FOA 1861 FINAL PROJECT BRIEFING
BIG DATA ANALYSIS OF SYNCHROPHASOR DATA

Combinatorial Evaluation of Physical Feature Engineering, Classical Machine Learning, and Deep Learning Models for Synchrophasor Data at Scale

Dr. Mohini Bariya
PingThings
mohini@pingthings.io
October 5, 2021

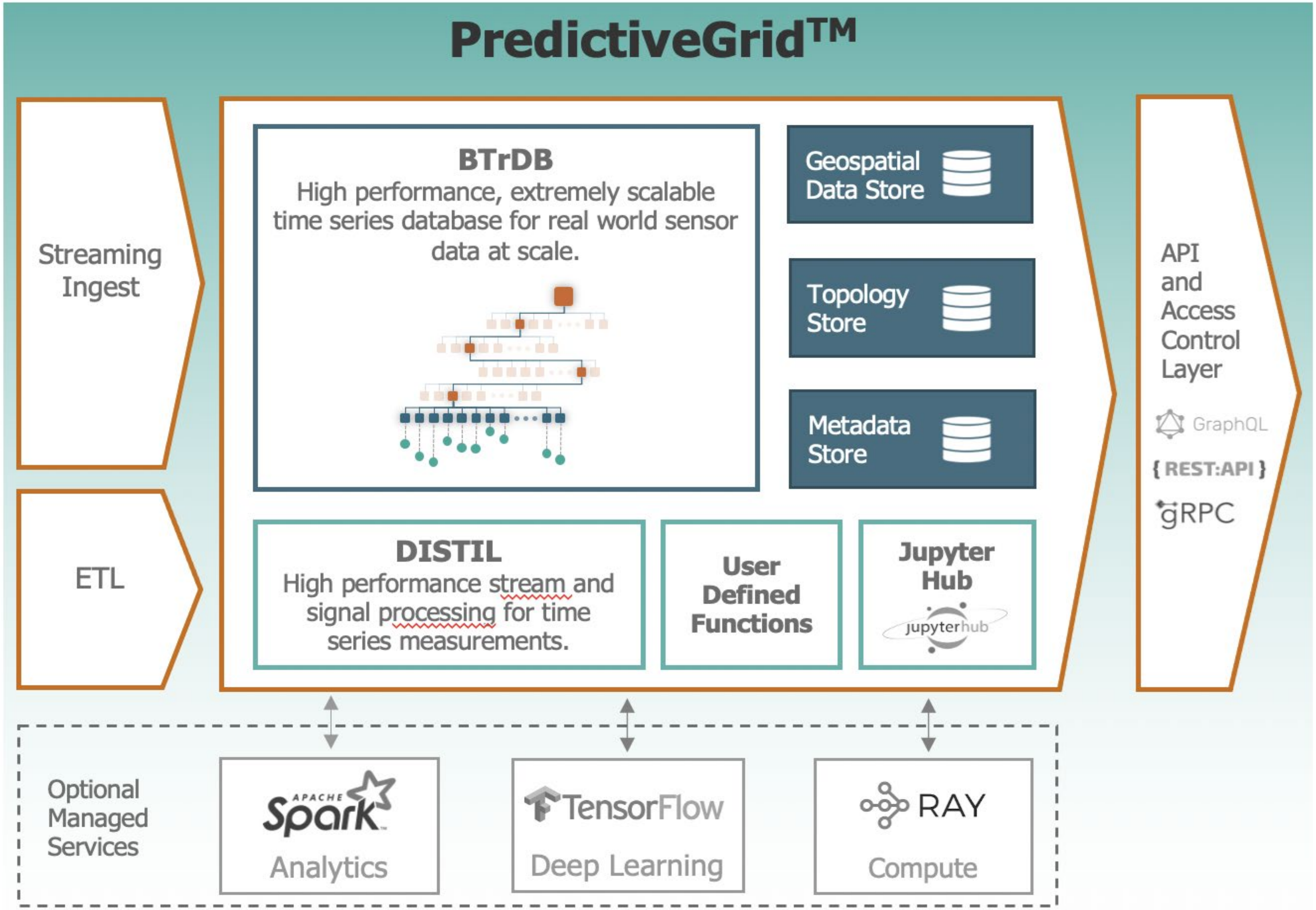
Project Partners

PingThings

Outline

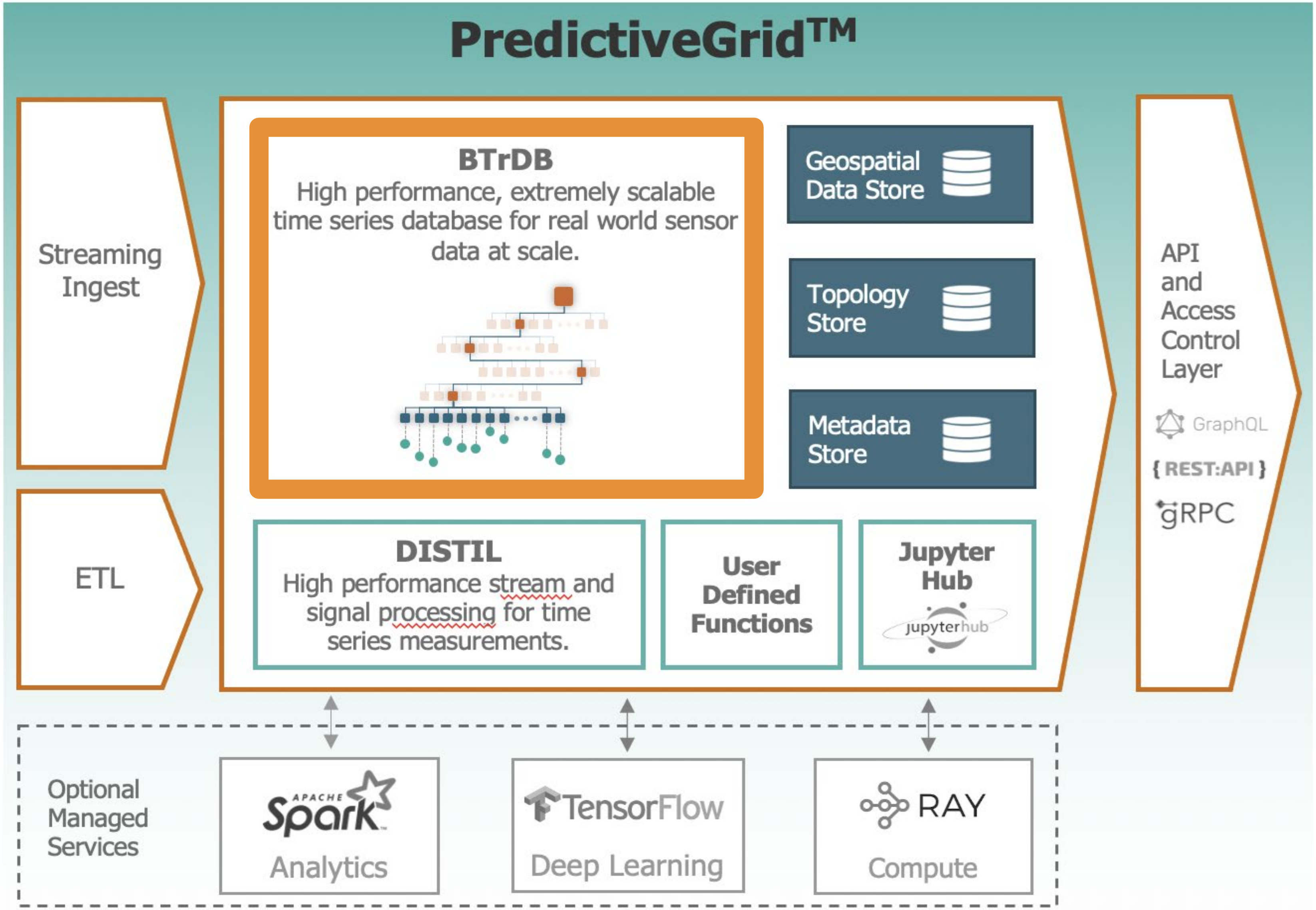
- Background
- Experimental Results
  - Statistical & ML Pipeline
  - Assessment of Datasets
  - Validation Results
