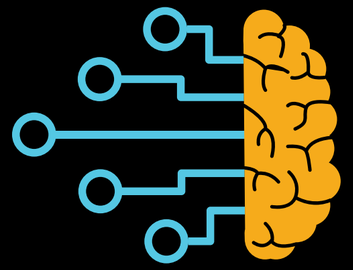
APRENDIZAJE AUTOMÁTICO.

Fundamentos y Aplicaciones en Meteorología del Espacio

DCAO, UBA
08-12 AGOSTO 2022



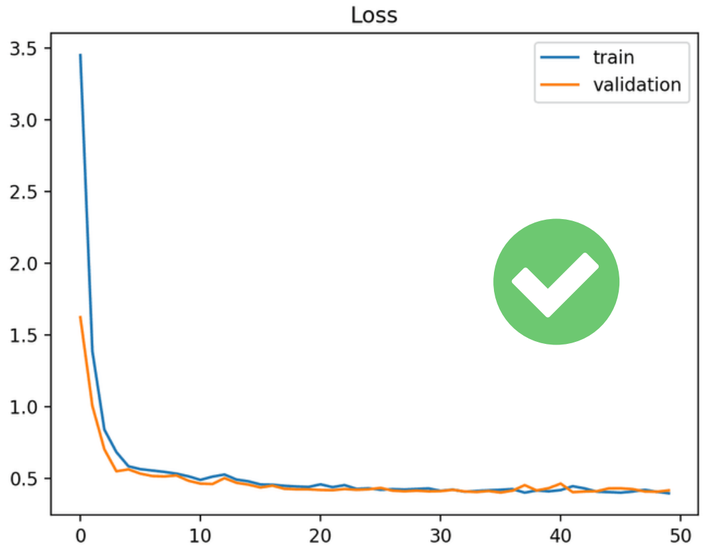
Dra María Graciela Molina
FACET-UNT / CONICET
Tucumán Space Weather Center - TSWC

https://spaceweather.facet.unt.edu.ar/
IG -> @spaceweatherargentina

gmolina@herrera.unt.edu.ar

Diagnose

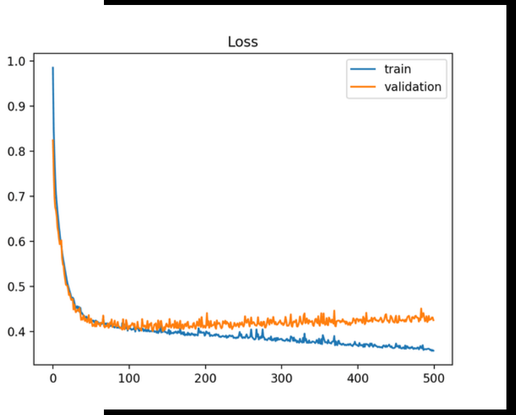
- The learning process



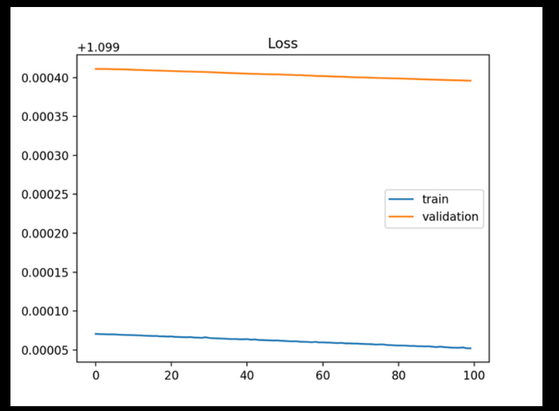
reallity

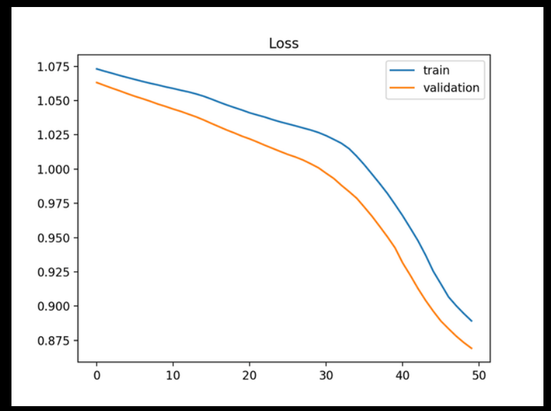
TSWC, 2022

overfitting



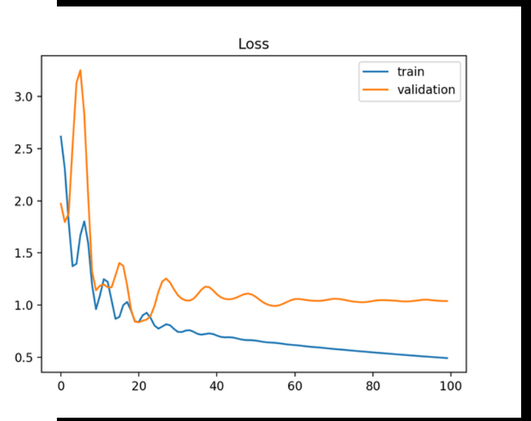
Model does not have a suitable capacity for the complexity of the dataset (underfit)





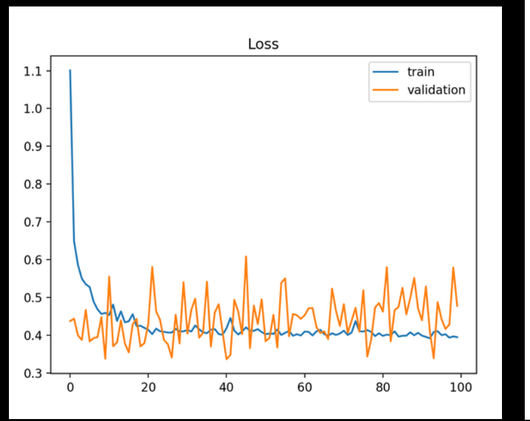
Model that requires further training (underfit)

