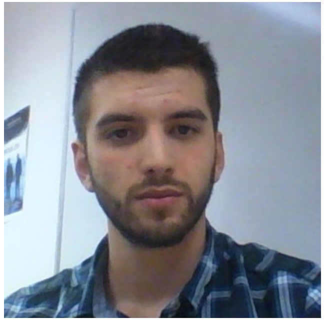
Medi chal

Synthetic Medical Data Generation

Isabelle Guyon, UPSud/INRIA & ChaLearn

The Team



Adrien Pavao
UPSud, Paris
(master student)

Thomas
Gespacher
ENSIMAG
Grenoble
(master student)

Andrew Yale
RPI, NY
(PhD student)

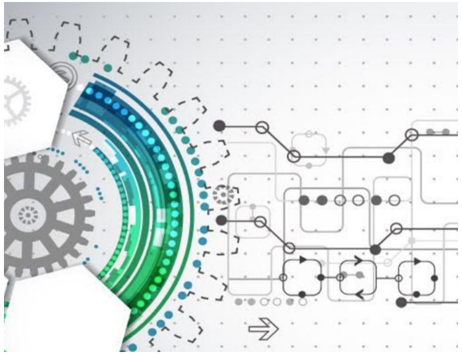
Saloni Dash
Birla Tech, India
(master student)

Ritik Dutta
IIT Gandhinagar
India
(CS, eng. student)

Objectives

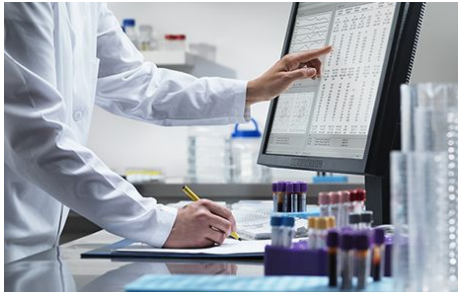
Training

students in health data analytics



Benchmarking new algorithms

Advancing medical research



Method

- Repo: https://github.com/Didayolo/medi-chal/tree/master/
- Data:
  - Target = OPTUM LABS
  - Practice = REPO/data (MIMIC data + classical ML datasets, all in standard format + artificial data)
- Generative models: REPO/code/generators
- Evaluation:
  - ID notebook: REPO/notebooks/auto_ml/ID_notebook.ipynb
  - Comparison notebook: REPO/notebooks/auto_ml/comparison_notebook.ipynb

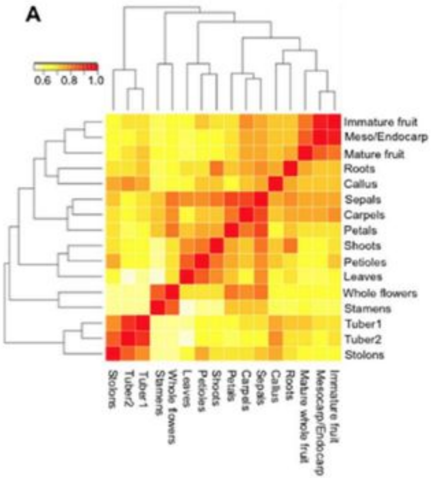
Requirements: retain utility and protect privacy

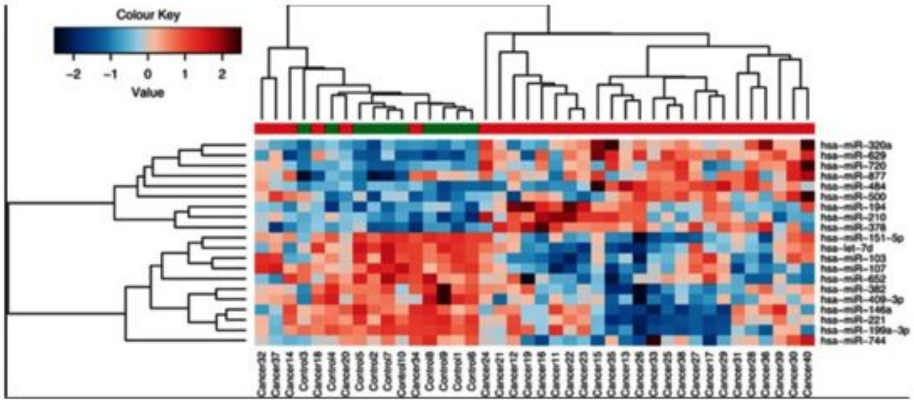
- Utility
  - **Distributions** similar (Similar marginals and multivariate dependencies)
  - **Application** results similar

