UNSAM
UNIVERSIDAD NACIONAL DE SAN MARTÍN

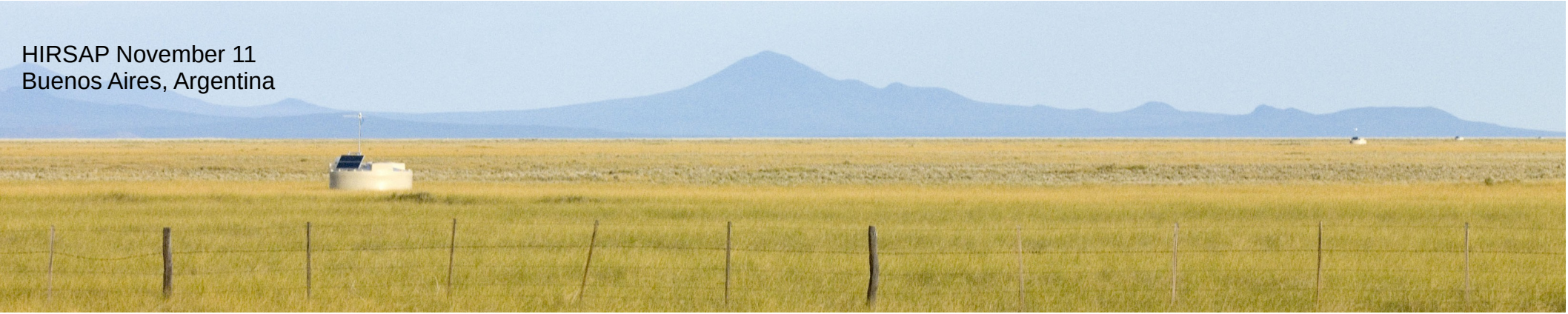
Double Doctoral degree in Astrophysics
DDAp

Karlsruhe Institute of Technology

Search for ultra-high energy photons with Machine Learning in the Auger Prime era

Ezequiel E. Rodriguez
Supervisors: Dr. Brian Wundheiler, Prof. Dr. Ralph Engel

HIRSAP November 11
Buenos Aires, Argentina

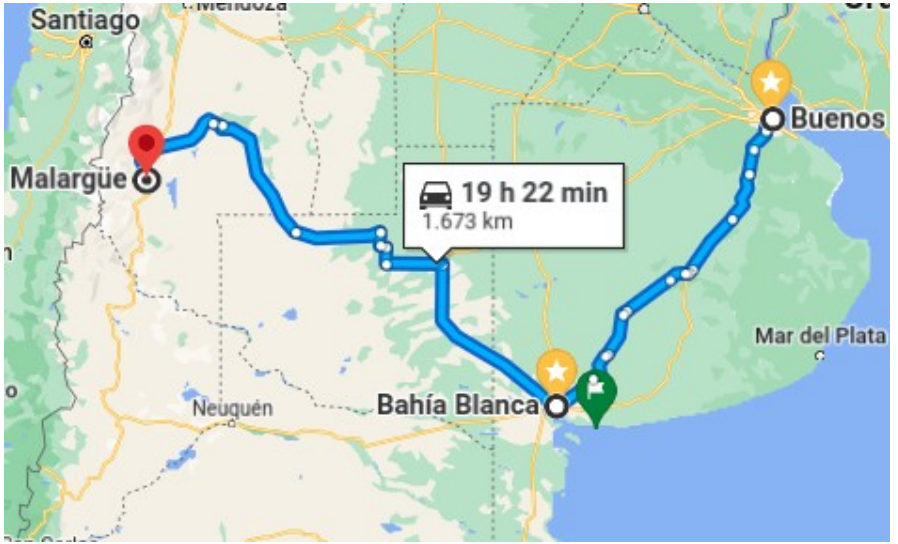


About me...

- Finished my undergraduate studies in 2019 in Bahia Blanca

- Worked as a Machine Learning model validator (1.6 years) and data scientist (6 months).
  For the most part engaged in fraud detection (small signals hidden in background).

- Started my PhD and joined the DDAp and HIRSAP in Abril



Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Photon search with the Underground Muon Detector (UMD)

