



**Synthetic Medical Data Generation**

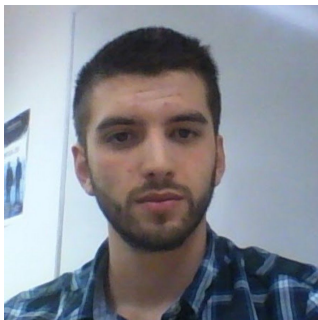
**Isabelle Guyon, UPSud/INRIA & ChaLearn**



# The Team



Andrew Yale  
RPI, NY  
(PhD student)



Adrien Pavao  
UPSud, Paris  
(master student)



Thomas  
Gespacher  
ENSIMAG  
Grenoble  
(master 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



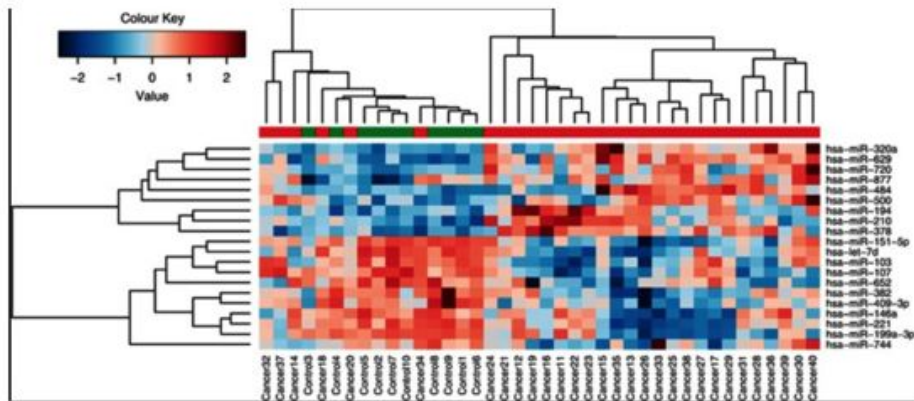
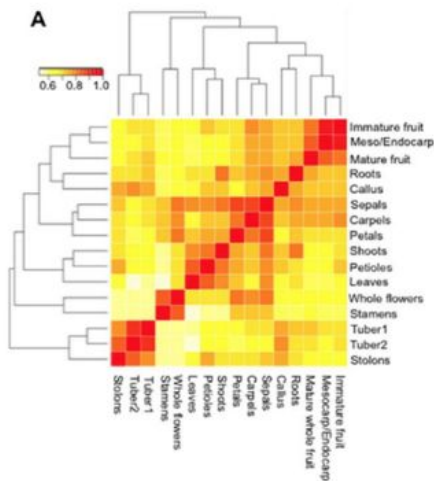
- Privacy

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

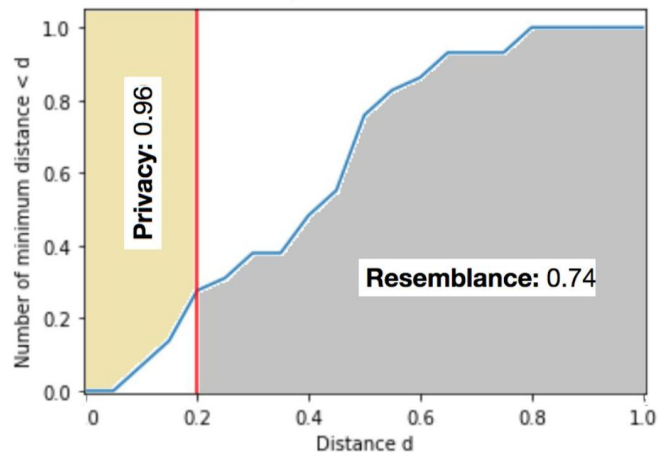


# A quick tour

- ID notebook



- Comparison notebook

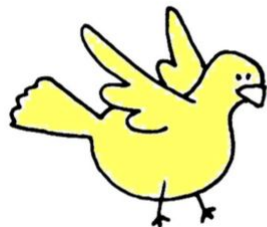


# Modeling workflow

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

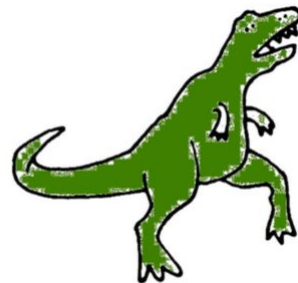
- None (remove categorical variables)
- Label (arbitrary numerical value)
- One hot (binary encoding)
- Feature hashing
- Mean target value
- Likelihood
- Frequency
- Cat2vec: DL embedding



I am a bird.  
I am yellow.  
I am awesome.