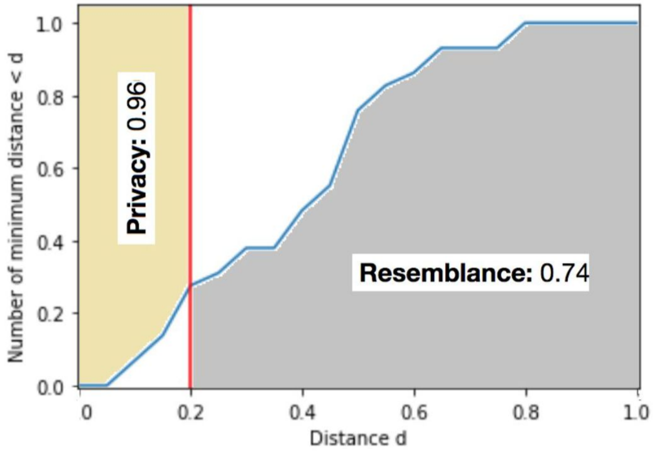
- Privacy
  - **Identity** hidden
  - **Sensitive information** protected

A quick tour

- [ID notebook]







- [Comparison notebook]

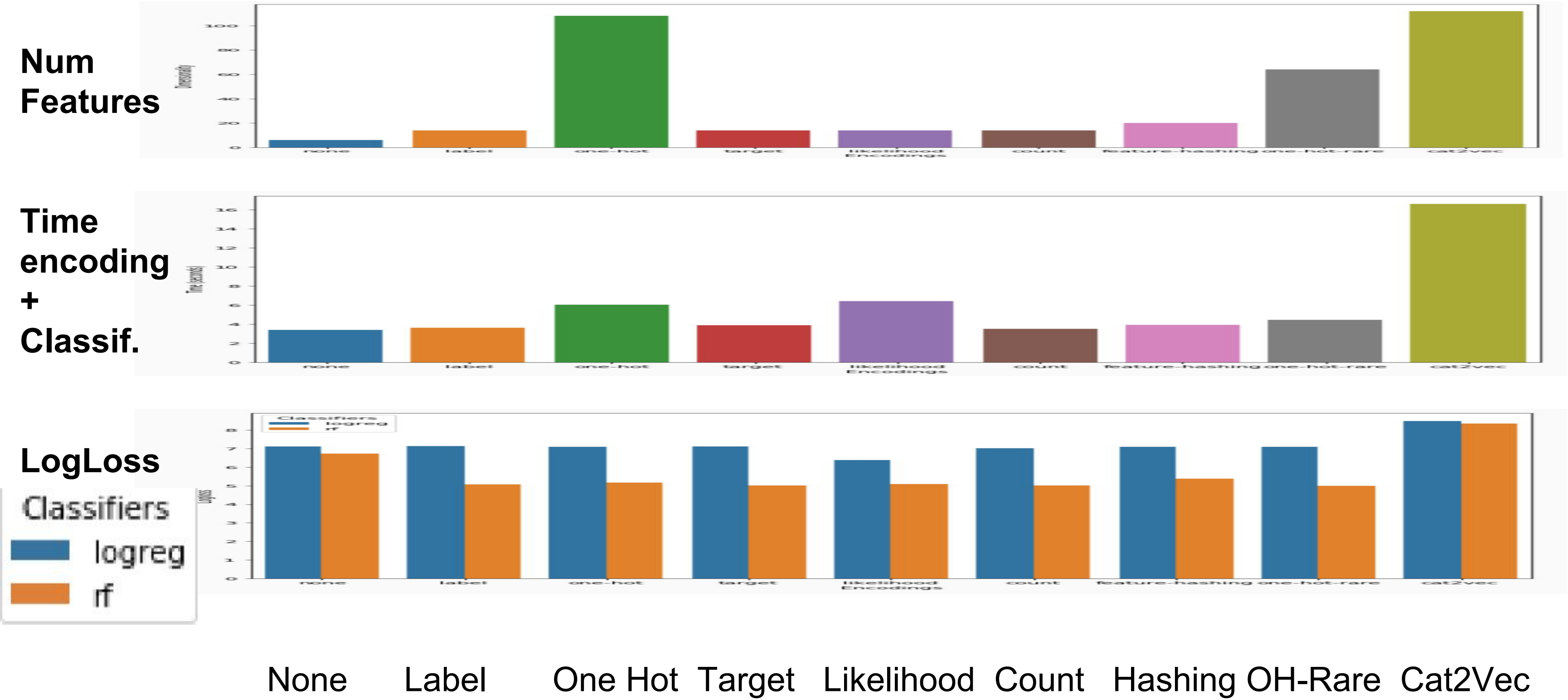
Modeling workflow

1. **Pre-processing:** privacy sanitizing, encoding categorical variables, imputing missing values, etc. [ID notebook]
2. **Modeling**: [Generator library]
   a. **Classical statistics** (e.g. Parzen windows and other kernel methods, multivariate Gaussians).