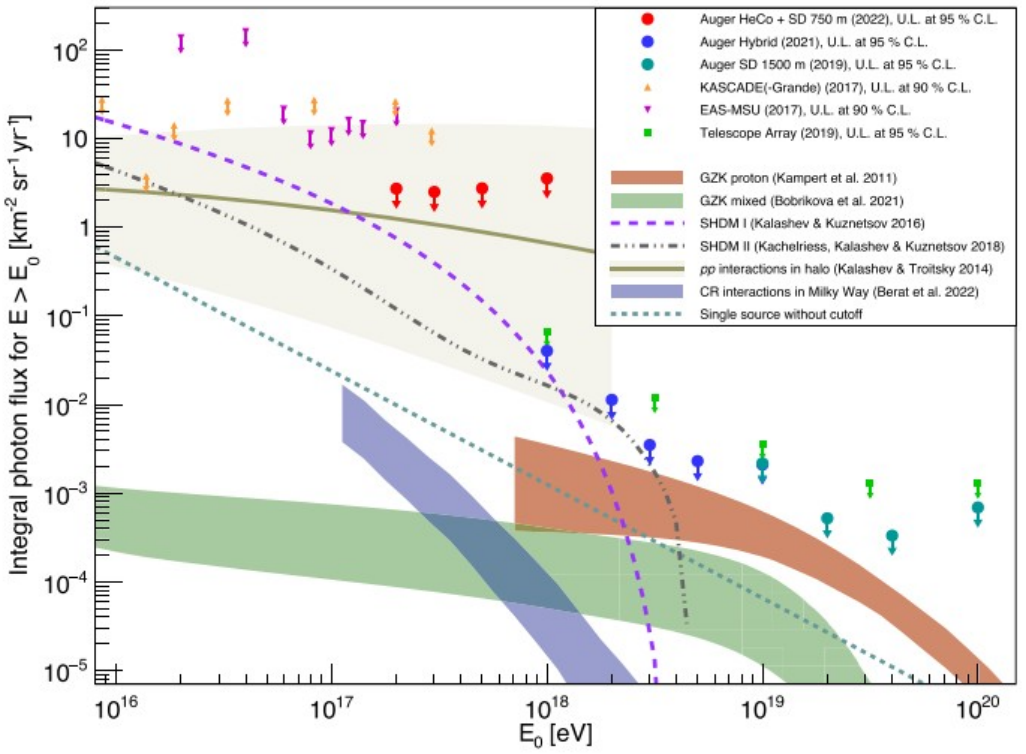
Motivations:

Why photons?

- Tracers of the highest-energy processes in the Universe
- Point back to their sources
- One of the “messengers” in the Multi-Messenger approach

The objective:

Estimate or constraint the diffuse photon flux at low energies → $\log_{10}\left(\frac{E_{\rm MC}}{\rm eV}\right) \leq 17.5$

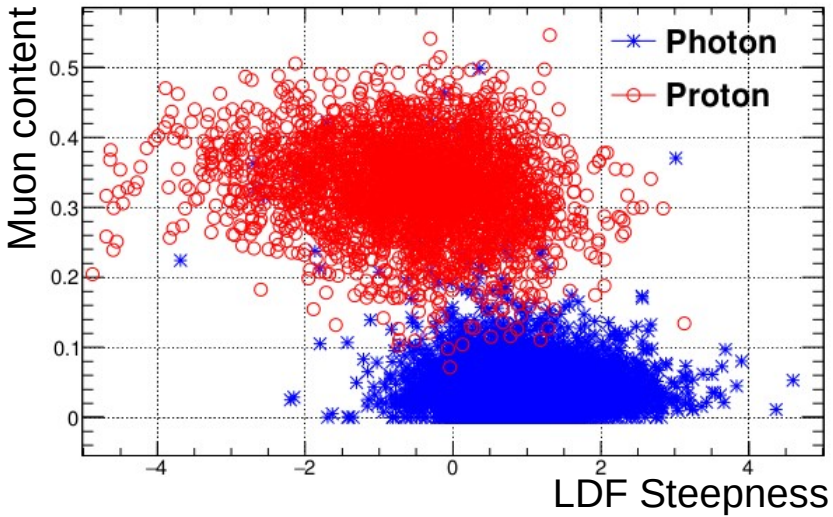


Auger Coll., ApJ 933 125, 2022

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

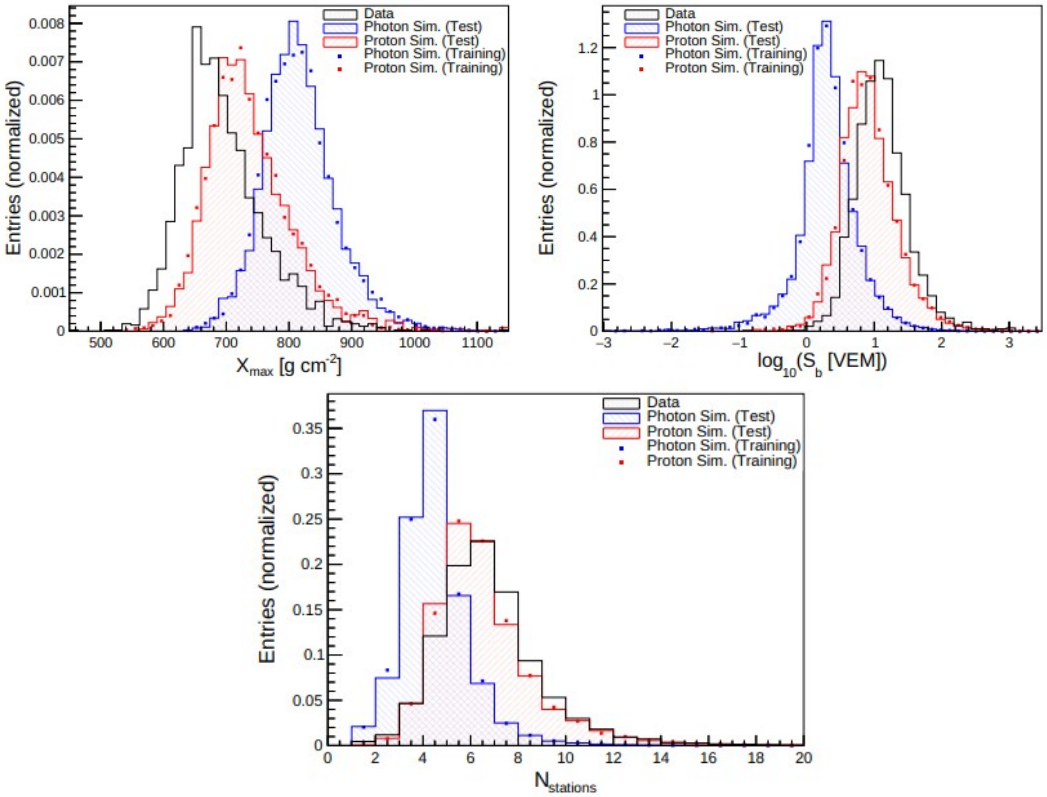
Multi-variate analysis in photon searches