I am a seahorse.  
I am orange.  
I am super awesome.

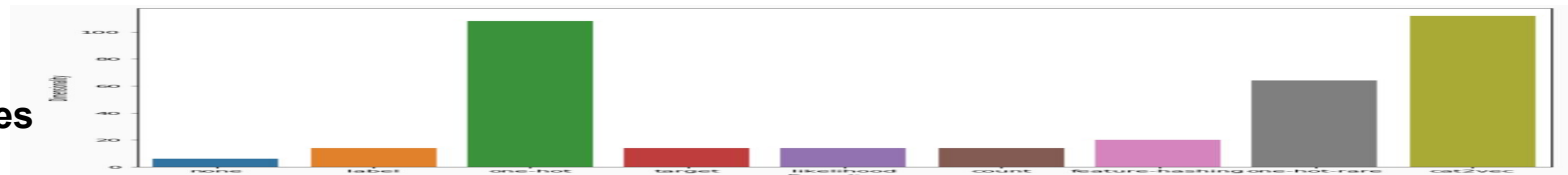


I am a T-rex.  
I am green.  
I am extinct.

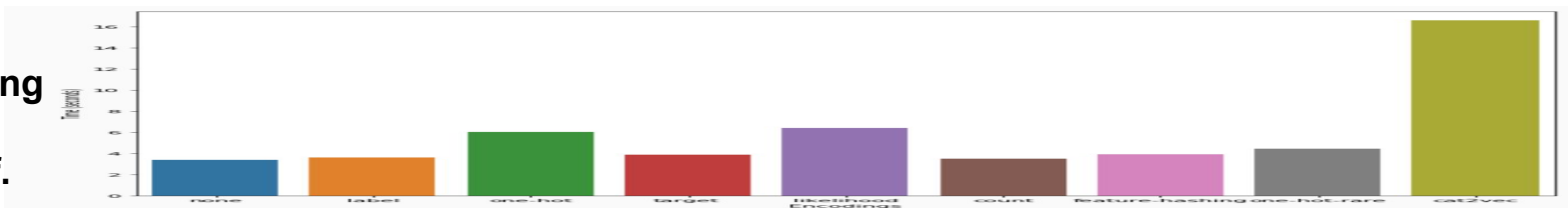


# Categorical variables: Comparison of encodings

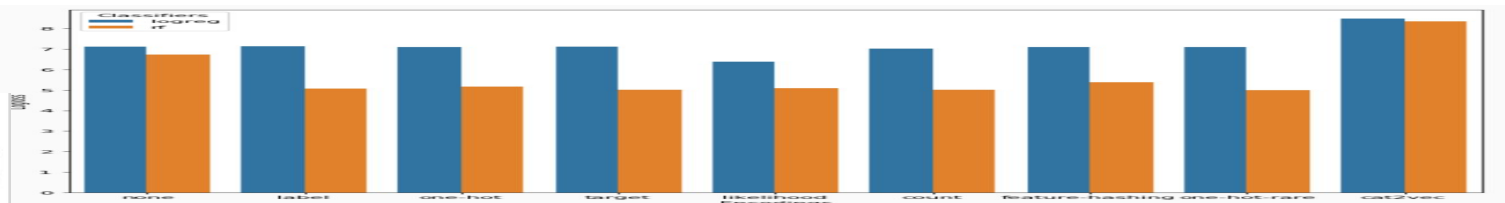
Num  
Features



Time  
encoding  
+  
Classif.