- Technical Accomplishments
- Value of Work
- Readiness for Commercialization
- Readiness for ML & BD Analytics
- Lessons Learned and Next Steps

Background - The Platform



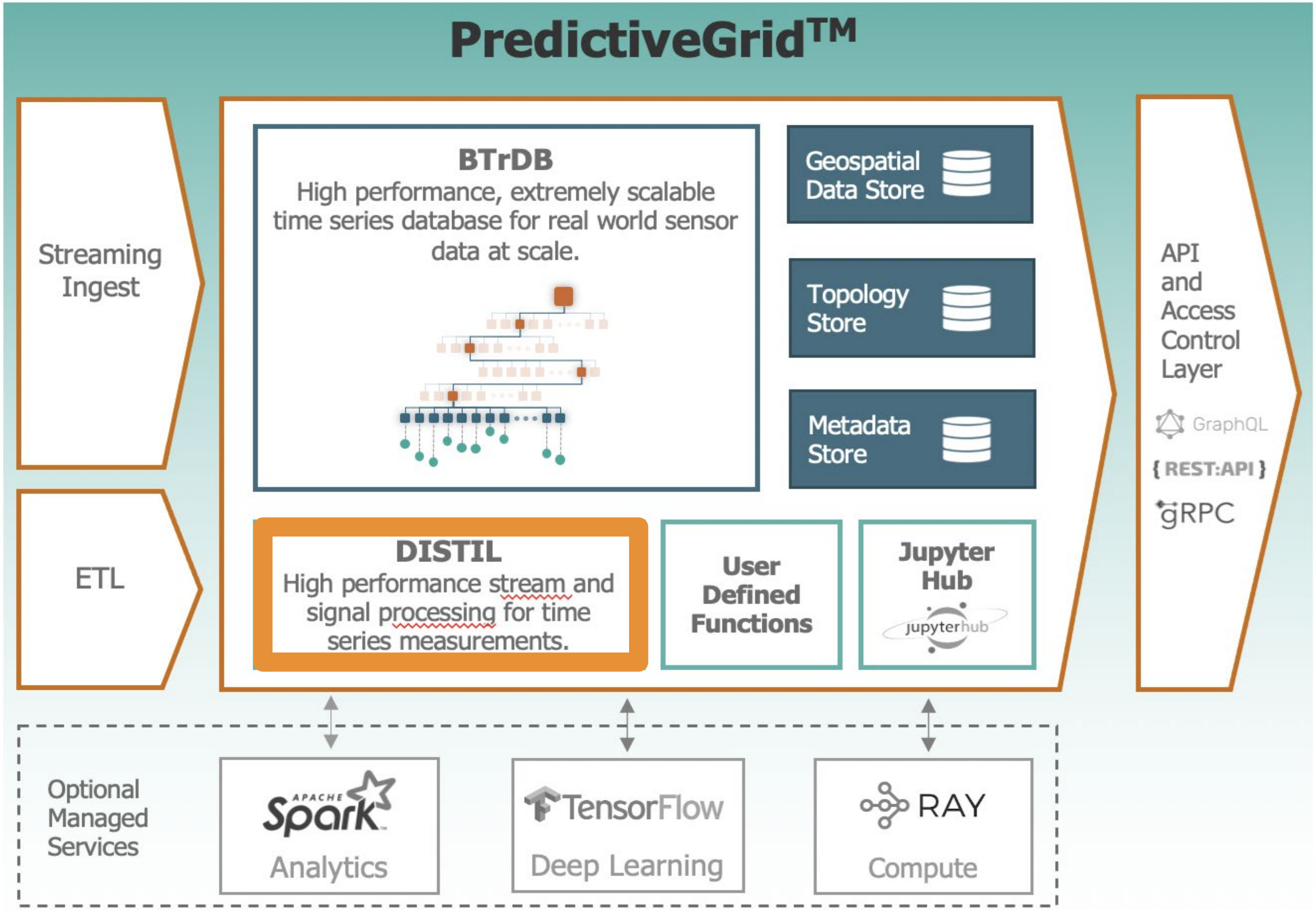
PingThings

Background - The Platform



PingThings

Background - The Platform



PingThings

Background - Our Approach

Aims

- Identify & classify events (within and outside utility logs)
- Identify & classify precursors
- Extract event signatures
- Discover seasonal & weather patterns