González N., PhD thesis, 2018



• Combined with Fisher's Linear Discriminant Analysis

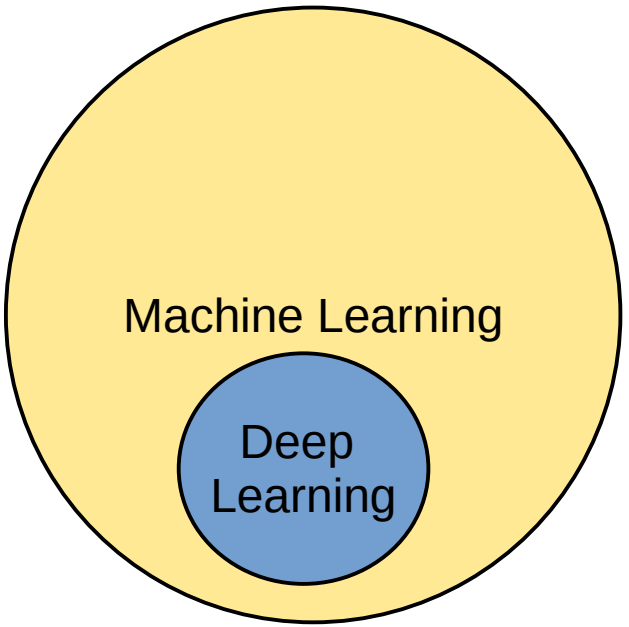
Auger Coll., ApJ 933 125, 2022



• Event observables sensitive to photon signatures
• Combined with Boosted Decision Trees

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Where we are headed



Our approach – Law of parsimony (Occam's razor)

Simple models

Logistic Regression, Gaussian Naive Bayes,
Decision trees, KNNs, etc.

Ensemble methods

Deep Learning

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Current strategy - Embracing uncertainty

- Represent events in tabular form
- Employ both *event-level* observables and *station-level* observables
- LDF-like observables are defined by binning the distance to the shower axis according to the core resolution

	rec_energy	rec_zenith	n_counters	n_stations	saturation_flag	mu_45	mu_75	mu_105	mu_135	mu_165	...
0	1.11e+17	27.56	3.0	13.0	1.0	NaN	NaN	NaN	NaN	0.07	...
1	1.21e+17	28.08	4.0	12.0	1.0	NaN	NaN	NaN	NaN	0.28	...
2	1.05e+17	26.21	4.0	13.0	1.0	NaN	NaN	0.38	NaN	NaN	...
3	1.11e+17	26.67	6.0	12.0	1.0	NaN	NaN	NaN	0.21	NaN	...
4	9.82e+16	27.41	5.0	12.0	1.0	NaN	NaN	NaN	0.10	NaN	...
5	1.10e+17	26.98	6.0	13.0	1.0	NaN	NaN	NaN	NaN	0.28	...
6	1.15e+17	27.89	4.0	13.0	1.0	NaN	NaN	NaN	NaN	NaN	...
7	1.14e+17	27.56	7.0	11.0	1.0	NaN	NaN	NaN	0.07	NaN	...
8	1.15e+17	27.28	3.0	12.0	1.0	NaN	NaN	NaN	NaN	NaN	...
9	1.13e+17	27.80	4.0	13.0	0.0	NaN	0.8	NaN	NaN	NaN	...
10	8.32e+16	30.78	3.0	11.0	2.0	NaN	0.6	NaN	NaN	NaN	...
11	8.69e+16	30.20	6.0	10.0	1.0	NaN	NaN	NaN	NaN	NaN	...
12	9.12e+16	29.75	6.0	10.0	1.0	NaN	NaN	NaN	0.10	NaN	...
13	8.10e+16	29.98	3.0	12.0	1.0	NaN	NaN	NaN	NaN	NaN	...
14	9.00e+16	30.10	3.0	12.0	1.0	NaN	NaN	NaN	NaN	0.07	...