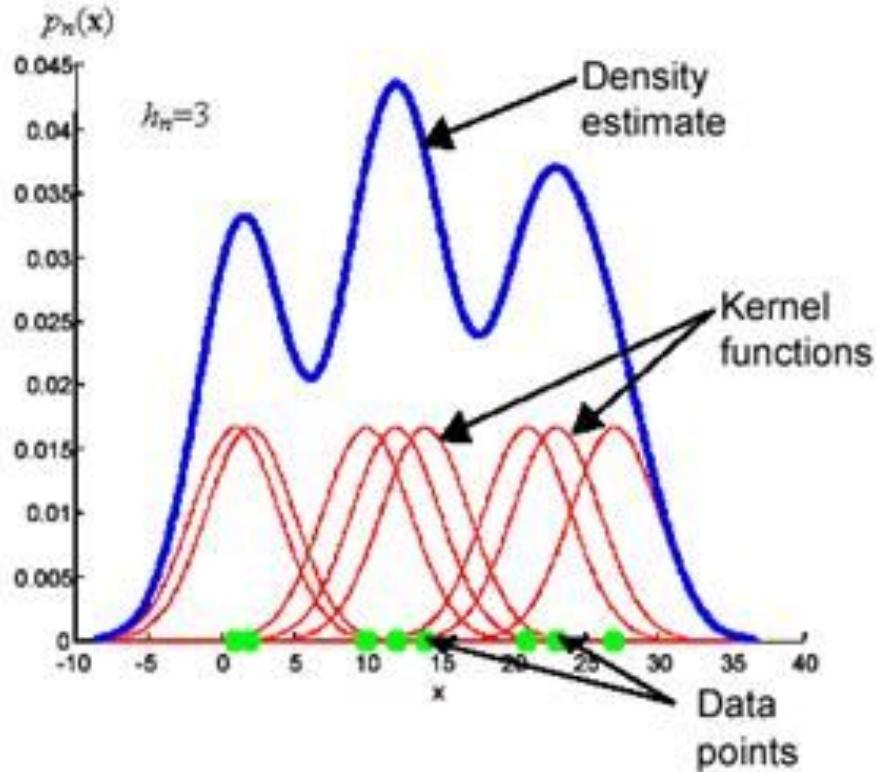


LogLoss  
Classifiers

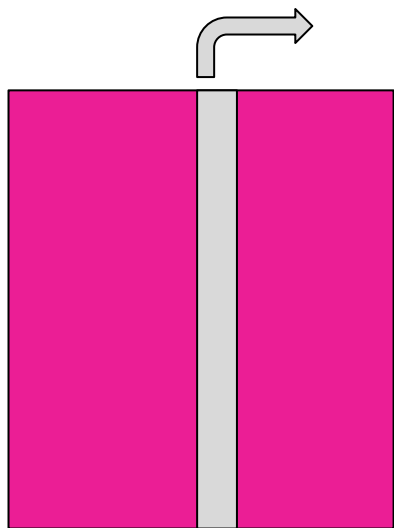


None    Label    One Hot    Target    Likelihood    Count    Hashing OH-Rare    Cat2Vec