unrepresentative training dataset

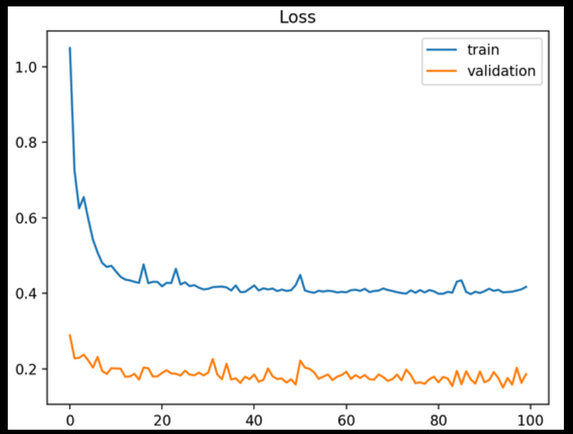


Dataset << relative to the validation dataset

unrepresentative validation dataset

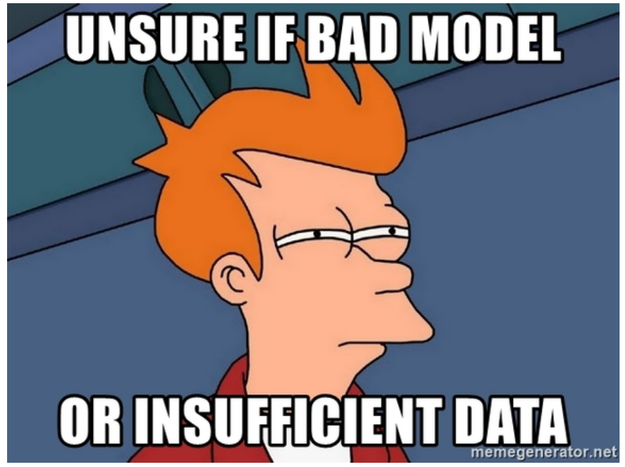


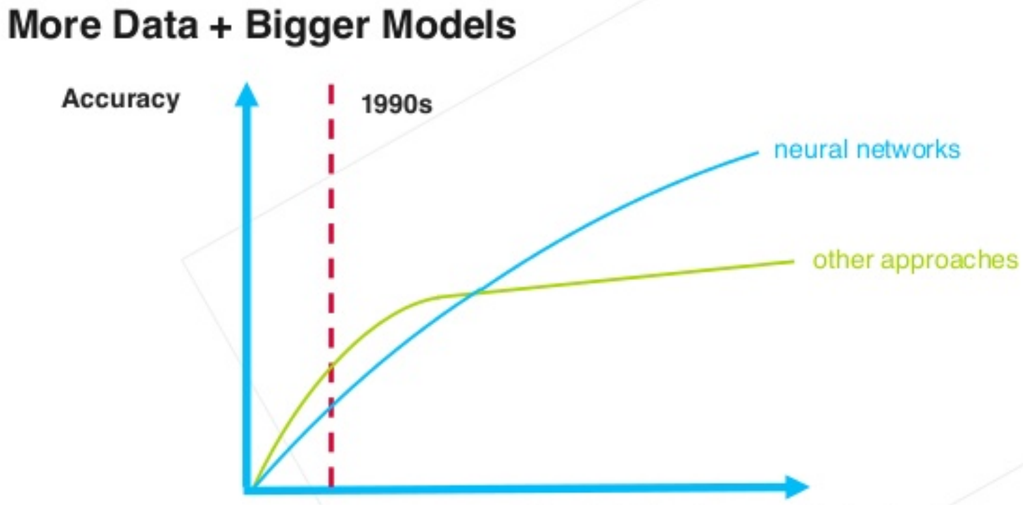
validation dataset does not provide sufficient information to evaluate the ability of the model to generalize

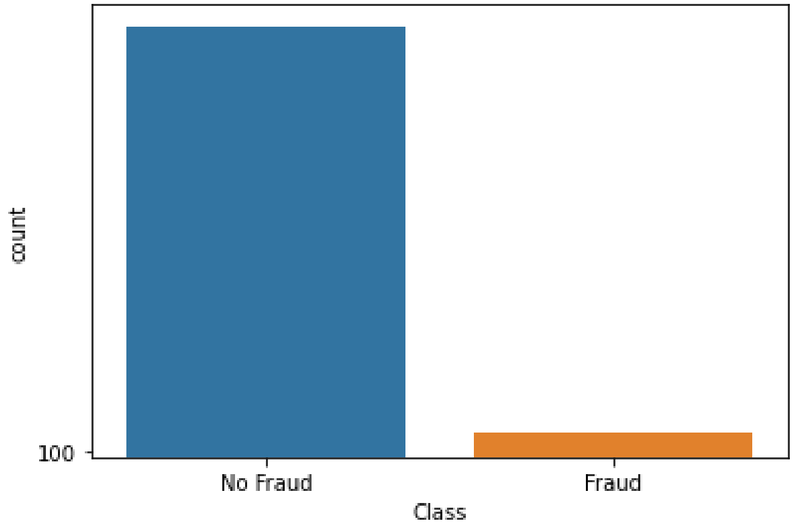


About the dataset

- Amount of data and data balancing





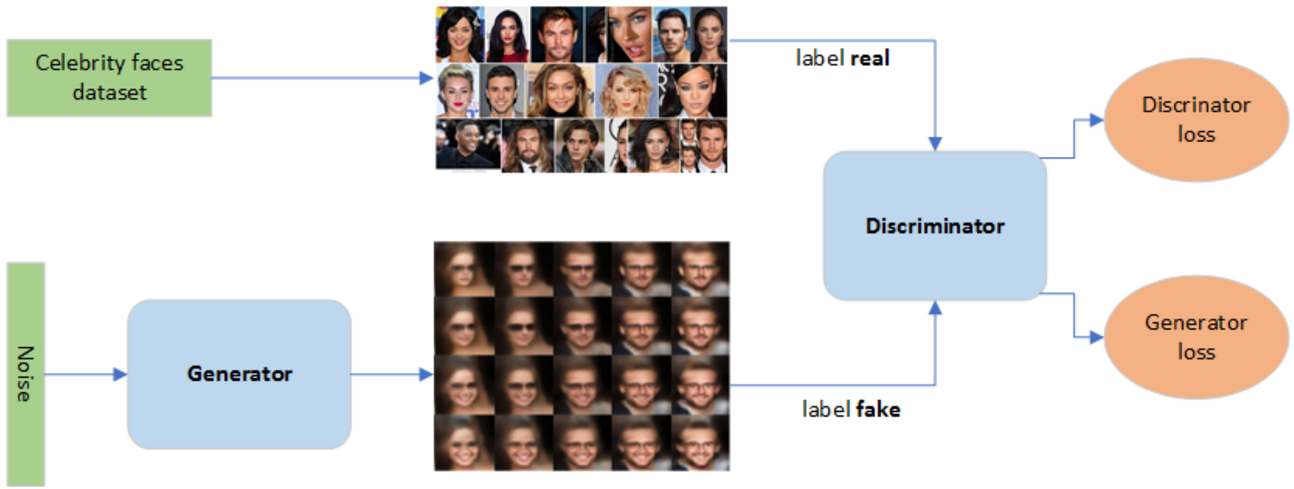


re sample (?)
generate more data (?)