   b. **Machine Learning** (e.g. imputation of missing values with RF, Generative Adversarial Networks -- GAN, Causal generative networks).
3. **Post-processing:** privacy tuning, marginal distribution back-fitting (Copula inspired); restore categorical variables.
4. **Quality control**: Utility and privacy [Comparison notebook]

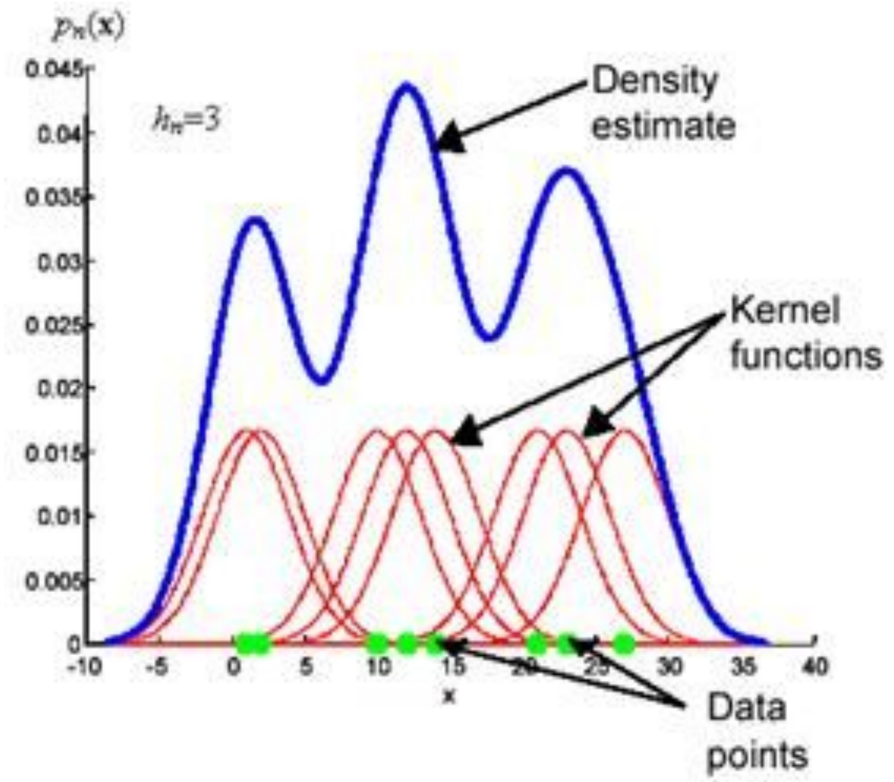
1. **Preprocessing:** Encoding categorical variables

- None (remove categorical variables)
- Label (arbitrary numerical value)
- One hot (binary encoding)
- Feature hashing
- Mean target value
- Likelihood
- Frequency
- Cat2vec: DL embedding