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

The dataset

Showers are from the Prague Library (GAP 2018-043)

$$E_{\mathrm{MC}} \propto E^{-1} \,\mathrm{and}\, \theta_{\mathrm{MC}} \propto \sin\theta \, \cos\theta$$

Hadronic models: EPOS-LHC + FLUKA

Conservative background: protons with muon-deficit

Quality Cuts:

- $16.5 \leq \log_{10}\left(\frac{E_{\mathrm{MC}}}{\mathrm{eV}}\right) \leq 17.5$
- $\theta_{\mathrm{MC}} \leq 45^{\circ}$

Events: approx. 45k (balanced classes)

Missing values: 70%

Initial stratified split

Training set (2/3 original dataset)
- Model training
- Hyperparameter tuning
- Model Selection

Testing set (1/3 original dataset)
- Unbiased Performance Estimation

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Feature extraction

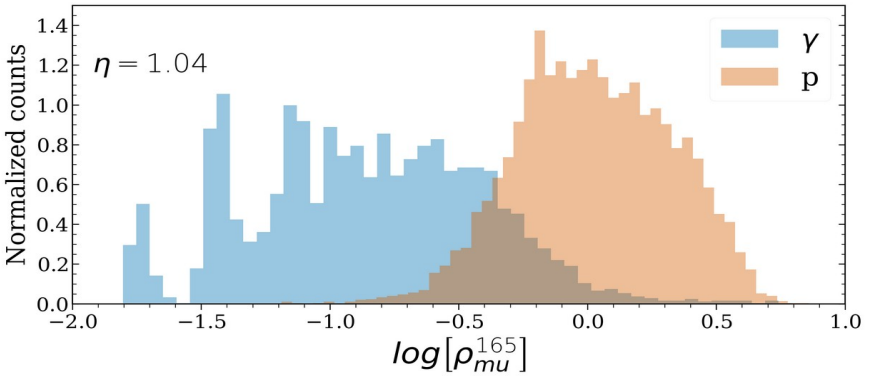
Normalized muon densities

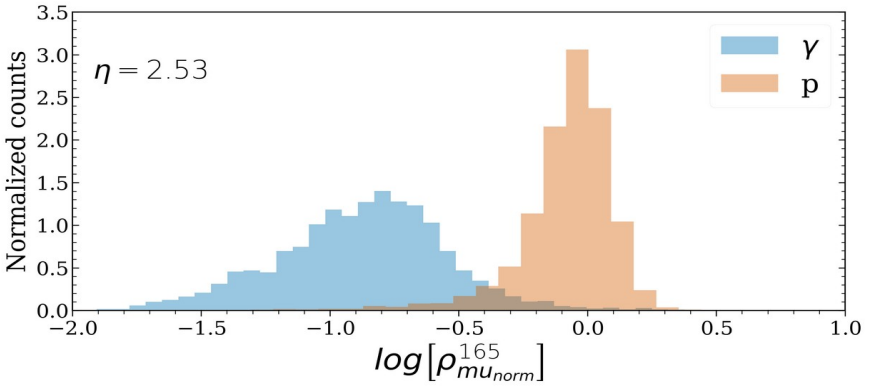
$$\rho^r_{\mu\ \mathrm{norm}} = \frac{\rho^r_\mu}{\rho^r_{\mu\ \mathrm{ref}}}$$

Parameters are fitted from a separate set of CORSIKA simulations for fixed combinations of E_{MC} and θ_{MC}

$$\rho^r_{\mu_{\mathrm{ref}}} = R^r_\mu(\theta)\left[\frac{E}{10^{16.7}\mathrm{eV}}\right]^{k^r_\mu}$$
$$\mathrm{R}^r_\mu(\theta) = ax^2 + bx + c$$
$$x = \sin^2\theta - \sin^2 30°$$

Distributions from Offline reconstructions





$$\eta = \frac{|Q^\gamma_{50} - Q^p_{50}|}{\sqrt{{\sigma_\gamma}^2 + {\sigma_p}^2}} \qquad \sigma^2 = \left(\frac{Q_{84} - Q_{16}}{2}\right)^2$$

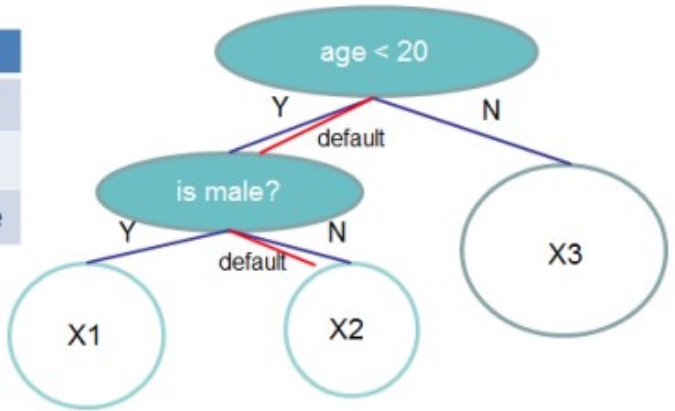
Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

The best contender – eXtreme Gradient Boosting (XGBoost)