About the dataset

Bertimas et.al., From Predictive Methods to Missing Data Imputation: An Optimization Approach (2018)

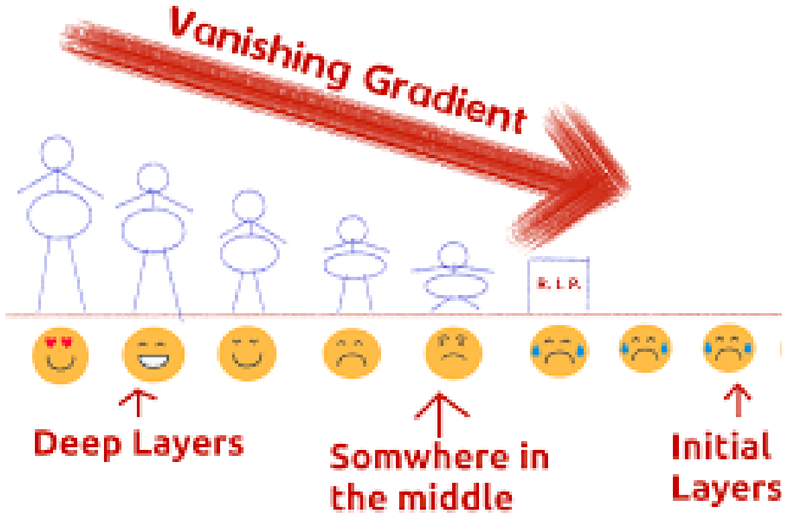
Method Name	Category	Software	Reference
Mean impute (`mean`)	Mean		Little and Rubin (1987)
Expectation-Maximization (`EM`)	EM		Dempster et al. (1977)
EM with Mixture of Gaussians and Multinomials	EM		Ghahramani and Jordan (1994)
EM with Bootstrapping	EM	`Amelia II`	Honaker et al. (2011)
K-Nearest Neighbors (`knn`)	K-NN	`impute`	Troyanskaya et al. (2001)
Sequential K-Nearest Neighbors	K-NN		Kim et al. (2004)
Iterative K-Nearest Neighbors	K-NN		Caruana (2001); Brás and Menezes (2007)
Support Vector Regression	SVR		Wang et al. (2006)
Predictive-Mean Matching (`pmm`)	LS	`MICE`	Buuren and Groothuis-Oudshoorn (2011)
Least Squares	LS		Bø et al. (2004)
Sequential Regression Multivariate Imputation	LS		Raghunathan et al. (2001)
Local-Least Squares	LS		Kim et al. (2005)
Sequential Local-Least Squares	LS		Zhang et al. (2008)
Iterative Local-Least Squares	LS		Cai et al. (2006)
Sequential Regression Trees	Tree	`MICE`	Burgette and Reiter (2010)
Sequential Random Forest	Tree	`missForest`	Stekhoven and Bühlmann (2012)
Singular Value Decomposition	SVD		Troyanskaya et al. (2001)
Bayesian Principal Component Analysis	SVD	`pcaMethods`	Oba et al. (2003); Mohamed et al. (2009)
Factor Analysis Model for Mixed Data	FA		Khan et al. (2010)



Goodfellow et.al. Generative Adversarial Networks (2014)

Vanishing gradient problem

- activation function, architecture, weight initialization, loss optimization algorithm, learning rate, ...



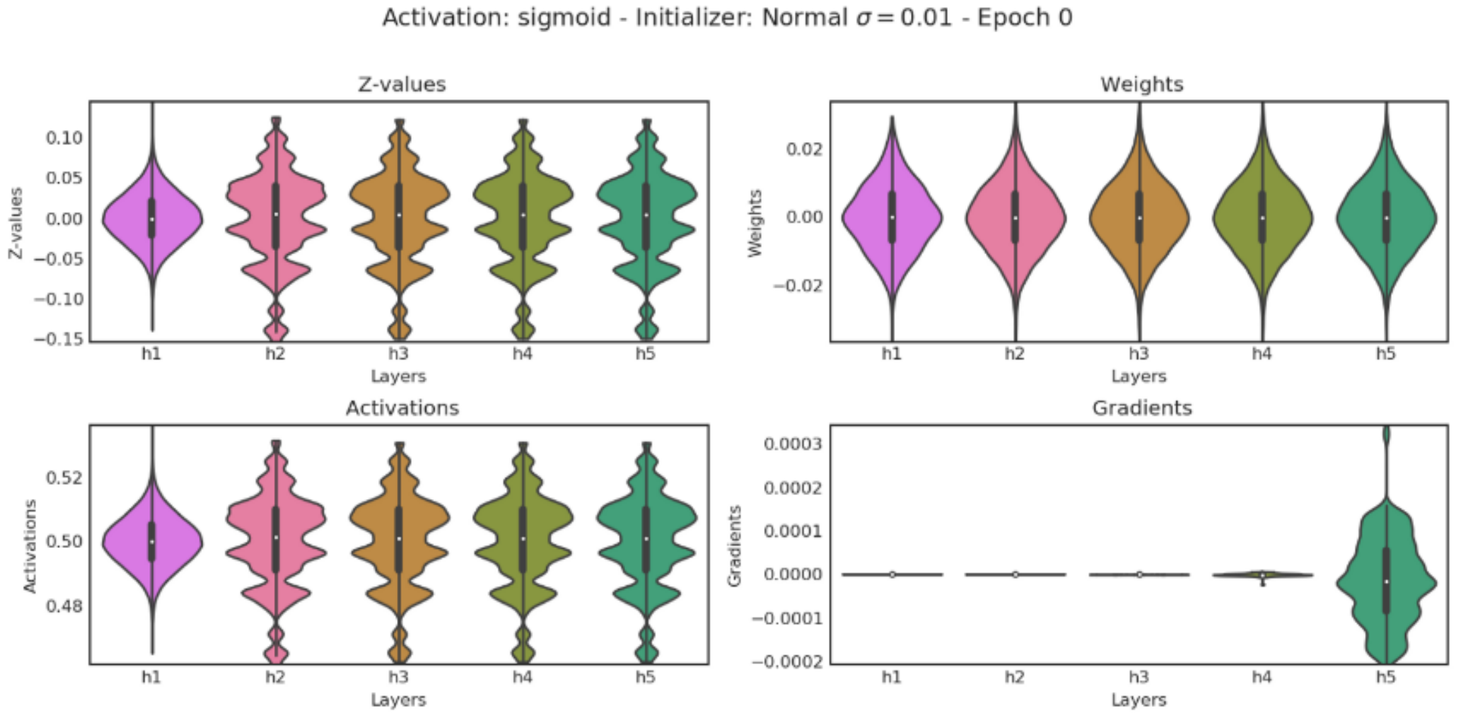


Figure 1. BLOCK model using sigmoid and naive initialization — don't try this at home!

https://towardsdatascience.com/hyper-parameters-in-action-part-ii-weight-initializers-35aee1a28404

Hyperparameters optimization

- Choose a set of optimal hyperparameters for a learning algorithm (maximizes the model performance)
- Hyperparameters are set before the learning process (#neurons, #cells, loss optimization algorithms, etc)

GRID SEARCH

1 —Identify the model's hyperparameters to optimizest.

2 — Asses error score for each combination in the hyperparameter grid.

3 — Select the hyperparameter combination with the best error metric.