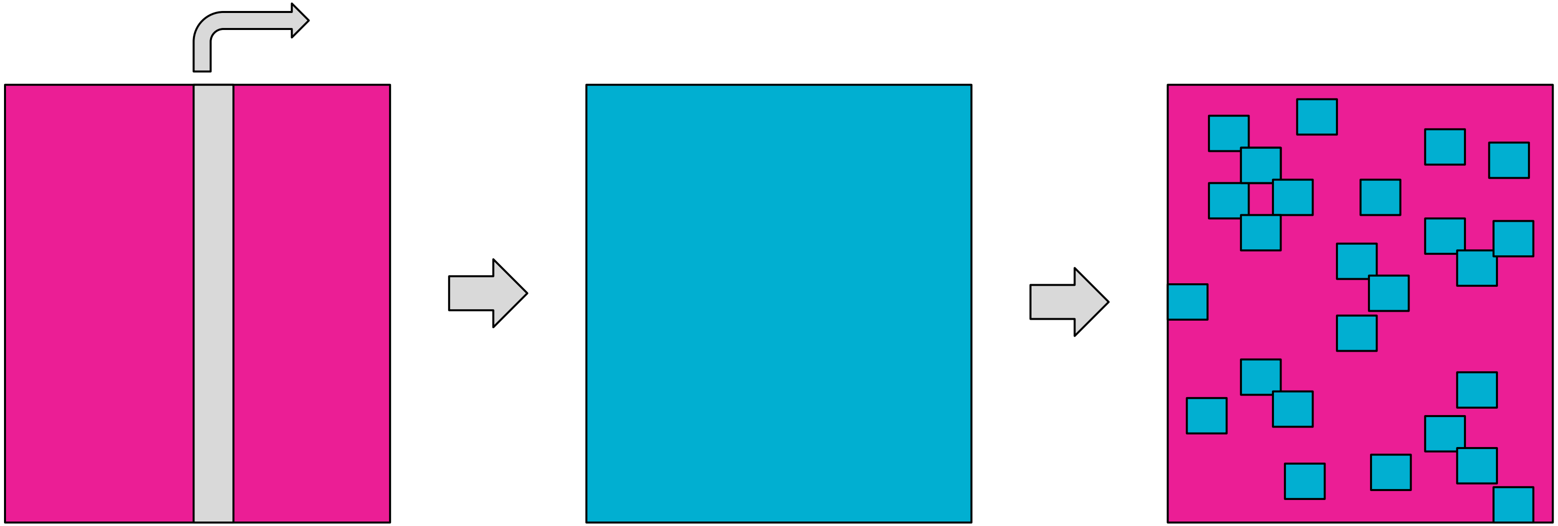
Categorical variables: Comparison of encodings



2. Modeling: Parzen windows



2. Modeling: Missing values imputation with RF



ORIGINAL DATA

(1) Build predictors of one column from the others..