- Introduced in 2016 (22k citations until last week)
- Highly optimized tree-based ensemble
- Handles missing values with Sparsity-aware Split Finding

Data

Example	Age	Gender
X1	?	male
X2	15	?
X3	25	female



Algorithm 3: Sparsity-aware Split Finding

Input: I, instance set of current node
Input: $I_k = \{i \in I | x_{ik} \neq \text{missing}\}$
Input: d, feature dimension
Also applies to the approximate setting, only collect statistics of non-missing entries into buckets
$gain \leftarrow 0$
$G \leftarrow \sum_{i \in I}, g_i, H \leftarrow \sum_{i \in I} h_i$
for $k = 1$ ***to*** m **do**
  // enumerate missing value goto right
  $G_L \leftarrow 0,\ H_L \leftarrow 0$
  for j *in sorted*$(I_k,$ *ascent order by* $\mathbf{x}_{jk})$ **do**
    $G_L \leftarrow G_L + g_j,\ H_L \leftarrow H_L + h_j$
    $G_R \leftarrow G - G_L,\ H_R \leftarrow H - H_L$
    $score \leftarrow \max(score, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda})$
  end
  // enumerate missing value goto left
  $G_R \leftarrow 0,\ H_R \leftarrow 0$
  for j *in sorted*$(I_k,$ *descent order by* $\mathbf{x}_{jk})$ **do**
    $G_R \leftarrow G_R + g_j,\ H_R \leftarrow H_R + h_j$
    $G_L \leftarrow G - G_R,\ H_L \leftarrow H - H_R$
    $score \leftarrow \max(score, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda})$
  end
end
Output: Split and default directions with max gain

Chen, T., & Guestrin, C. (2016, August). *Xgboost: A scalable tree boosting system.*

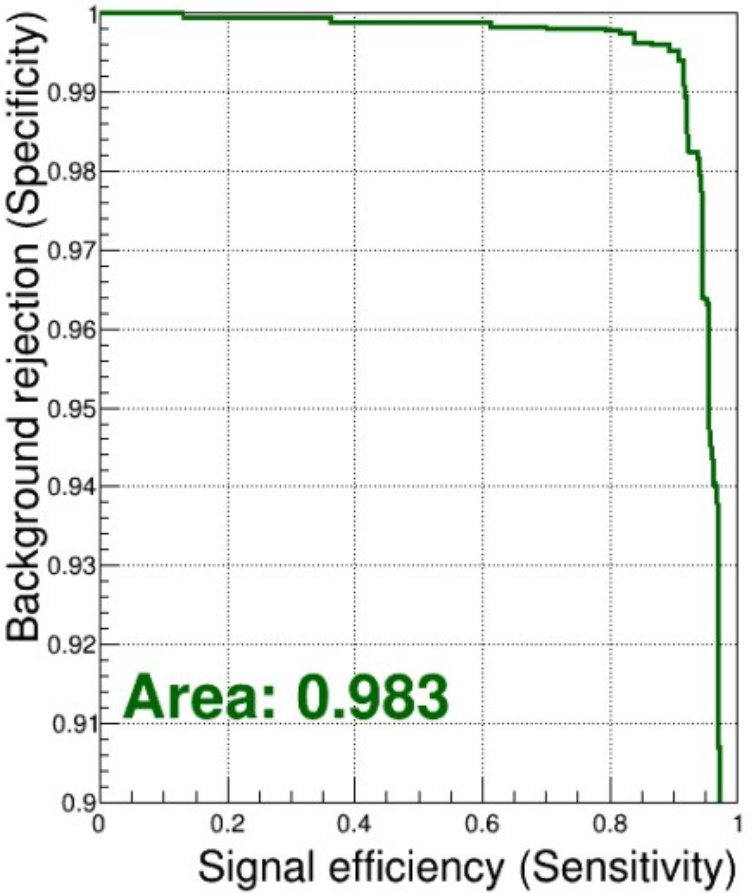
Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Model Training

Stratified 5-fold Cross-Validation (CV)
Low-level optimization task

output

Mean Area Under the Curve (AUC) for Background Rejection (BR) Vs Signal Efficiency (SE)



Savina, P., Bleve, C., & Perrone, L. , PoS (ICRC2019), 414.

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Training – Hyperparameter tuning – Model selection

Model Selection
Select a set of features from the BO distributions of AUCs

Sets:

$\{\rho^r_{\mu}\}$

$\{\rho^r_{\mu_{norm}}\}$

$\{\rho^r_{\mu_{norm}}, S^r\}$

Number of trained Classifiers
3 x 1100 x 5 = 16500

Hyper-parameter tuning

What is the best configuration for XGBoost?
Too many hyperparameters (non-trainable)