Hyperparameters optimization

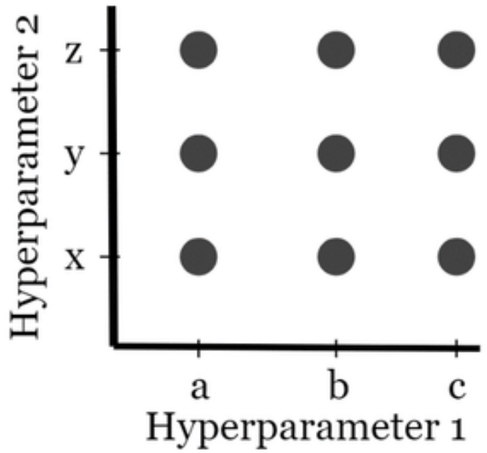
```
# learning_rate choices
 learning_rates = [ 0.1, 0.2, 0.3, 0.4, 0.5,
                    0.01, 0.02, 0.03, 0.04, 0.05 ]
 # iterations choices
 iterations = [ 100, 200, 300, 400, 500 ]

 parameters = []
 for i in learning_rates :
     for j in iterations :
         parameters.append( ( i, j ) )

 print("Available combinations : ",  parameters )

 # Applying linear searching in list of available combination
 # to achieved maximum accuracy on CV set

 for k in range( len( parameters ) ) :
   #  model = METHOD(..., learning_rate = parameters[k][0],iterations = parameters[k][1] )
   # ...
```

Hyperparameters optimization

Grid Search

Pseudocode

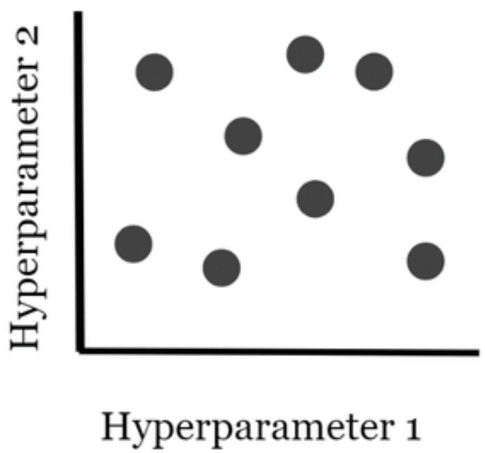
```
Hyperparameter_One = [a, b, c]
Hyperparameter_Two = [x, y, z]
```



Random Search

Pseudocode

```
Hyperparameter_One = random.num(range)
Hyperparameter_Two = random.num(range)
```

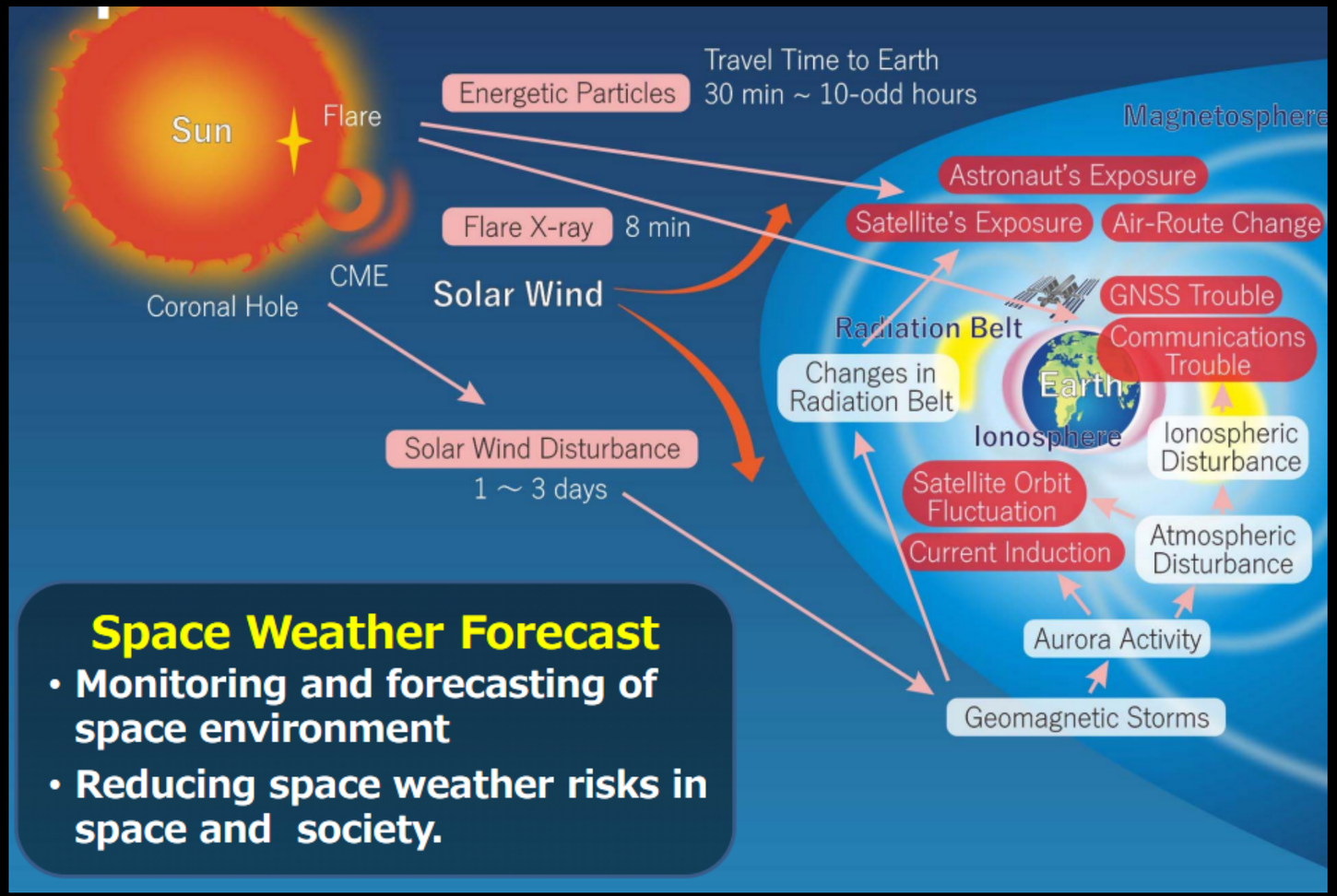


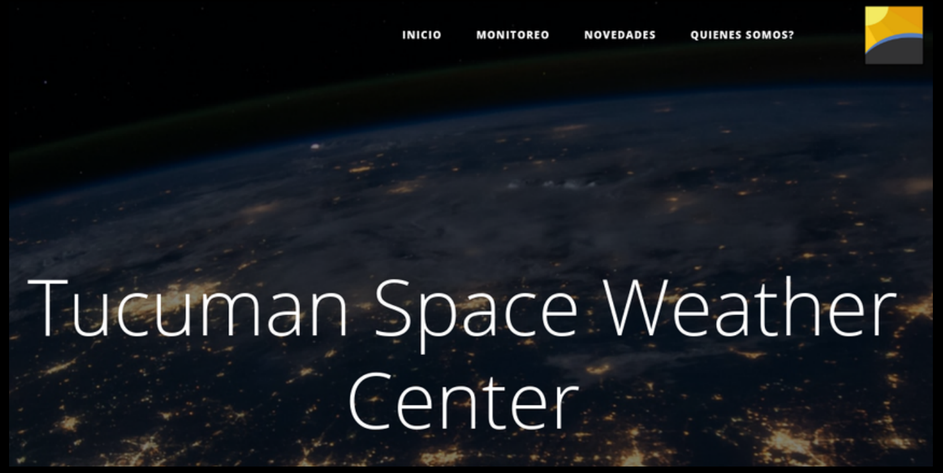
Other methods

- Bayesian Optimization
- Evolutionary Optimization.