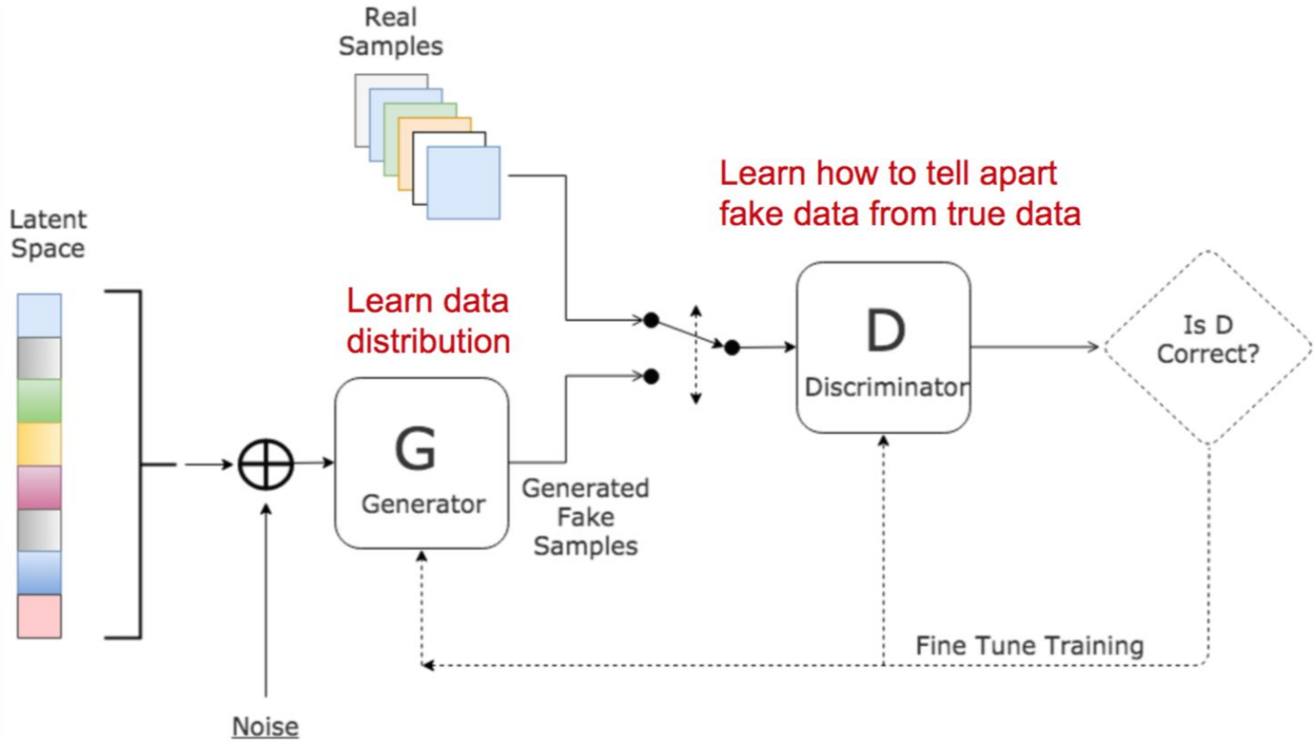
PREDICTED DATA

(2) Make a matrix of predictions.

SYNTHETIC DATA

(3) Mix original and predicted data randomly in a certain proportion p.

2. Modeling: GAN (Generative Adversarial Networks)



Two flavors: MedGAN (Andrew) and SAM (Diviyan)

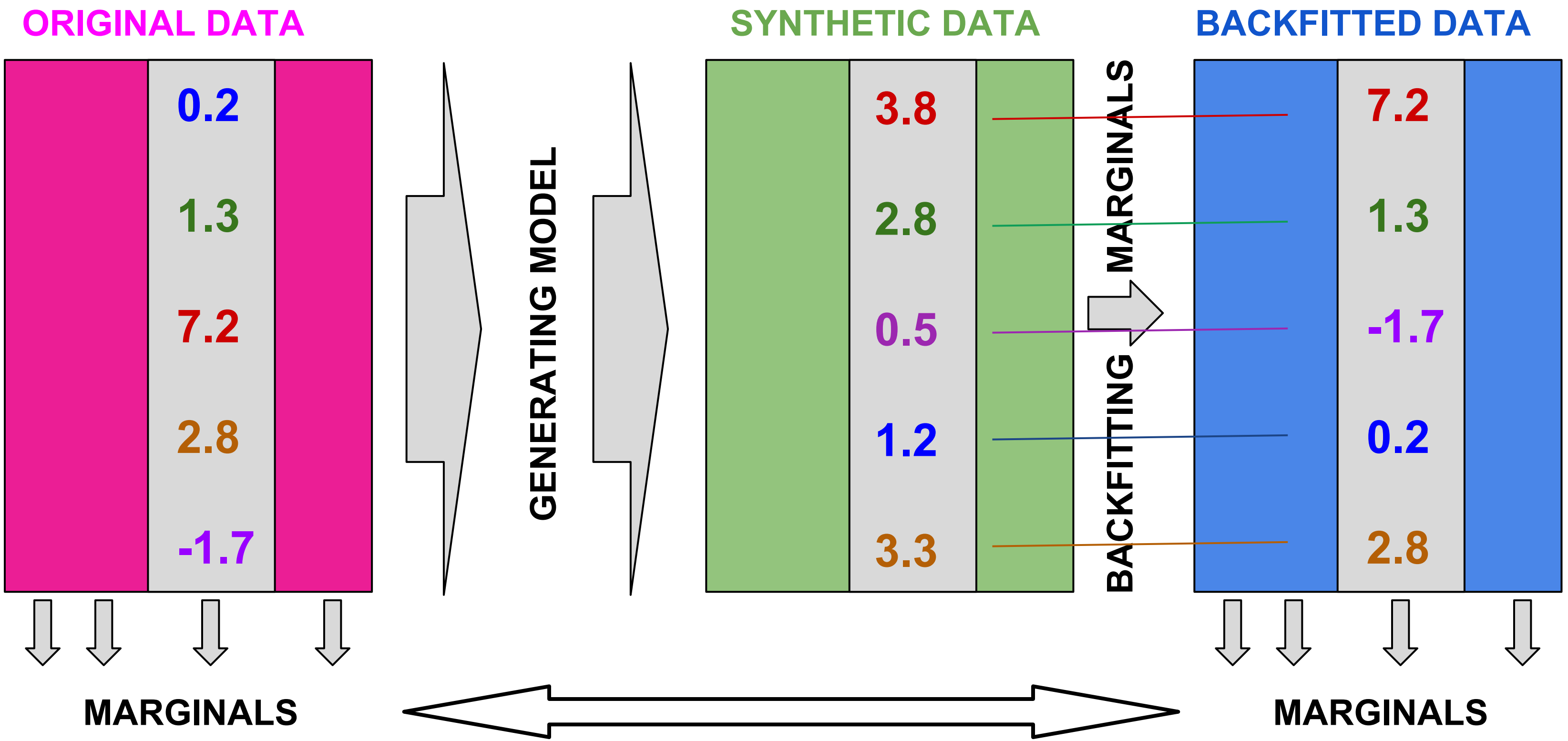
2. Modeling: Copula

Copula = multivariate distribution with uniform marginals

Sklar's theorem = Every multivariate distribution can be expressed in terms of its marginals and a copula.