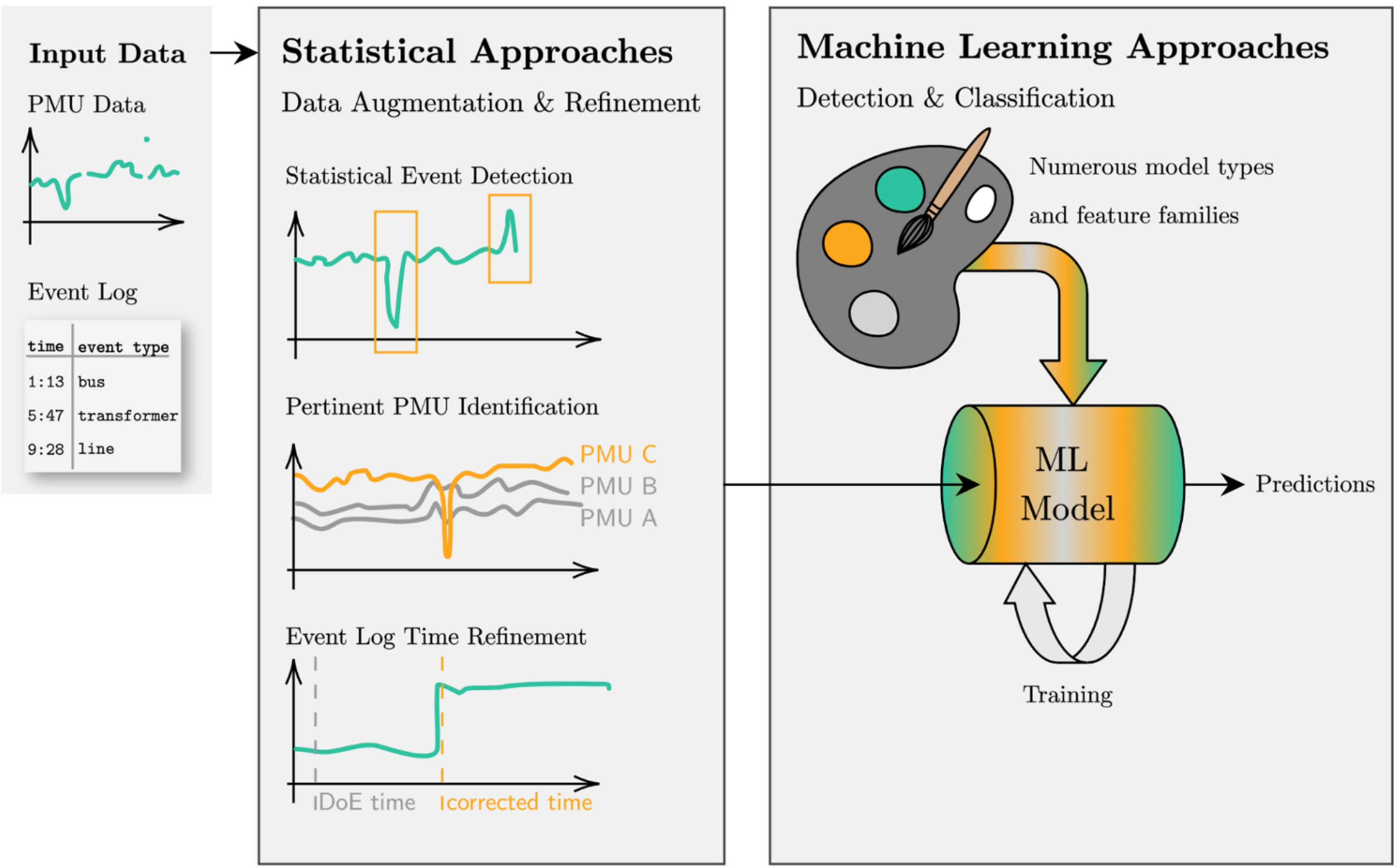
Strategy

- Broad survey & assessment of algorithms for these tasks.

Significance

- Development of broadly useful tooling for pain free analytics pipeline.
- Algorithm discoveries

