate inputs and cell states (⚠️ size)
  - Use additional DeepLayers to reduce size (⚠️ training)
  - Transform the current LSTM cell to accept multiple inputs
- Investigation is ongoing - first analytical results due ASAP.

Conclusion and future work

- GANs are current state-of-the-art technique for population synthesis (outperforms previous approaches)

- Proposed directed acyclic graph structure addresses existing limitations of TGAN

- Future work:
  - DATGAN: Finalise implementation
  - Validation: Define more robust metrics for assessing aggregate representativity

Thank you!

Questions?

email: gael.lederrey@epfl.ch