# To be or not to be?
# Mortality Prediction Challenge

Adrien Pavao

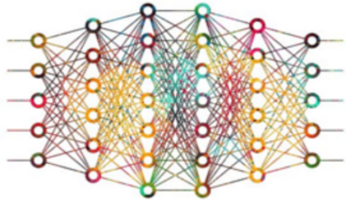
July 30 2018

Laboratoire de Recherche en Informatique

# Introduction

## Project overview

**Data generation**



**Metrics**


**Challenges**

# Introduction

What did we learn from the mini-challenges organized?

- **Chems**:
  `https://competitions.codalab.org/competitions/18751`
- **Mortality**:
  `https://competitions.codalab.org/competitions/19365`

  To be, or not to be?
  Mortality Prediction Challenge

**No end dates. Go for it!**

# Use of synthetic data

**What we want from generated data:**

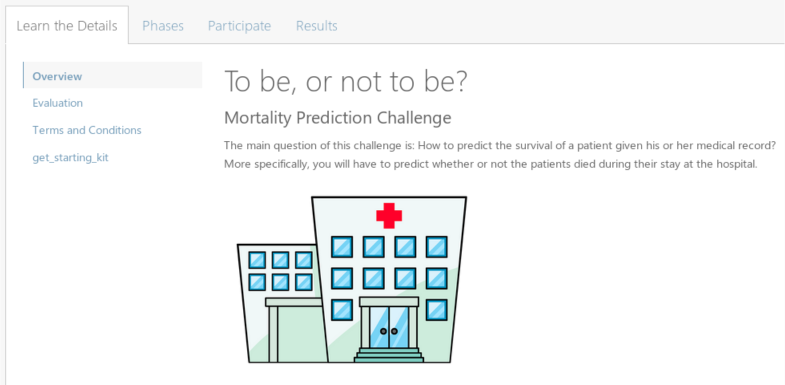
- Respect of privacy
- Same behaviour

**3 levels of synthetic data:**

1. **Student**
2. Research Machine Learning
3. Research scientific discovery

# Challenge presentation

# Mortality prediction challenge




- Synthetic medical data
- Imbalanced binary classification
- Scoring metric: balanced accuracy

# Original data

## MIMIC dataset

| HADM_ID | ADMITTIME | DISCHTIME | INSURANCE | LANGUAGE | RELIGION | MARITAL_STATUS | ETHNICITY | GENDER | ... |
|---|---|---|---|---|---|---|---|---|---|
| 152223 | 2153-09-03_07:15:00 | 2153-09-08_19:10:00 | Medicare | NaN | CATHOLIC | MARRIED | WHITE | M | ... |
| 129635 | 2160-11-02_02:06:00 | 2160-11-05_14:55:00 | Private | NaN | UNOBTAINABLE | MARRIED | WHITE | M | ... |
| 197661 | 2126-05-06_15:16:00 | 2126-05-13_15:00:00 | Medicare | NaN | CATHOLIC | SINGLE | UNKNOWN/NOT_SPECIFIED | M | ... |
| 162569 | 2177-09-01_07:15:00 | 2177-09-06_16:00:00 | Medicare | NaN | CATHOLIC | MARRIED | WHITE | M | ... |
| 104557 | 2172-10-14_14:17:00 | 2172-10-19_14:37:00 | Medicare | NaN | CATHOLIC | MARRIED | UNKNOWN/NOT_SPECIFIED | M | ... |

**Figure 1:** First rows of MIMIC dataset

Class: “DIED” binary variable.

# Wasserstein GAN

Wasserstein GAN[1, 2, 3]

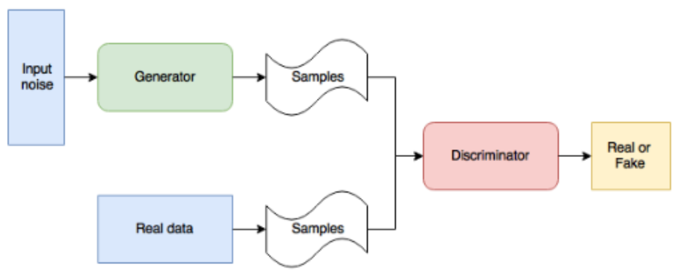



**Figure 2:** GAN architecture

- Discriminator replaced by “Earth Move” loss
- Main hyper-parameters: batch size, neural architecture