https://towardsdatascience.com/hyperparameters-tuning-from-grid-search-to-optimization-a09853e4e9b8

Something we are working on ...





https://spaceweather.facet.unt.edu.ar/
Instagram: /spaceweatherargentina

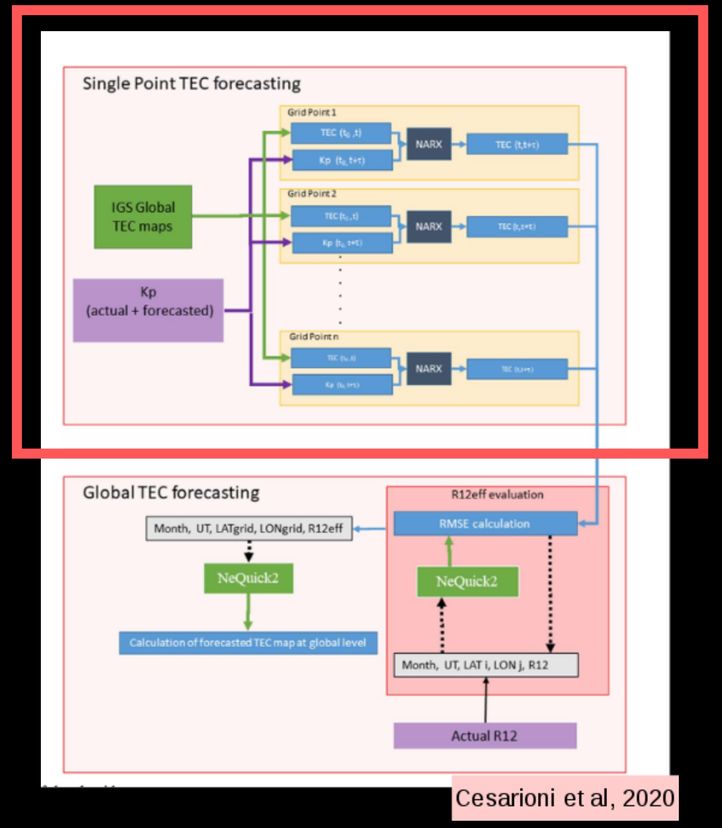
The problem: 24hs forecasting of TEC given information regarding geomagnetic conditions

GLOBAL TEC FORECASTING

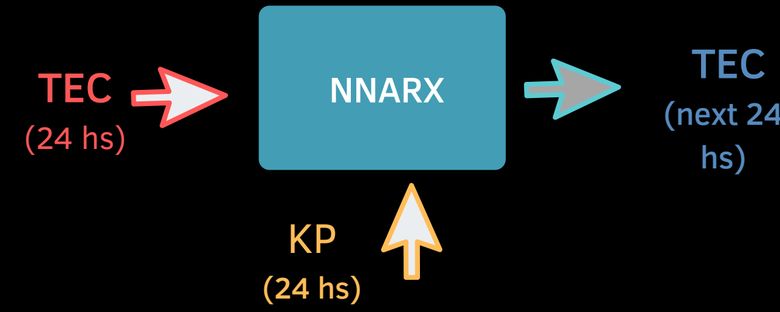
In prep

TSWC, 2022

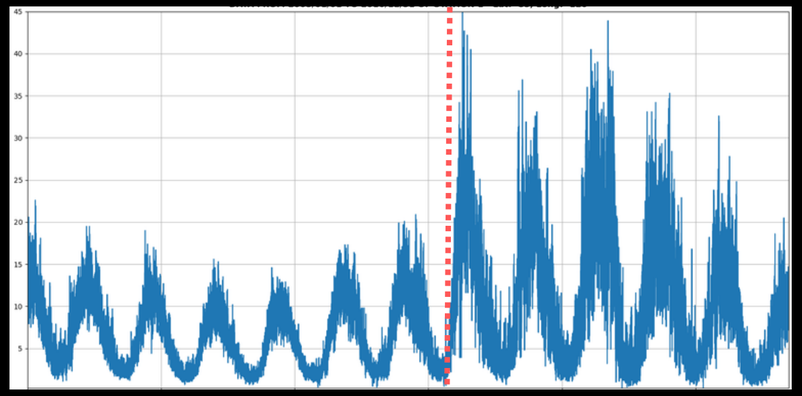
- 24 hs ahead
- Dataset: 2005 - 2016
- **Input**: Global Ionospheric Map (GIM) from IGS. Spatial-temporal res 2.5° (lat) - 5° (lon) - 2 h
- **External forcing** (* SWx): Kp index
- Loosely physics-informed ML

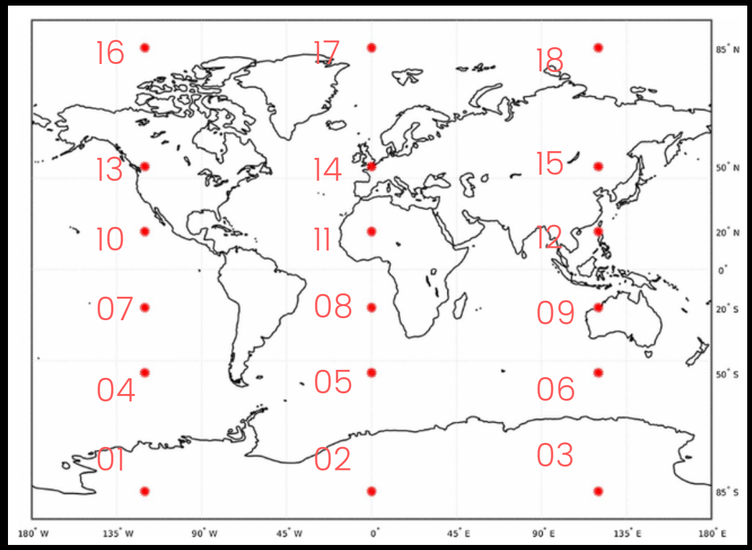


Cesarioni et al, 2020



St 01 - dataset (2005 - 2016)



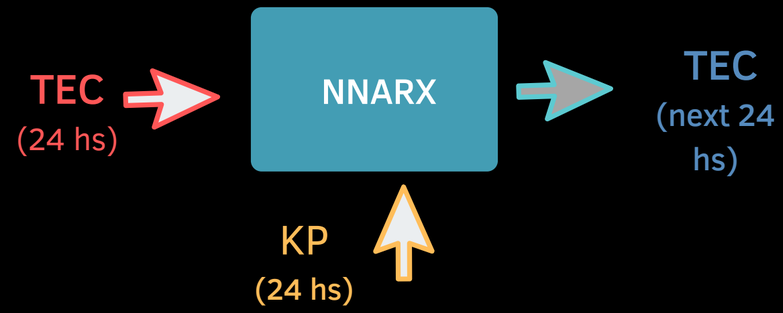


(Molina et al, ESWW 2021)