## 2. Modeling: Parzen windows

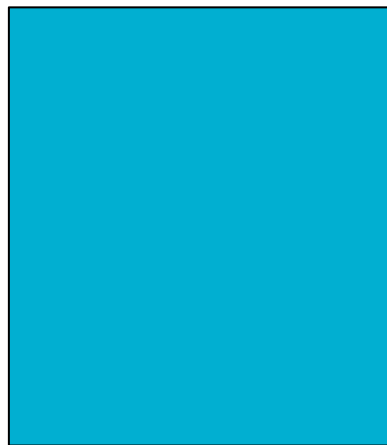


## 2. Modeling: Missing values imputation with RF



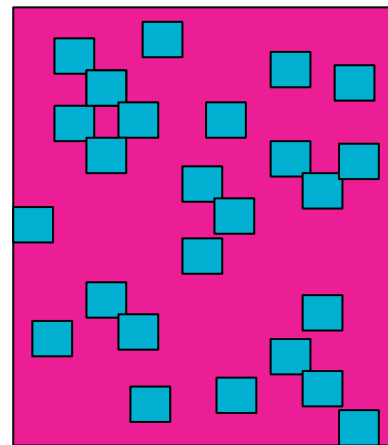
**ORIGINAL DATA**

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



**PREDICTED DATA**

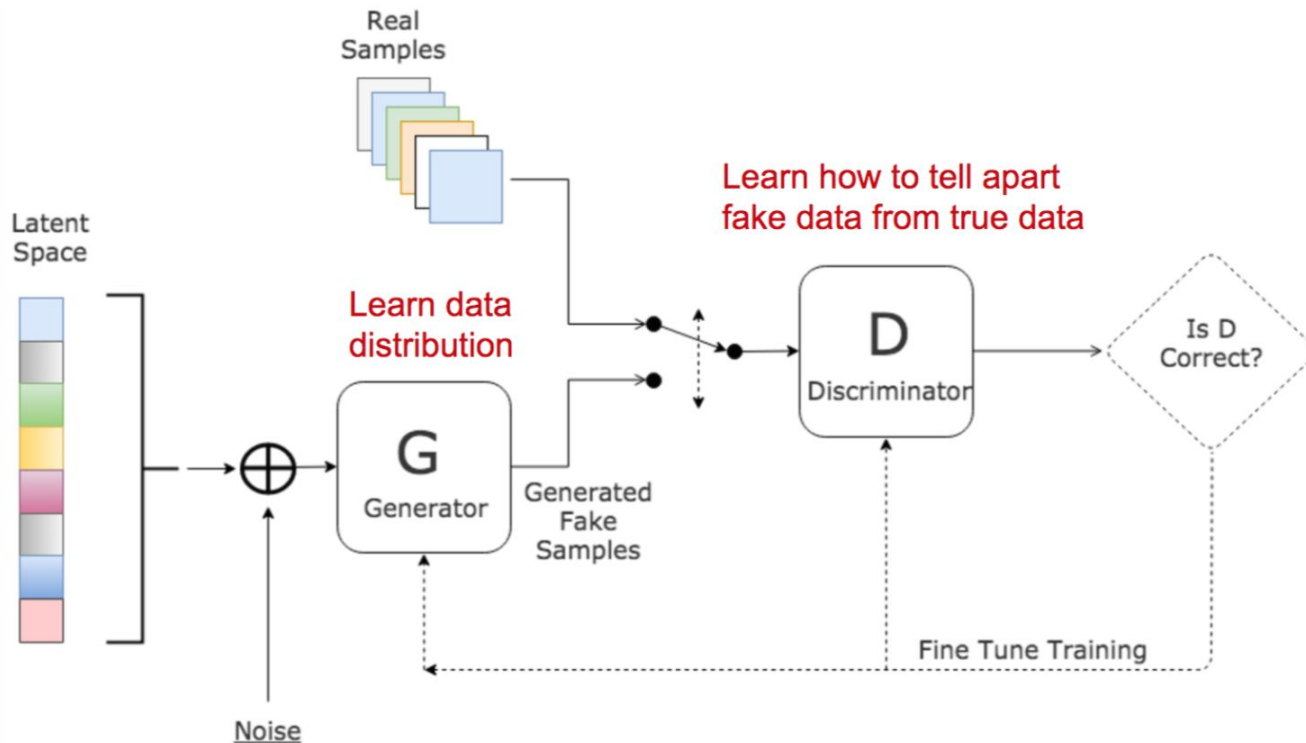
(2) Make a matrix of predictions.