Bayesian Optimization (BO) ← BRvsSE AUC

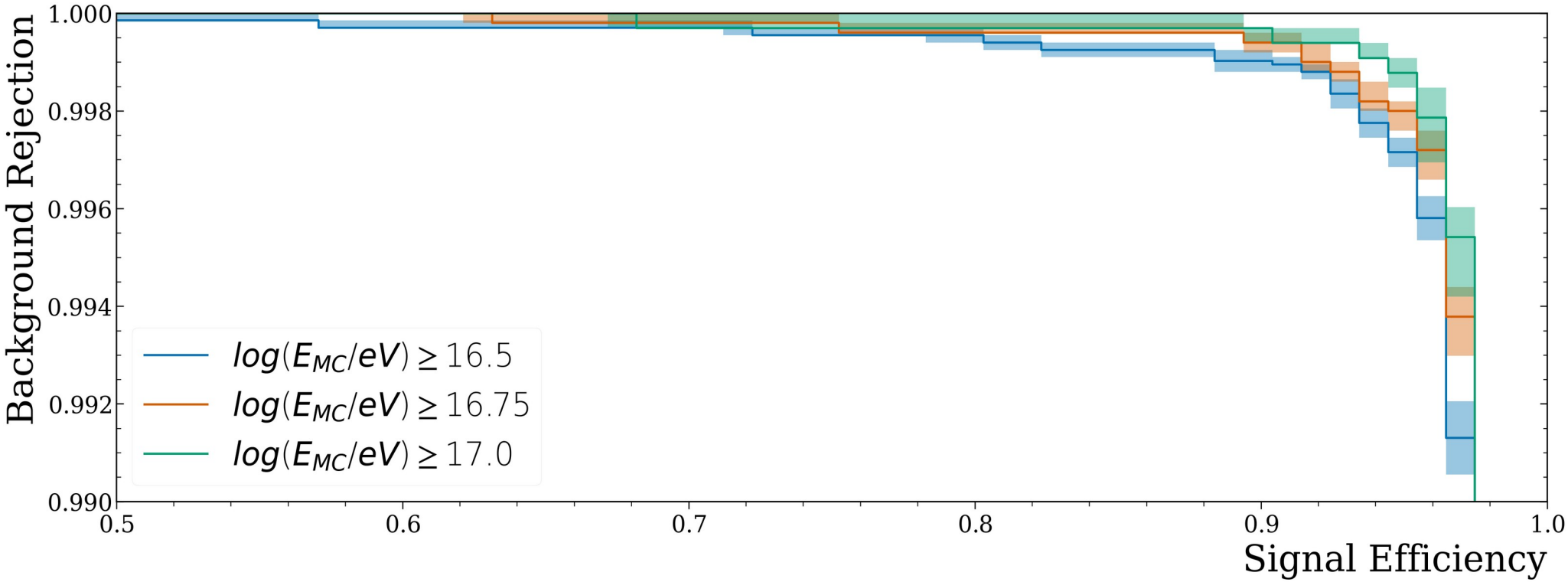
Meta-optimization in a

high-dimensional space (1100 trials)

Model Training
with 5-fold CV

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

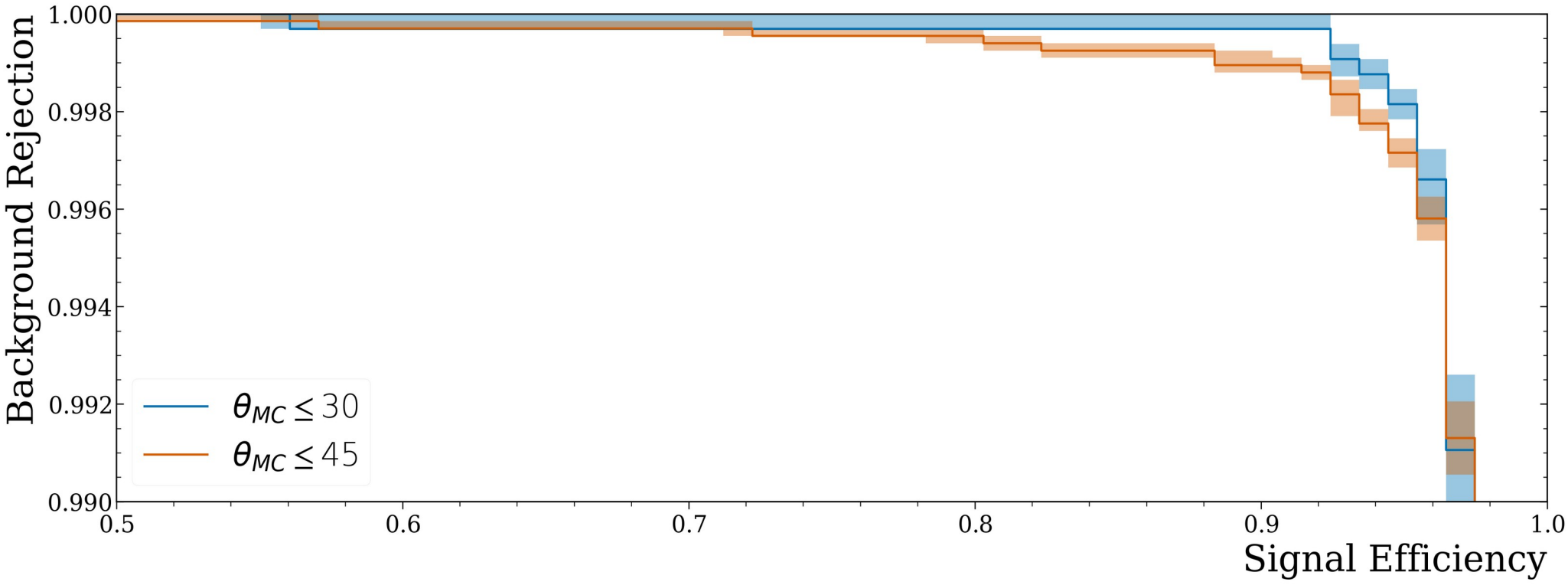
Out-of-sample performance estimation with bootstrap