Procedure Copula modeling:
> Make marginals uniform (replace variables by their rank)
> Model the distribution
> Back-fit the marginals to the original marginals

3. Post-processing: Backfitting marginals (Copula trick)

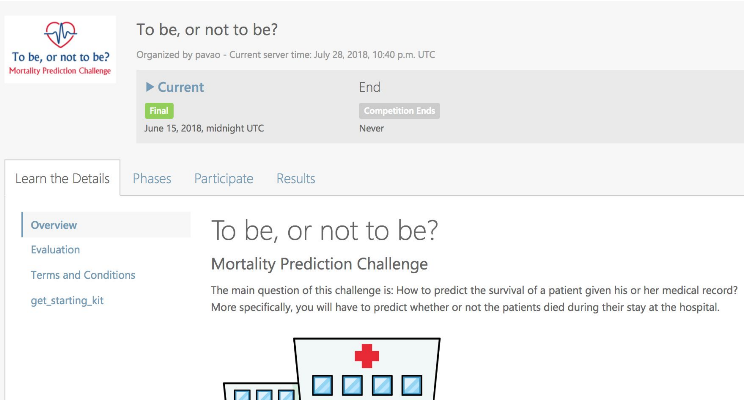


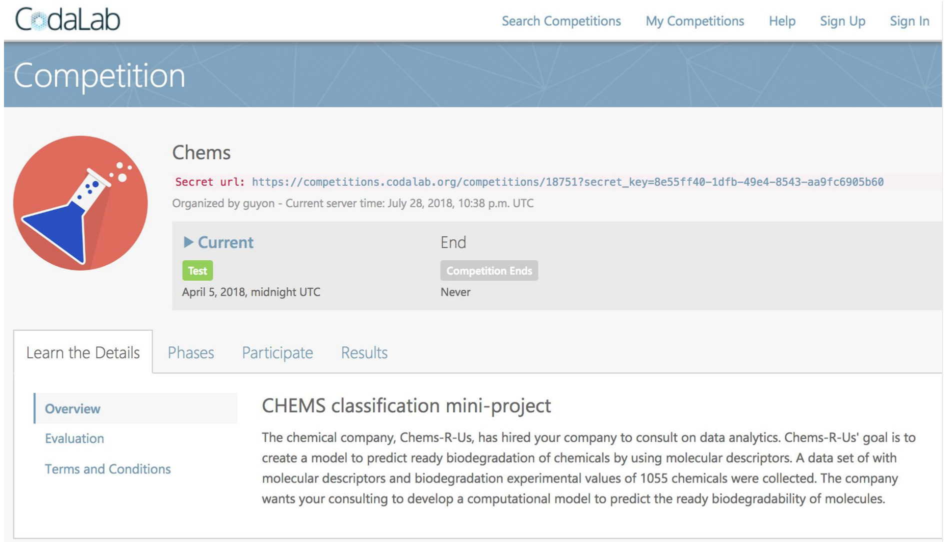
On-going work

Systematic study to compare **Utility** and **Privacy**.

Methods / Datasets	Pazen (or other kernel method)	Gaussian multivariate	RF multiple imputation	GAN(s)
Iris				
Boston housing				
Adult (census)				
Mimic				

Challenge organization





- Chems: Predict ready biodegradation of chemicals by using molecular descriptors.
- Mortality: Predict the survival of a patient given his or her medical record using synthetic MIMIC data.
- Chems 2: with feature selection (in preparation)
- Survival analysis (in preparation)

Conclusion

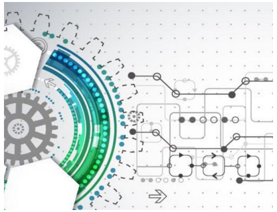
- With our synthetic data we already started

training students in health data analytics

- We are working on:

- improving data quality and

- designing ML challenges for students and researchers.



- Using synthetic data for discovery is further

  down the road: Utility / Privacy tradeoff.