**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

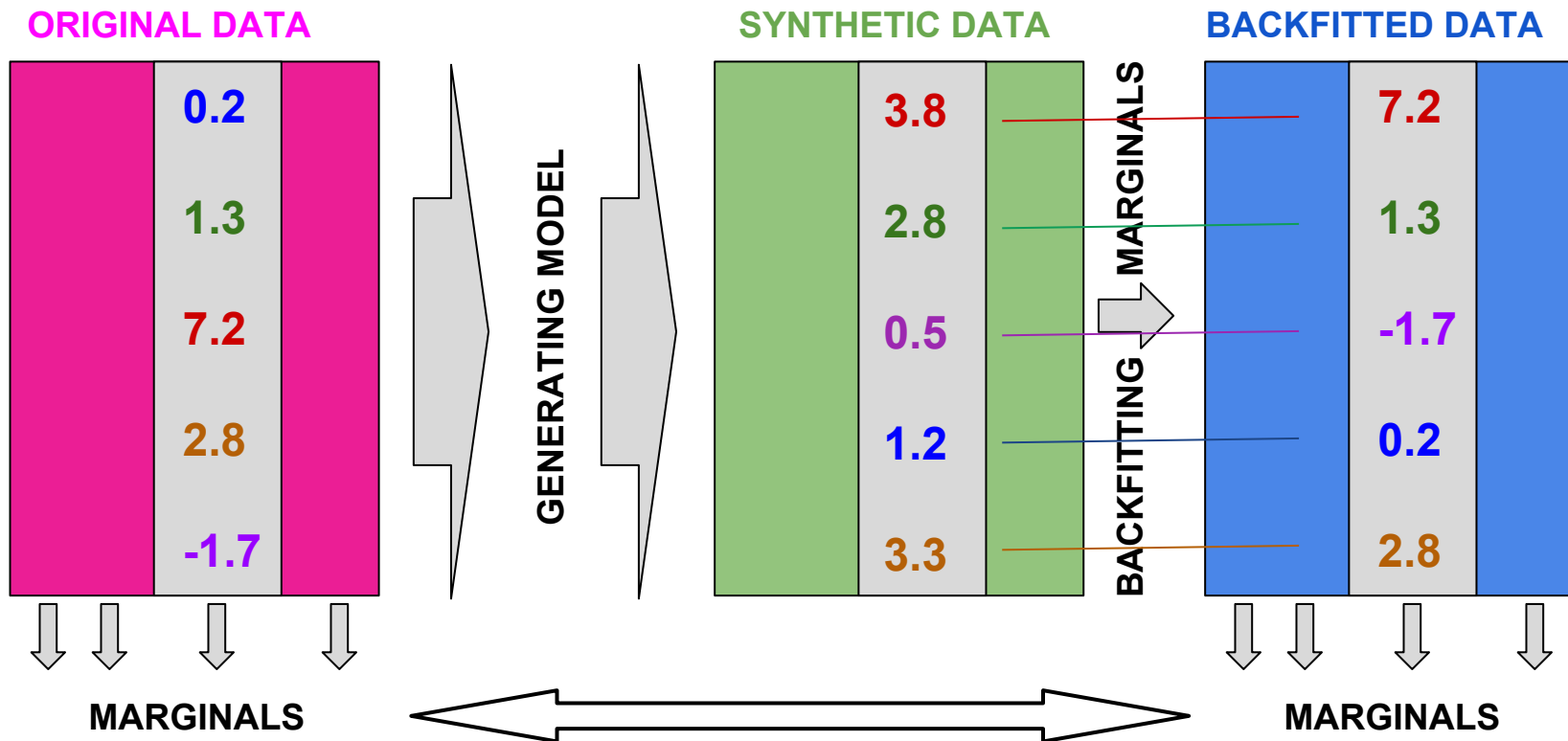
**Copula** = multivariate distribution with uniform marginals