As we contemplate events with lower primary energy performance drops

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Out-of-sample performance estimation with bootstrap

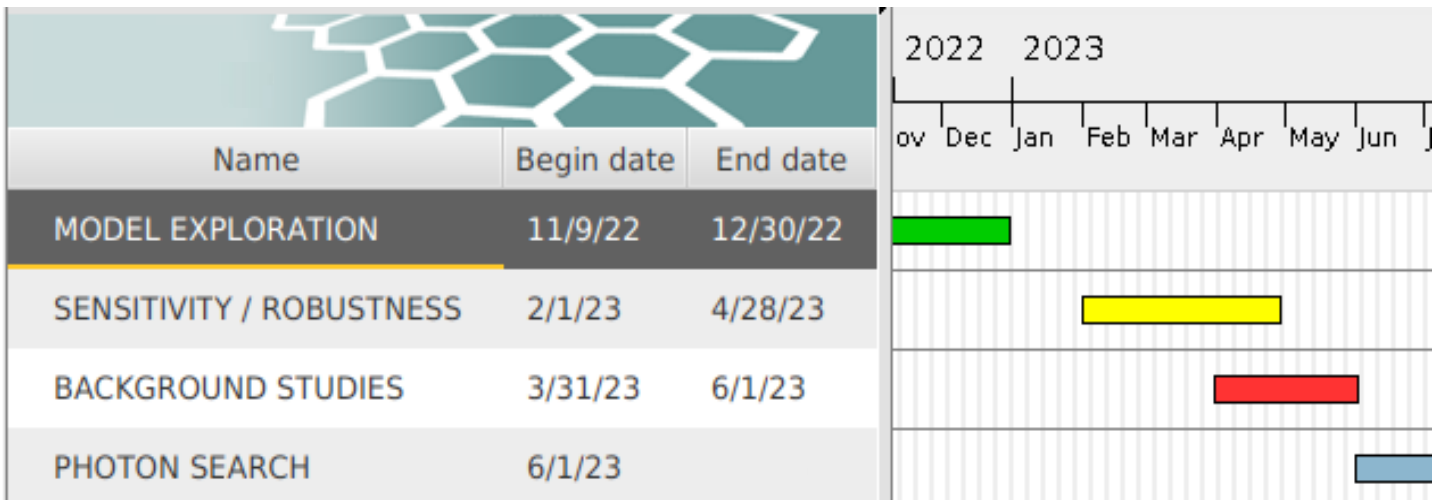


As we contemplate events with higher zenith angle performance drops

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Summary and Outlook

- Developed observables as input for to MVA classifiers
- Initial model development with “traditional” Machine Learning models
- Estimation of model performance



Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Backup slides

Bayesian Optimization - Tree Parzen Estimator (1)

- Sequential Model-Based Optimization(SMBO) -> Tree Parzen Estimator (TPE)

Algorithms for Hyper-Parameter Optimization

SMBO(f, M_0, T, S)

1 $\mathcal{H} \leftarrow \emptyset$,
2 For $t \leftarrow 1$ **to** T,
3 $x^* \leftarrow \text{argmin}_x\ S(x, M_{t-1})$,
4 Evaluate $f(x^*)$, ▷ *Expensive step*
5 $\mathcal{H} \leftarrow \mathcal{H} \cup (x^*, f(x^*))$,
6 Fit a new model M_t to $\mathcal{H}$.
7 **return** $\mathcal{H}$

Figure 1: The pseudo-code of generic Sequential Model-Based Optimization.

$$\text{EI}_{y^*}(x) := \int_{-\infty}^{\infty} \max(y^* - y, 0) p_M(y|x) dy.$$

$$\text{EI}_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x) dy = \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)p(y)}{p(x)} dy$$

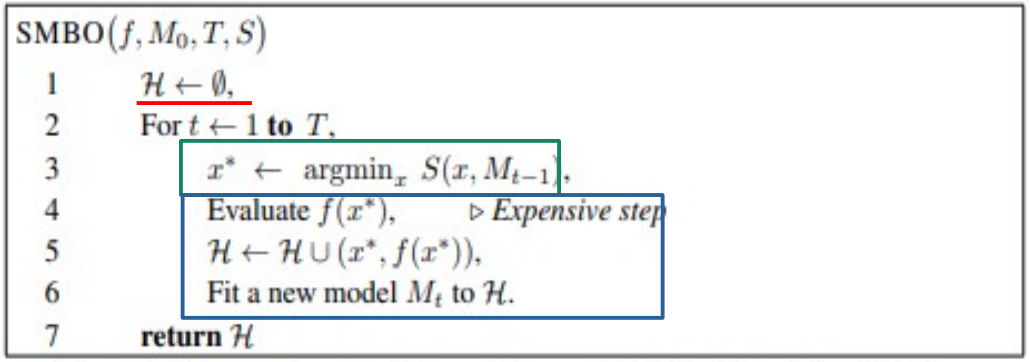
Some steps later...