# Wasserstein distance

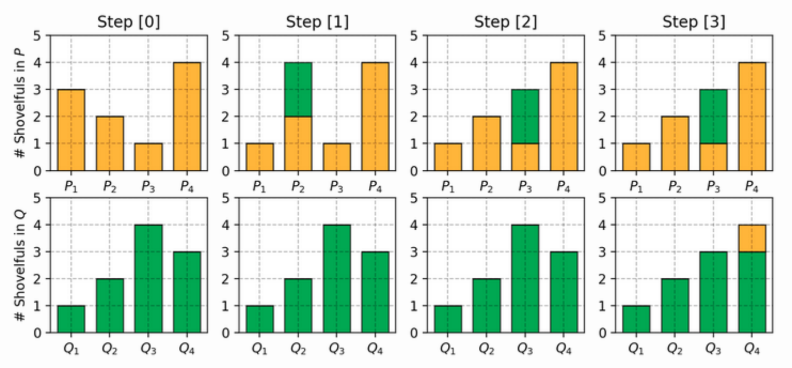



**Figure 3:** Step-by-step plan of moving dirt between piles in P and Q to make them match.

**Continuous** $\rightarrow$ gradient everywhere.

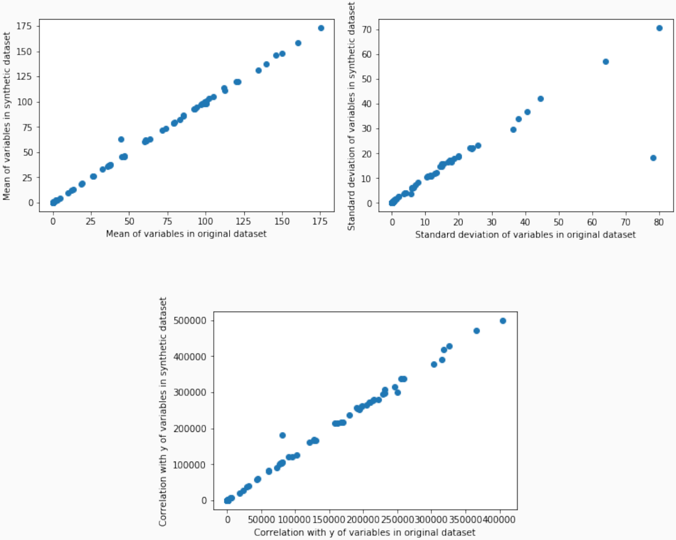
# Generated data

| HADM_ID | ADMITTIME | DISCHTIME | INSURANCE | LANGUAGE | RELIGION | MARITAL_STATUS | ETHNICITY | GENDER | ... |
|---|---|---|---|---|---|---|---|---|---|
| 108398 | 2128-05-15_23:42:00 | 2132-07-23_15:00:00 | Private | ENGL | CATHOLIC | DIVORCED | WHITE | F | ... |
| 186416 | 2134-03-17_03:59:00 | 2113-03-06_12:05:00 | Private | ENGL | UNOBTAINABLE | SINGLE | WHITE | M | ... |
| 126413 | 2164-04-05_17:32:00 | 2180-09-20_16:30:00 | Medicaid | SPAN | CATHOLIC | WIDOWED | OTHER | M | ... |
| 109355 | 2102-09-08_00:58:00 | 2166-06-26_15:30:00 | Medicare | ENGL | NOT_SPECIFIED | MARRIED | WHITE | M | ... |
| 123784 | 2163-08-06_12:07:00 | 2147-01-14_18:40:00 | Medicare | ENGL | UNOBTAINABLE | MARRIED | UNKNOWN/NOT_SPECIFIED | F | ... |

**Figure 4:** First rows of synthetic MIMIC dataset

- 100,000 rows
- Encoded and decoded as in [4]