GLOBAL TEC FORECASTING

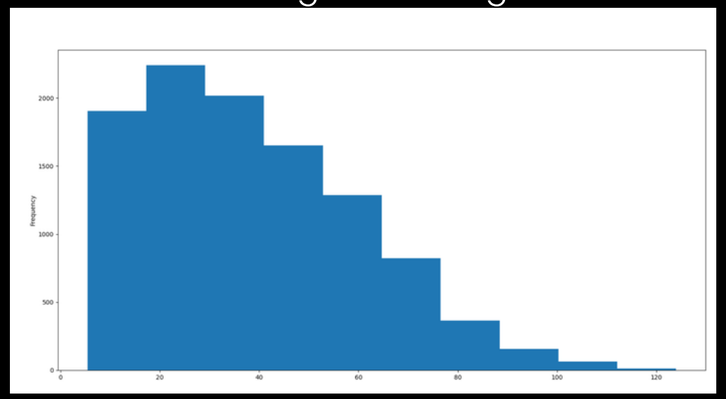
In prep

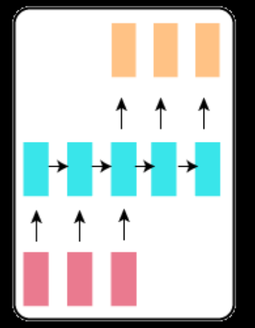
TSWC, 2022



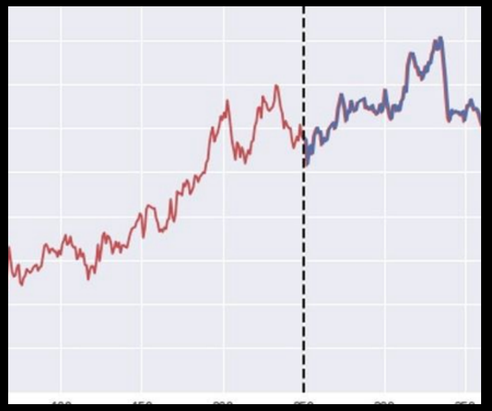
- Train: 99 %(99/1) Test: 1 %
- Re-sampling:
  - GIM 2 hs resolution
  - Kp 3hs resolution > K Nearest-neighbor interpolation
- No missing values
- Kernel initializer: GlorotNormal distribution

TEC - single ST Histogram





- DL modeling
  - 24 hs (before) to forecast 24 hs (ahead) - 24 hs = 12 samples
  - supervised ML
  - 3 methods
  - time-series

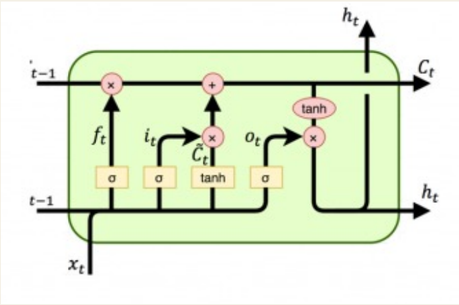


(Molina et al, ESWW 2021)

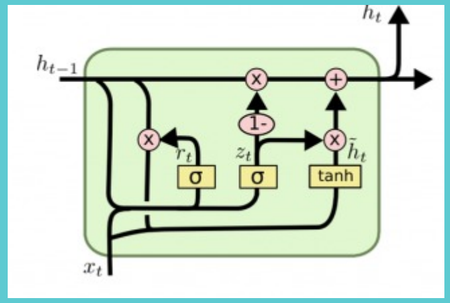
GLOBAL TEC FORECASTING

TSWC, 2022

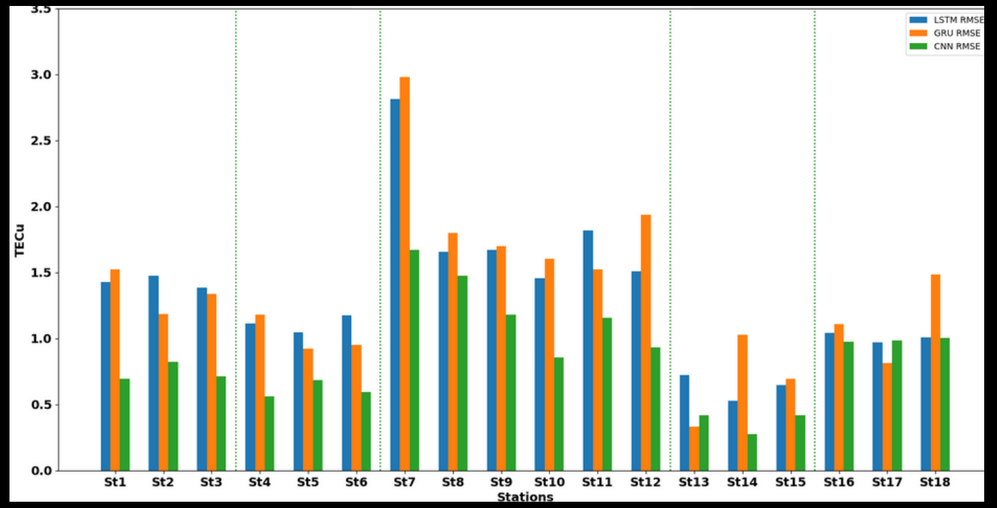
LSTM



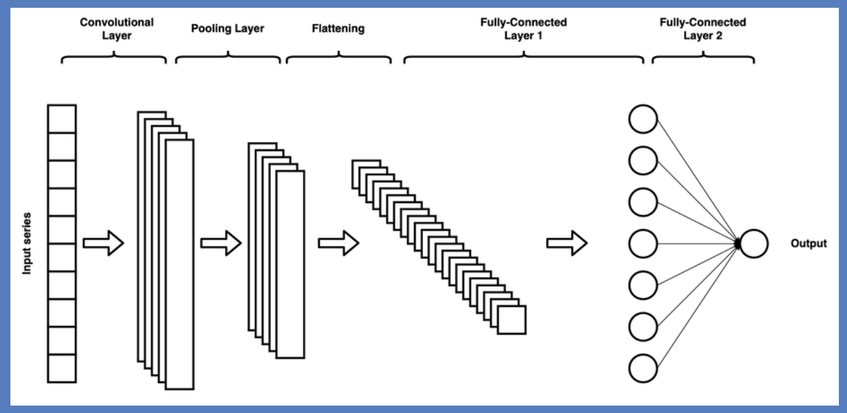
GRU

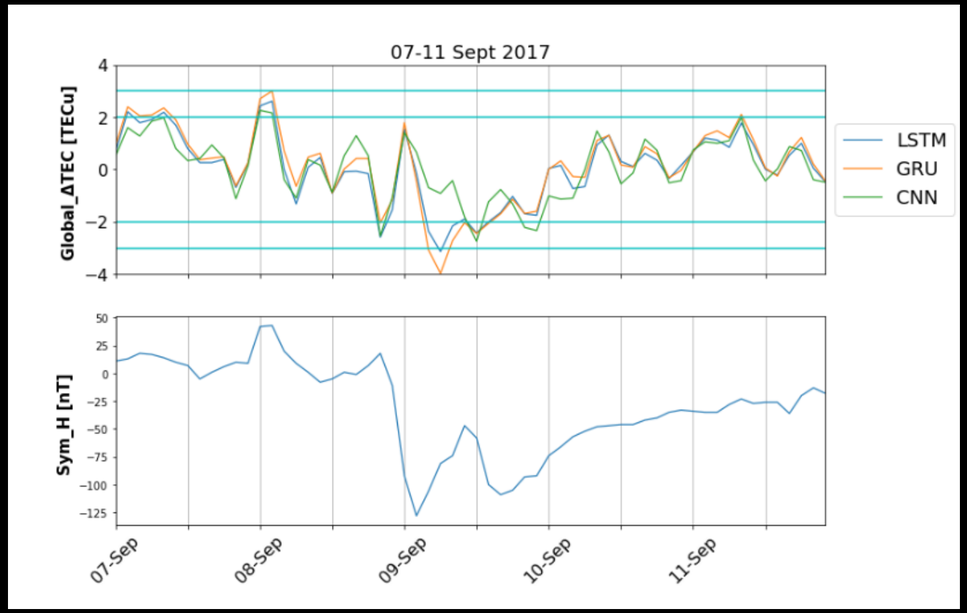


(Molina et al, ESWW 2021)