Statistical & ML Pipeline



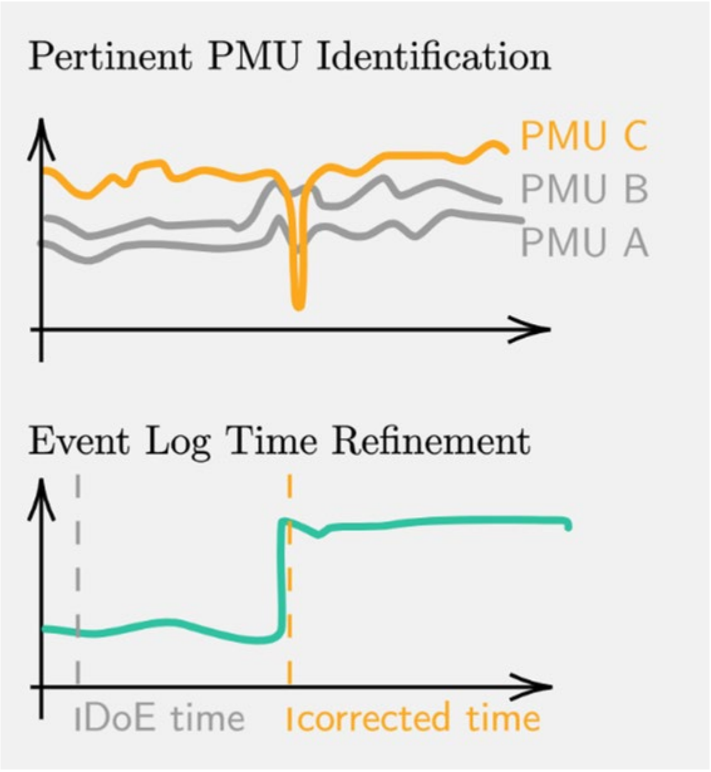
Dataset Assessment

Measurements

- Missing measurements
- Bad values
- Mislabeling

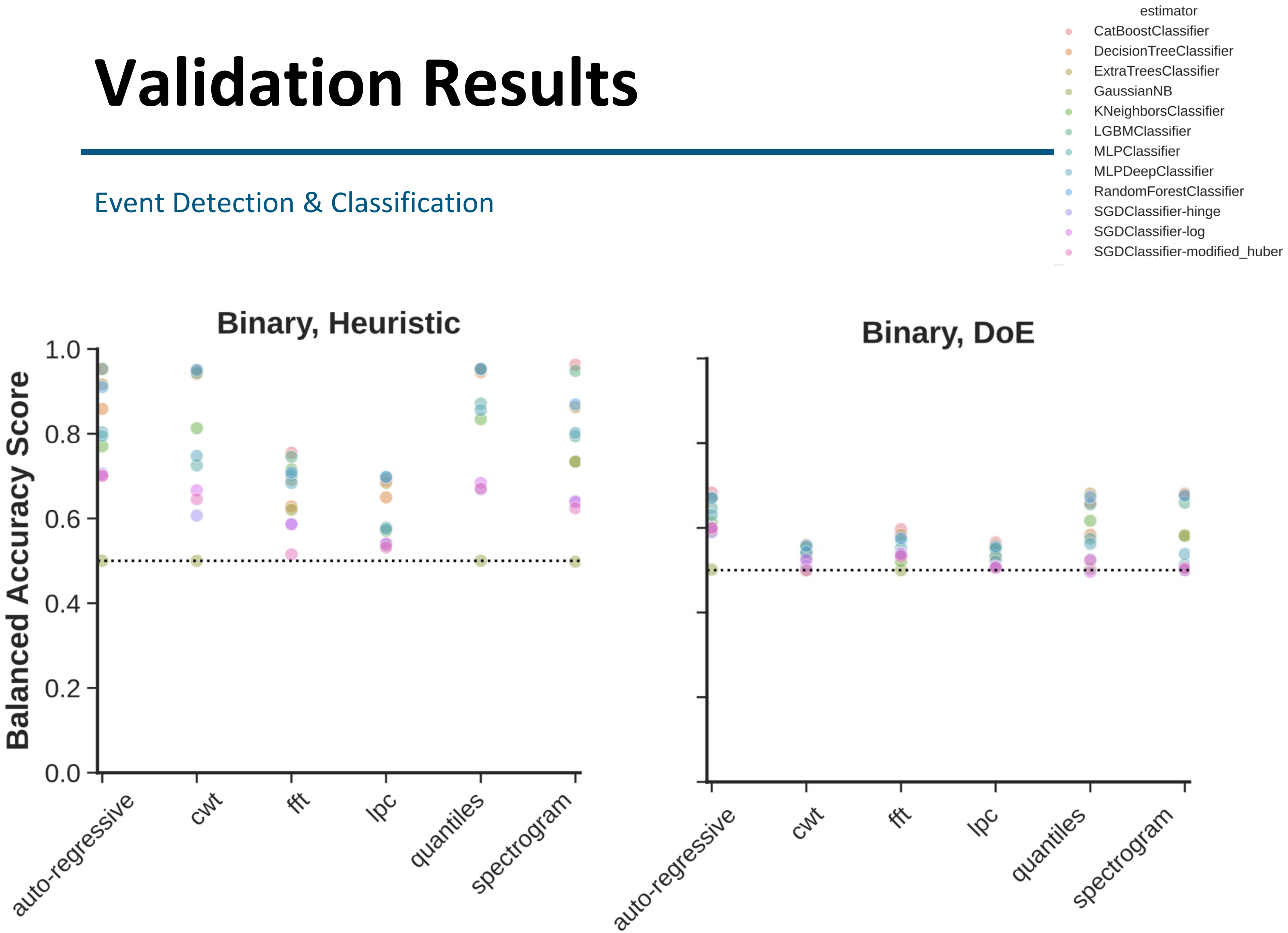
Event Logs

- Event times inaccurate
- No spatial information - topology or event location

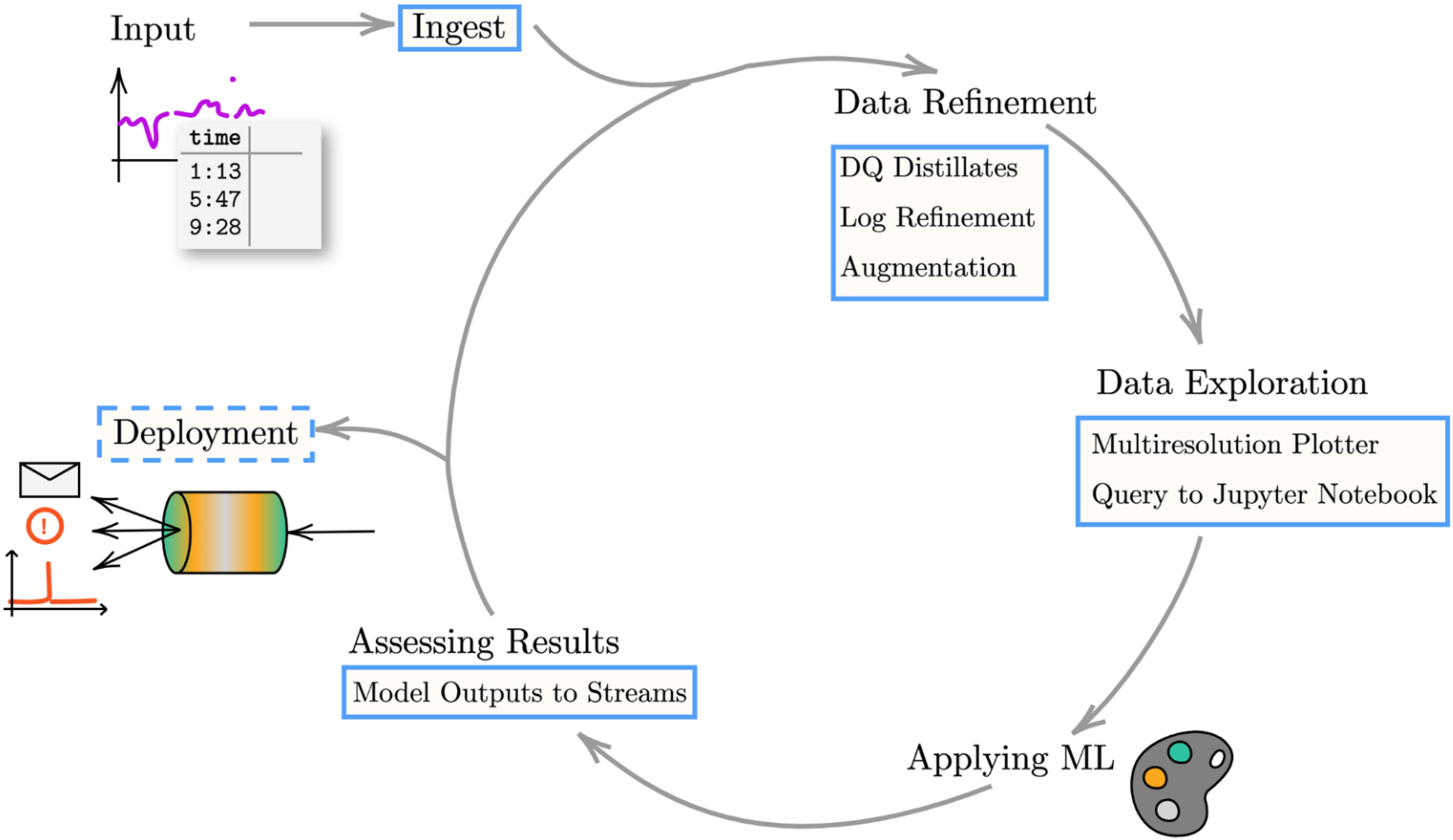


Validation Results

Event Detection & Classification

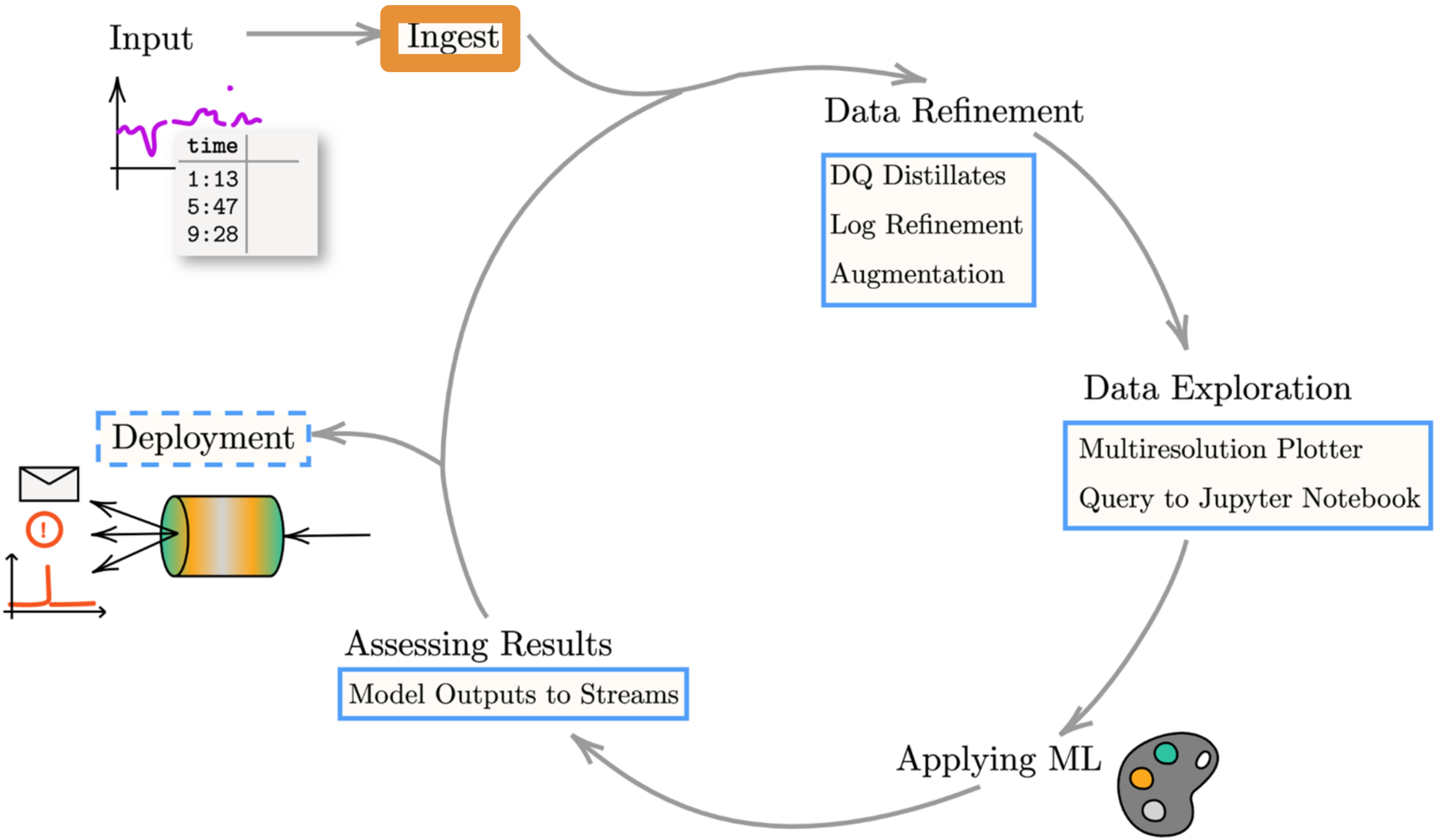