# Datasets comparison



# Classes distribution

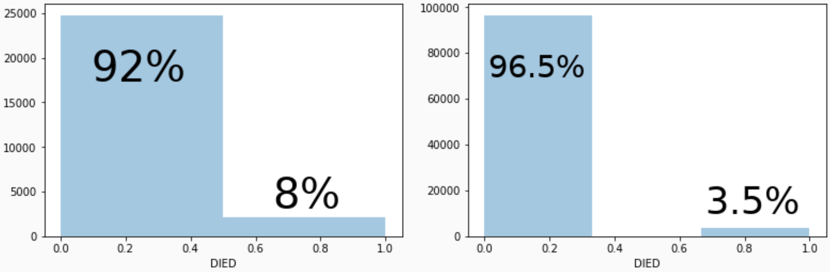



**Figure 5:** Classes distribution in original dataset (left) and synthetic dataset (right)

# Results

# Leaderboard

| RESULTS | | | | | |
|---|---|---|---|---|---|
| # | User | Entries | Date of Last Entry | Team Name | Accuracy ▲ |
| 1 | tianhaogu | 8 | 07/17/18 | Unstoppable league | 0.77 (1) |
| 2 | daiy3 | 7 | 07/17/18 | | 0.76 (2) |
| 3 | Nik.G | 7 | 07/17/18 | | 0.59 (3) |
| 4 | roterj | 1 | 07/18/18 | | 0.52 (4) |
| 5 | Fitztory | 1 | 07/17/18 | PlayerUnknown's Databases | 0.52 (4) |

**Figure 6:** Leaderboard top 5 scores

# Models score

| Model | Train on original<br>Test on original | Train on synthetic<br>Test on synthetic | Train on original<br>Test on synthetic |
|---|---|---|---|
| LogReg | 0.60 | 0.52 | 0.53 |
| GradBoost 150 | 0.61 | 0.52 | 0.53 |
| RF 100 | 0.51 | 0.50 | 0.50 |
| MLP [100] | 0.50 $\rightarrow$ 0.80 | 0.51 | 0.51 |
| MLP [100, 100] | 0.53 $\rightarrow$ 0.91 | 0.51 | 0.51 |

**Table 1:** Balanced accuracy for various models

# Oversampling

| Model | Train on original<br>Test on original | Train on synthetic<br>Test on synthetic | Train on original<br>Test on synthetic |
|---|---|---|---|
| LogReg | 0.76 | 0.76 | 0.77 |
| GradBoost 150 | *0.91* | *0.87* | *0.65* |
| RF 100 | 0.50 | 0.50 | 0.50 |
| MLP [100] | 0.67 $\rightarrow$ 0.82 | 0.65 $\rightarrow$ 0.80 | 0.61 $\rightarrow$ 0.80 |
| MLP [100, 100] | 0.78 $\rightarrow$ 0.91 | 0.54 $\rightarrow$ 0.94 | 0.55 $\rightarrow$ 0.92 |

**Table 2:** Balanced accuracy for various models **after oversampling**

# Conclusion and future work

# Conclusion

**Data generation**

**Metrics**


**Challenges**

# Chems challenge

- Predict biodegradability of molecules
- Training medical students from RPI
- Future improvement: feature selection

| X1 | X2 | X3 | fake1 | fake2 | fake3 | fake4 | fake5 | fake6 |
|---|---|---|---|---|---|---|---|---|
| 3 | 4 | 5 | 5 | 4 | 3 | 5 | 6 | 7.2 |
| 5 | 6 | 7 | 1 | 2 | 5 | 3 | 4 | 2.9 |
| 1 | 2 | 3 | 3 | 6 | 7 | 1 | 2 | 5.4 |

**Table 3:** Adding fake features to data to create a feature selection problem

# References i


- Martín Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN". In: *CoRR* abs/1701.07875 (2017). arXiv: 1701.07875. URL: http://arxiv.org/abs/1701.07875.
- Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 2017, pp. 5769–5779. URL: http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.
- Ian J. Goodfellow et al. "Generative Adversarial Networks". In: *CoRR* abs/1406.2661 (2014). arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. "The Synthetic Data Vault". In: *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016*. 2016, pp. 399–410. DOI: 10.1109/DSAA.2016.49. URL: https://doi.org/10.1109/DSAA.2016.49.