CNN - 1D



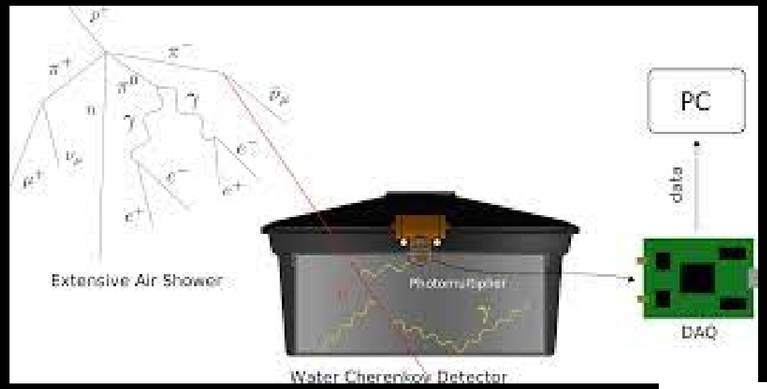


In prep

Ticiano Torres Peralta

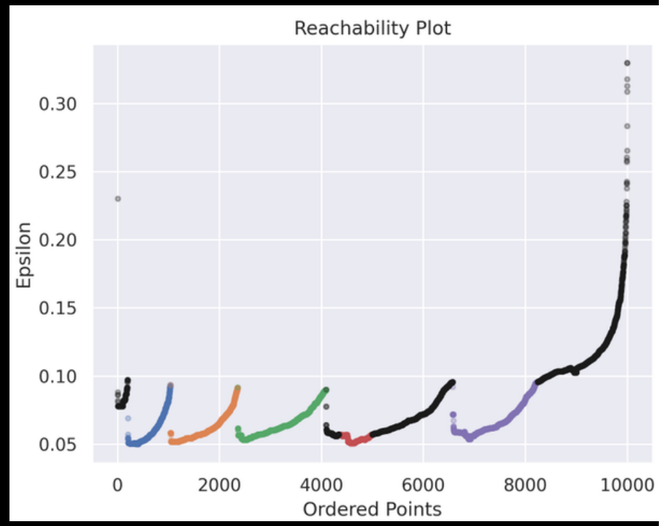
Something (else) we are working on ...

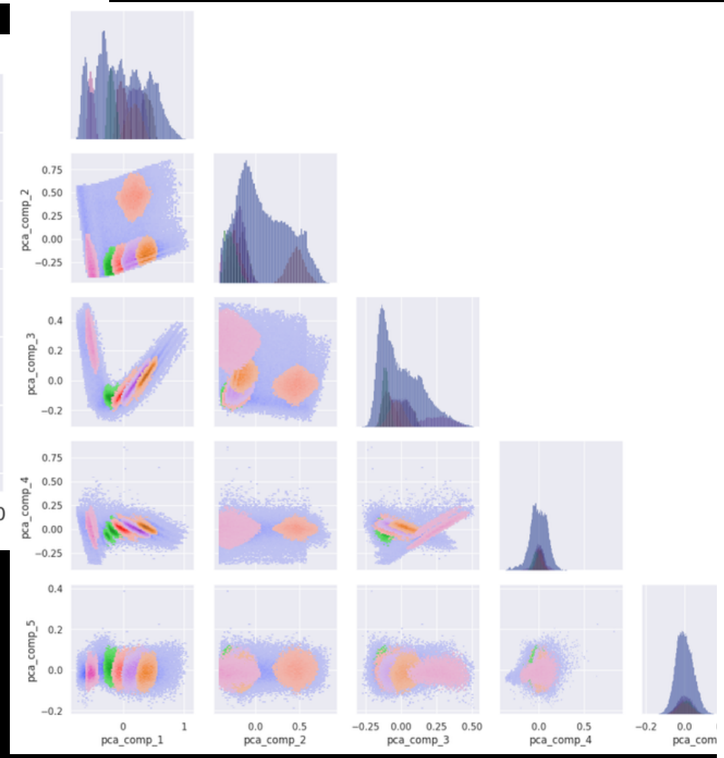
Clustering: Ordering points to identify the clustering structure (OPTICS)