An AI Platform for Grid Data



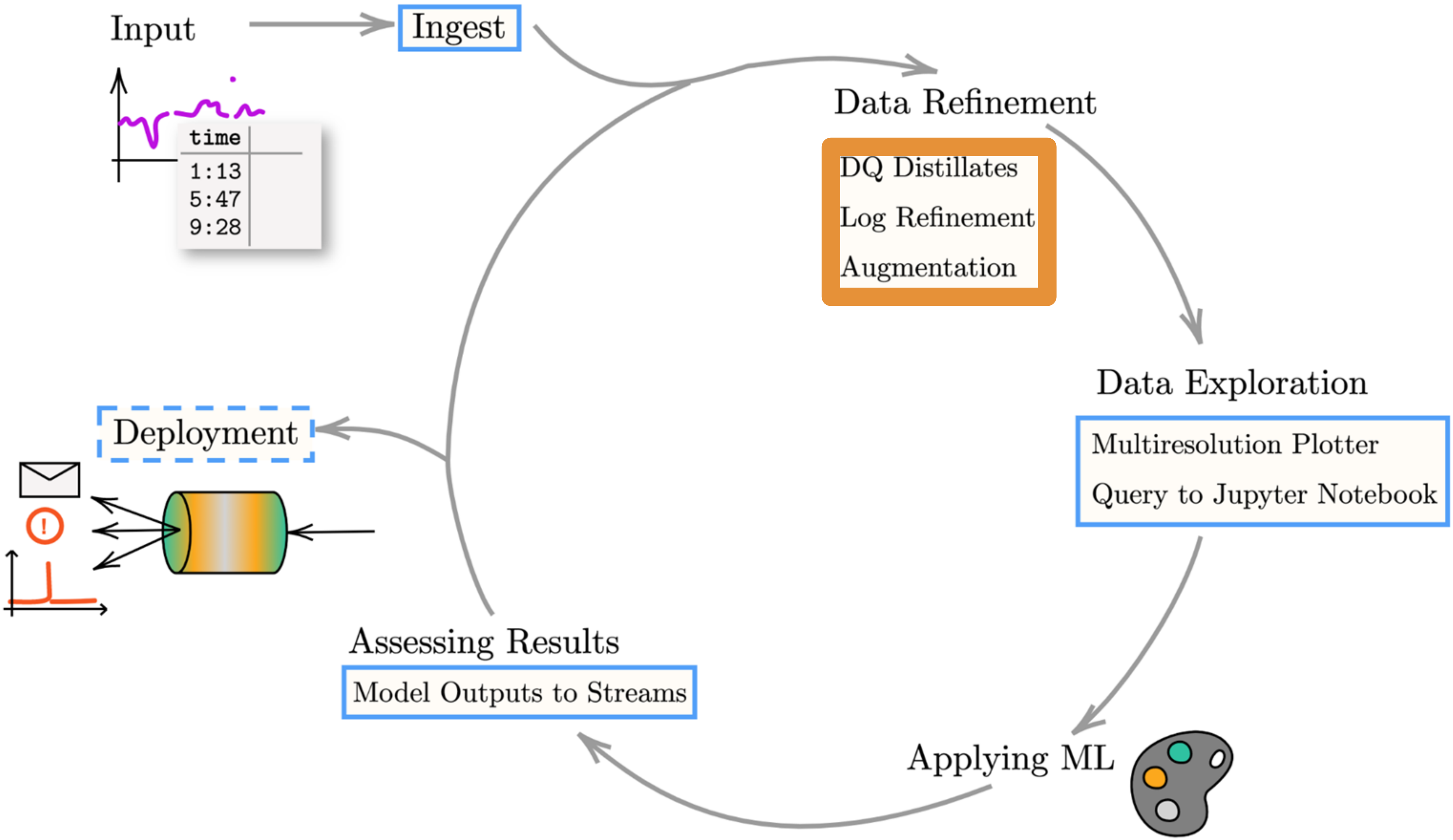
PingThings

An AI Platform for Grid Data



PingThings

An AI Platform for Grid Data



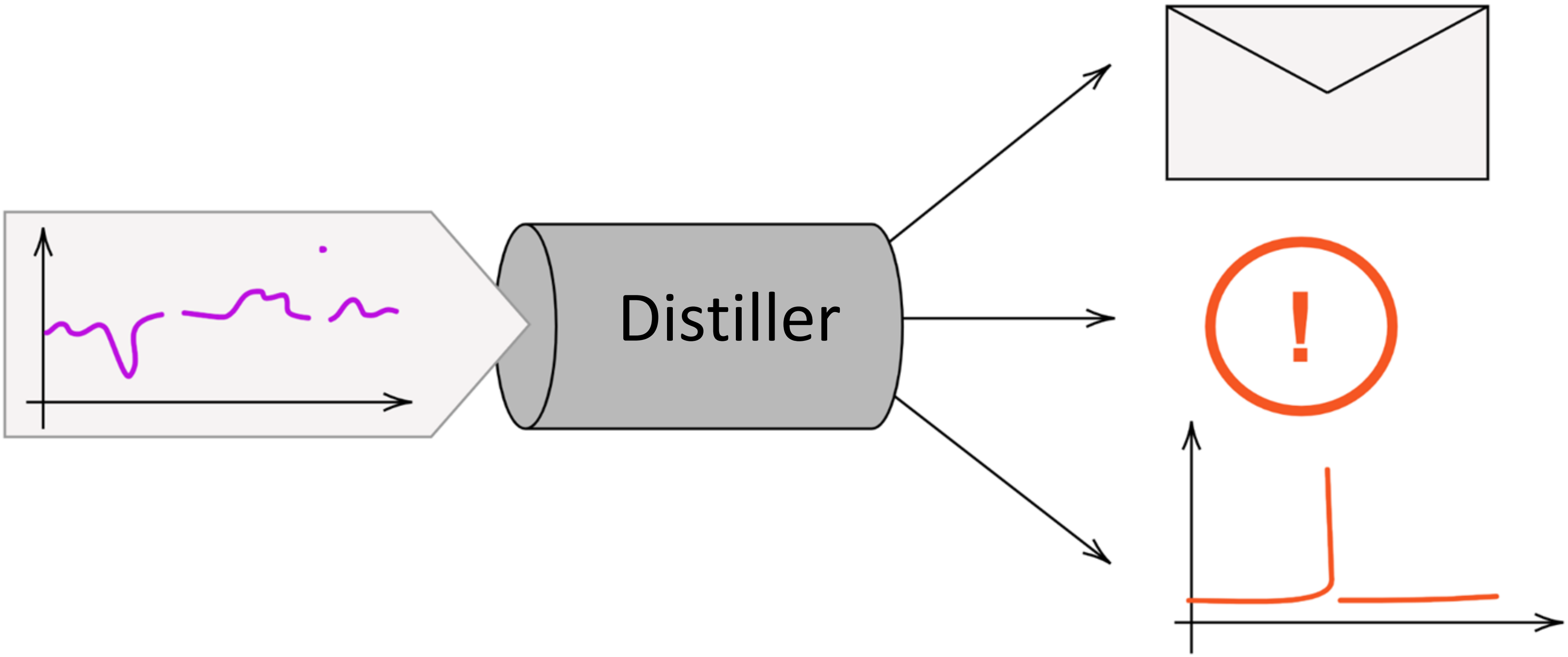
PingThings

Data Quality Distillates - Understand & Flag DQ Issues

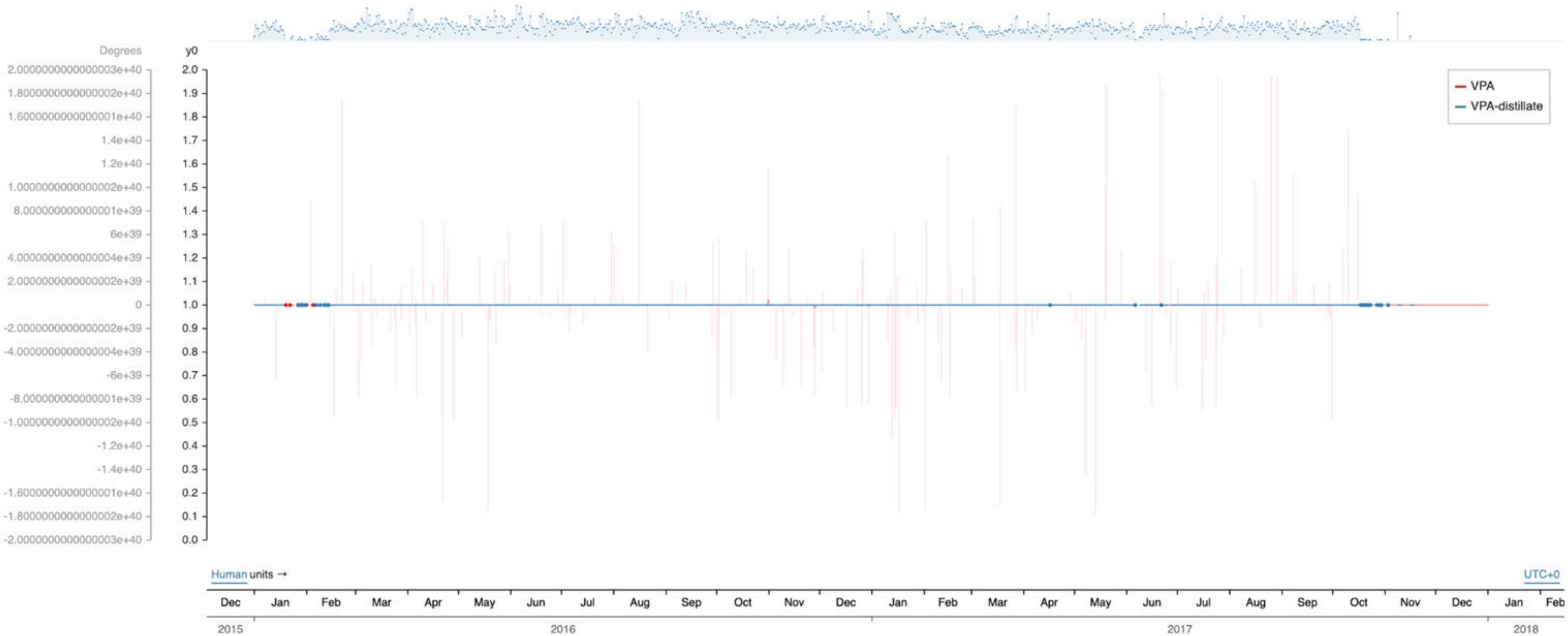
PingThings

NETL
NATIONAL ENERGY TECHNOLOGY LABORATORY

Data Quality Distillates - Understand & Flag DQ Issues