$$EI_{y^*}(x) = \frac{\gamma y^* \ell(x) - \ell(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma \ell(x) + (1-\gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{\ell(x)}(1-\gamma)\right)^{-1}$$

$$p(x|y) = \begin{cases} \ell(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^*, \end{cases} \qquad \gamma = p(y < y^*)$$

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Bayesian Optimization - Tree Parzen Estimator (2)



SMBO(f, M_0, T, S)

1 $\mathcal{H} \leftarrow \emptyset$,
2 For $t \leftarrow 1$ **to** T,
3 $x^* \leftarrow \text{argmin}_x\ S(x, M_{t-1})$,
4 Evaluate $f(x^*)$, ▷ *Expensive step*
5 $\mathcal{H} \leftarrow \mathcal{H} \cup (x^*, f(x^*))$,
6 Fit a new model M_t to $\mathcal{H}$.
7 **return** $\mathcal{H}$

Figure 1: The pseudo-code of generic Sequential Model-Based Optimization.

- Warm-up Stage with random trials

Initial kernel density estimation

$$p(x|y) = \begin{cases} \ell(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^*, \end{cases} \quad \gamma = p(y < y^*)$$

Update likelihoods

$$x^* = argmin_x \frac{g(x)}{l(x)}$$

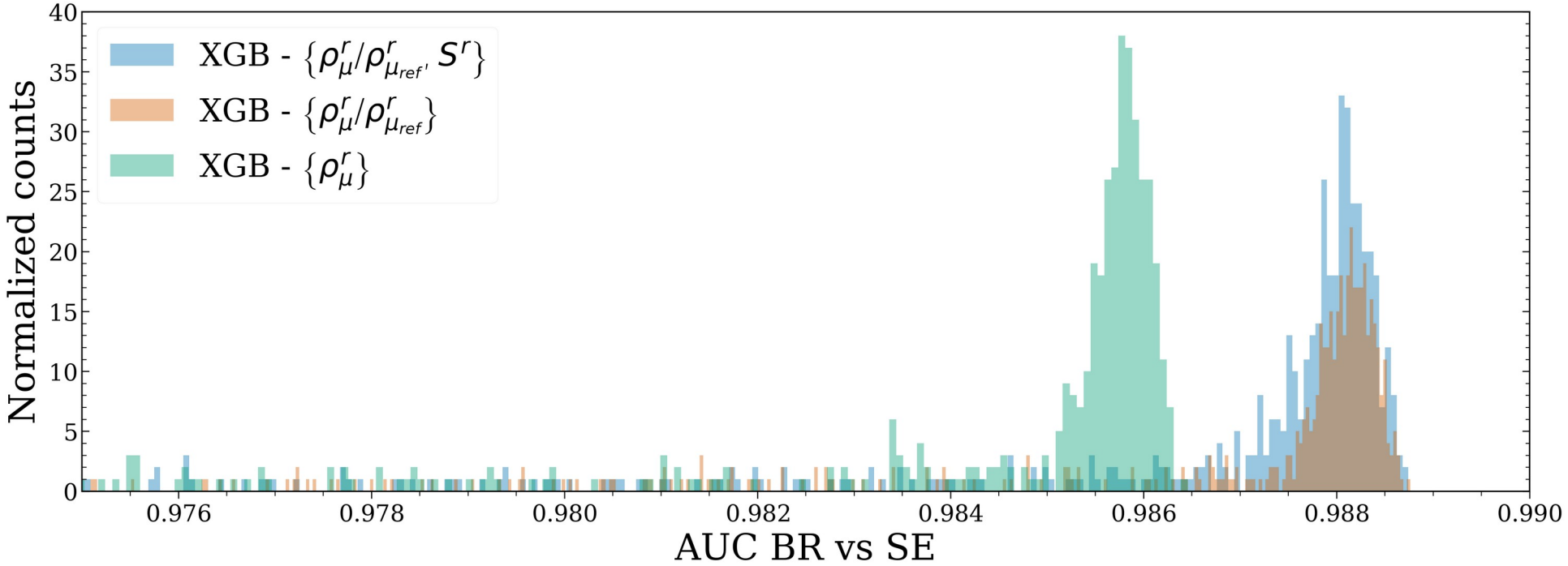
- Train the model with x*
- Append your metric to the search history
- Update likelihood

Side Notes

- The hyperparameter search space follows a tree-like structure
- Likelihoods are modeled like:
  - Mix of truncated gaussians for continuous hyperparameters
  - PMFs for integer and categorical hyperparameters

Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar

Results from Bayesian Optimization



Ezequiel E. Rodriguez – ezequiel.rodriguez@iteda.cnea.com.ar