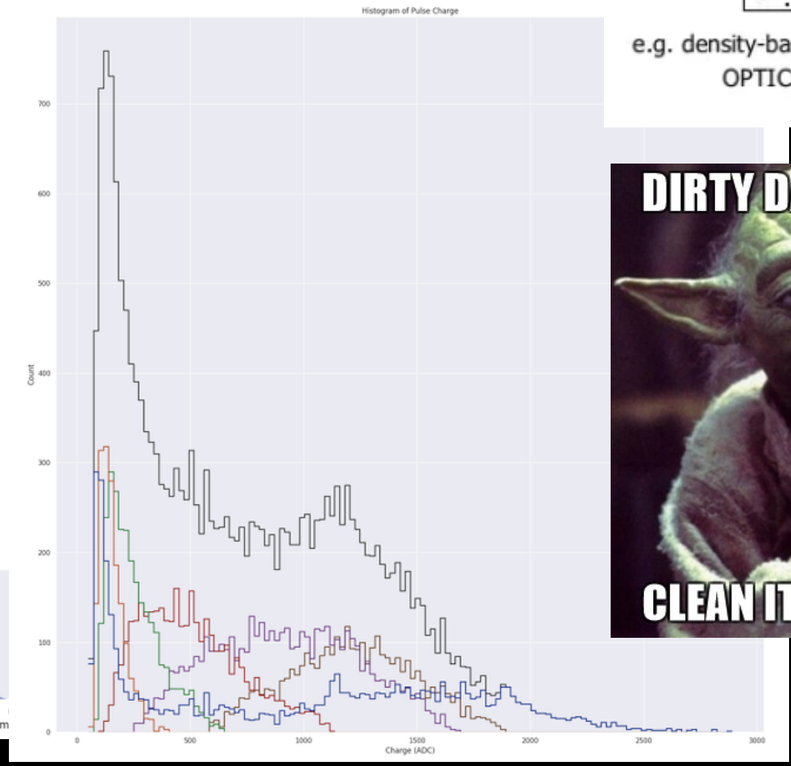


http://lagoproject.net/

In prep

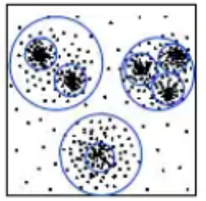






Hierarchical Clustering

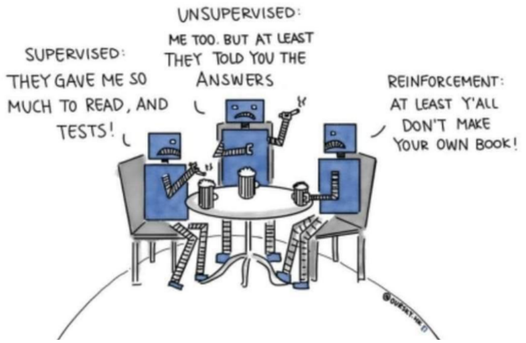
nested clusters



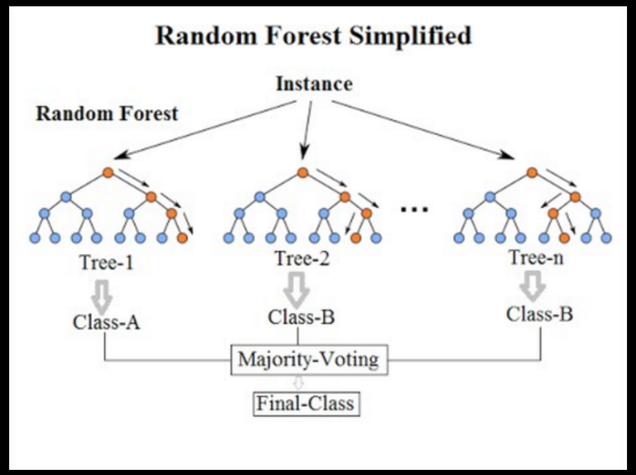
e.g. density-based clustering algorithm
OPTICS [SIGMOD 99]

https://indico.ictp.it/event/9840/

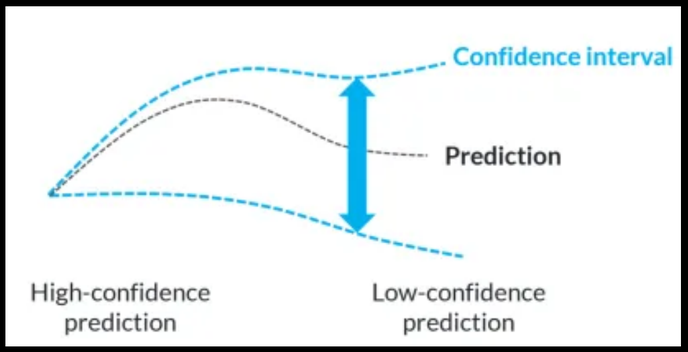
MACHINE LEARNING



- Ensemble techniques

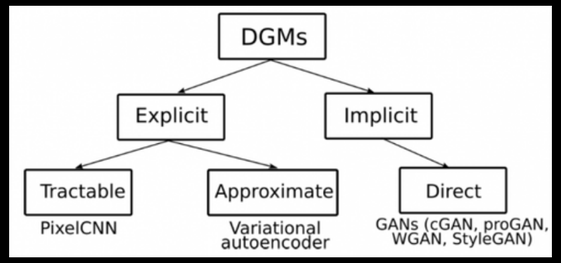


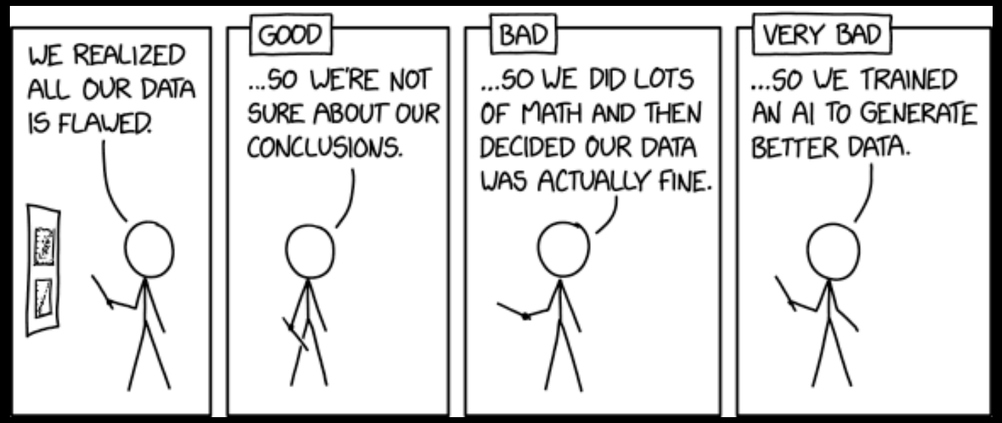
- Uncertanty Quantification (BNNs) - trustworthiness

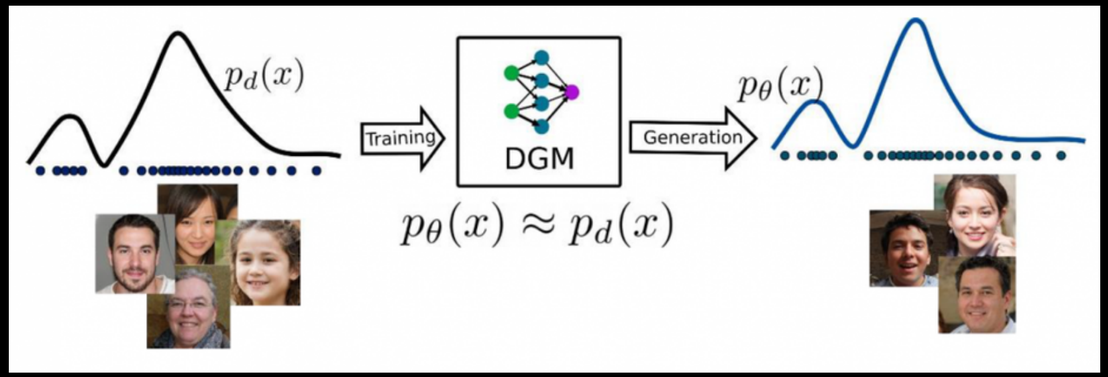


- Transformers
- Transfer Learning !

- Deep Generative Models







- (ML in production) Real -time -> re-training, incremental training, etc

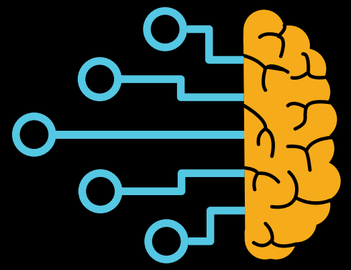
and much more!!!

OPEN DISCUSSION

APRENDIZAJE AUTOMÁTICO.

Fundamentos y Aplicaciones en Meteorología del Espacio

DCAO, UBA
08-12 AGOSTO 2022



Dra María Graciela Molina
FACET-UNT / CONICET
Tucumán Space Weather Center - TSWC

https://spaceweather.facet.unt.edu.ar/
IG -> @spaceweatherargentina

gmolina@herrera.unt.edu.ar