**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)



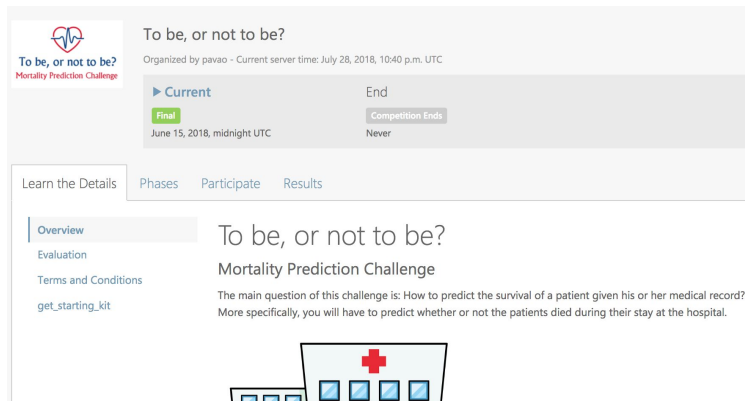
# On-going work



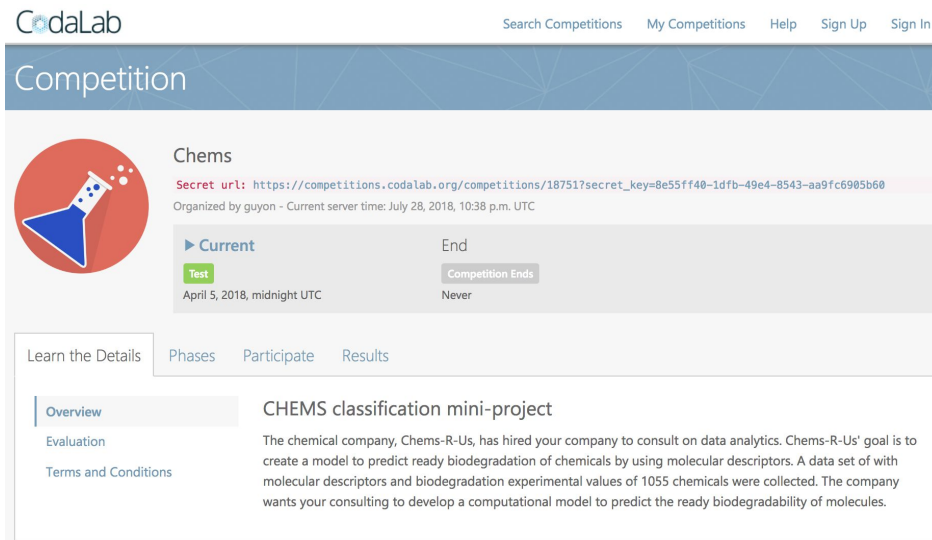
Systematic study to compare **Utility** and **Privacy**.

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

# Challenge organization



The screenshot shows the 'To be, or not to be?' challenge page. At the top, it says 'To be, or not to be? Mortality Prediction Challenge' with a heart icon. Below this, it states 'Organized by pavao - Current server time: July 28, 2018, 10:40 p.m. UTC'. A progress bar shows 'Current' and 'End' phases. The 'Current' phase is active, with a 'Final' button and a 'Competition Ends' button. The 'Final' button is green and says 'June 15, 2018, midnight UTC'. The 'Competition Ends' button is grey and says 'Never'. Below the progress bar, there are tabs: 'Learn the Details', 'Phases', 'Participate', and 'Results'. The 'Learn the Details' tab is selected, showing a sidebar with 'Overview', 'Evaluation', 'Terms and Conditions', and 'get\_starting\_kit'. The main content area shows the challenge title 'To be, or not to be?' and 'Mortality Prediction Challenge'. Below this, it says 'The main question of this challenge is: How to predict the survival of a patient given his or her medical record? More specifically, you will have to predict whether or not the patients died during their stay at the hospital.' At the bottom, there is a small icon of a hospital building with a red cross.



The screenshot shows the 'Chems' challenge page. At the top, it says 'Chems' with a flask icon. Below this, it states 'Secret url: https://competitions.codalab.org/competitions/18751?secret\_key=Be55ff40-1dfb-49e4-8543-aa9fc6985b60' and 'Organized by guyon - Current server time: July 28, 2018, 10:38 p.m. UTC'. A progress bar shows 'Current' and 'End' phases. The 'Current' phase is active, with a 'Test' button and a 'Competition Ends' button. The 'Test' button is green and says 'April 5, 2018, midnight UTC'. The 'Competition Ends' button is grey and says 'Never'. Below the progress bar, there are tabs: 'Learn the Details', 'Phases', 'Participate', and 'Results'. The 'Learn the Details' tab is selected, showing a sidebar with 'Overview', 'Evaluation', and 'Terms and Conditions'. The main content area shows the challenge title 'Chems' and 'CHEMS classification mini-project'. Below this, it says 'The chemical company, Chems-R-U, has hired your company to consult on data analytics. Chems-R-U's goal is to create a model to predict ready biodegradation of chemicals by using molecular descriptors. A data set of with molecular descriptors and biodegradation experimental values of 1055 chemicals were collected. The company wants your consulting to develop a computational model to predict the ready biodegradability of molecules.'

- 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.