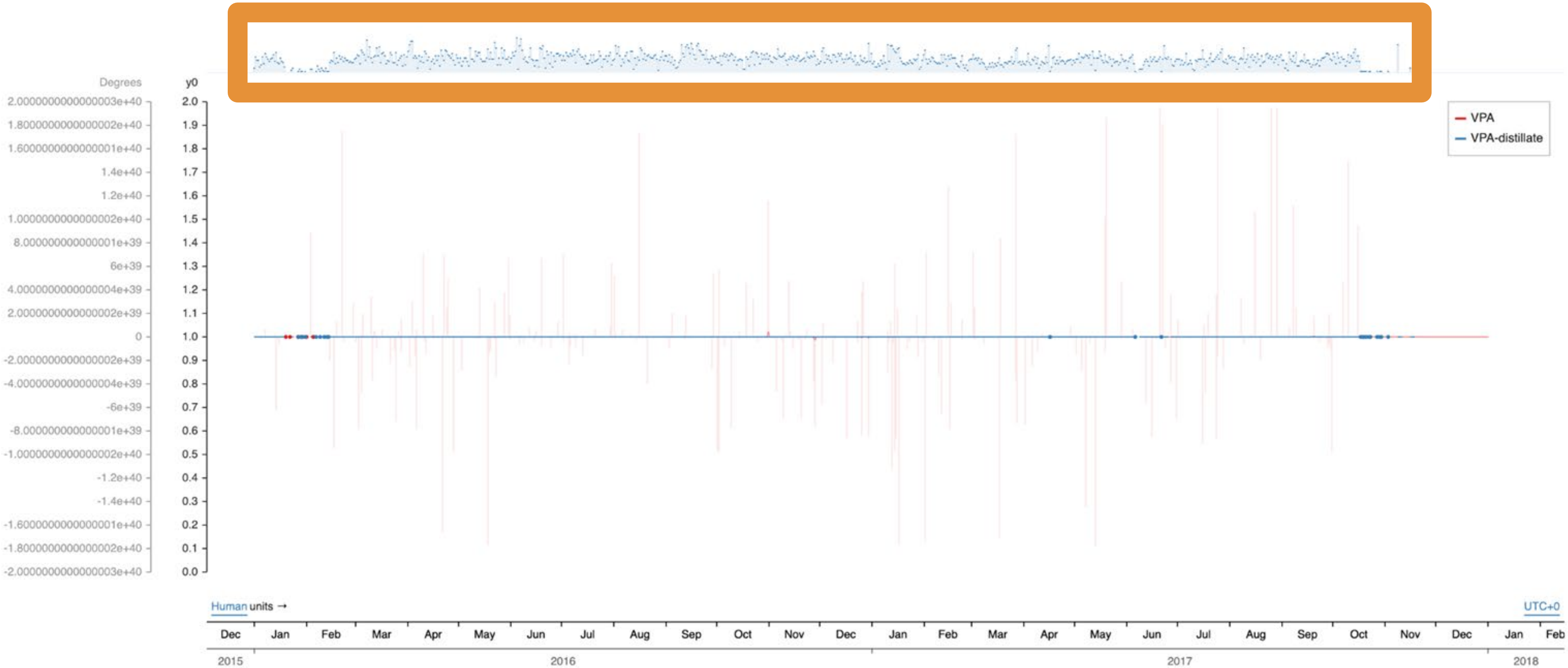


Data Quality Distillates - Understand & Flag DQ Issues



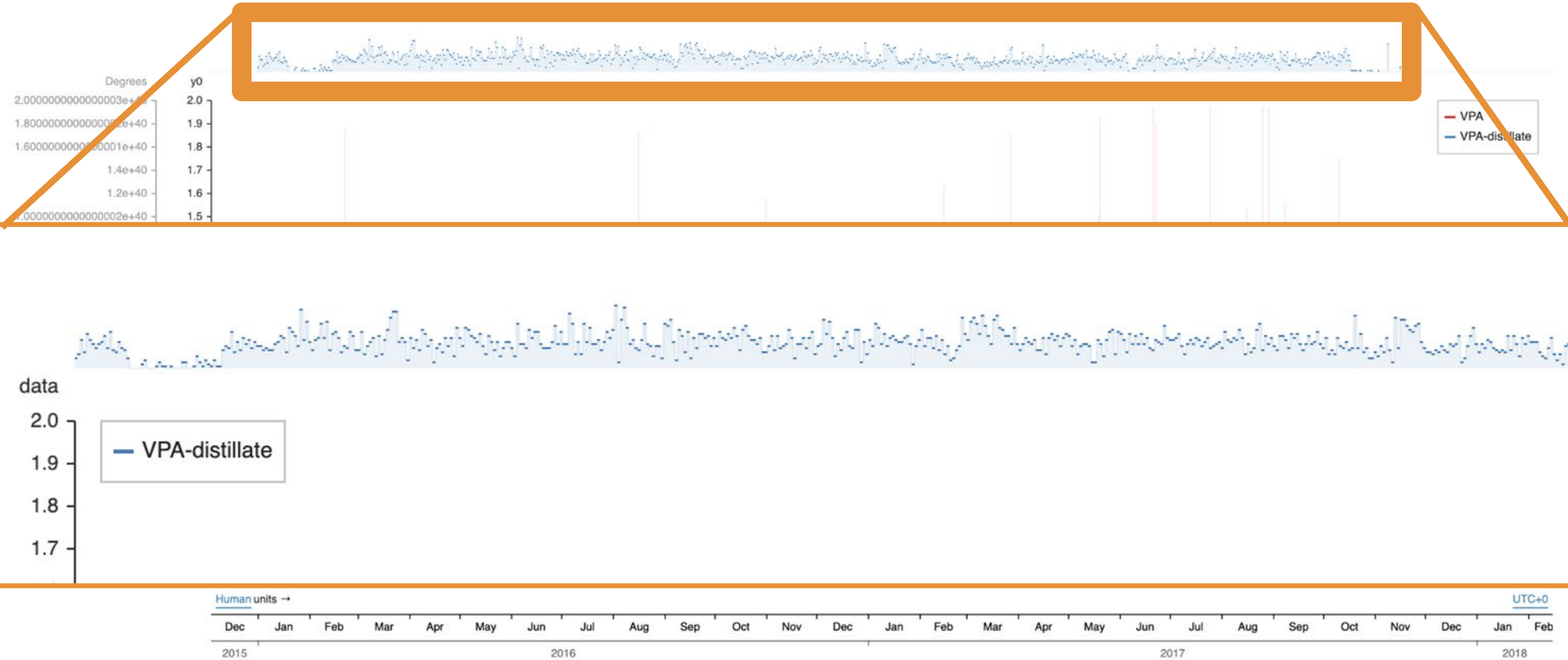
PingThings

Data Quality Distillates - Understand & Flag DQ Issues



PingThings

Data Quality Distillates - Understand & Flag DQ Issues



PingThings

Augmenting Logs - Labelling App

In [1]:
```python
from labeling import *
l = MyLabels(user='john')
```

You have been assigned batches 1 to 12. Each has 25 events

You are at the beginning of your labeling exercise. No previous saved work found

1. Did the naive algo separate the streams correctly?

In [2]:
```python
labels_q1 = annotate(**l.get_kwargs_q1())
```

Current PlotId: 0

yes | no | almost | mostly not | I dont see an event

prev | skip

See plot here: https://jupyter.collab.ptpg.dev/user/ramiro/view/notebooks_playground/labeling/plotid0.png

2. Select a `time-range` (x-axis) and a `feature-range` (y-axis) that captures the PMUs that recorded the event (if any):

In [3]:
```python
labels_q2 = annotate(**l.get_kwargs_q2())
```

Current PlotId: 0

x1 0.00

x2 0.00

y1 0.00

y2 0.00

submit | prev | skip

0

3. Which feature (eg `Voltage`, `Current`, etc) do the x,y coordinates provided correspond to? (you can choose multiple features)

In [4]:
```python
labels_q3 = annotate(**l.get_kwargs_q3())
```

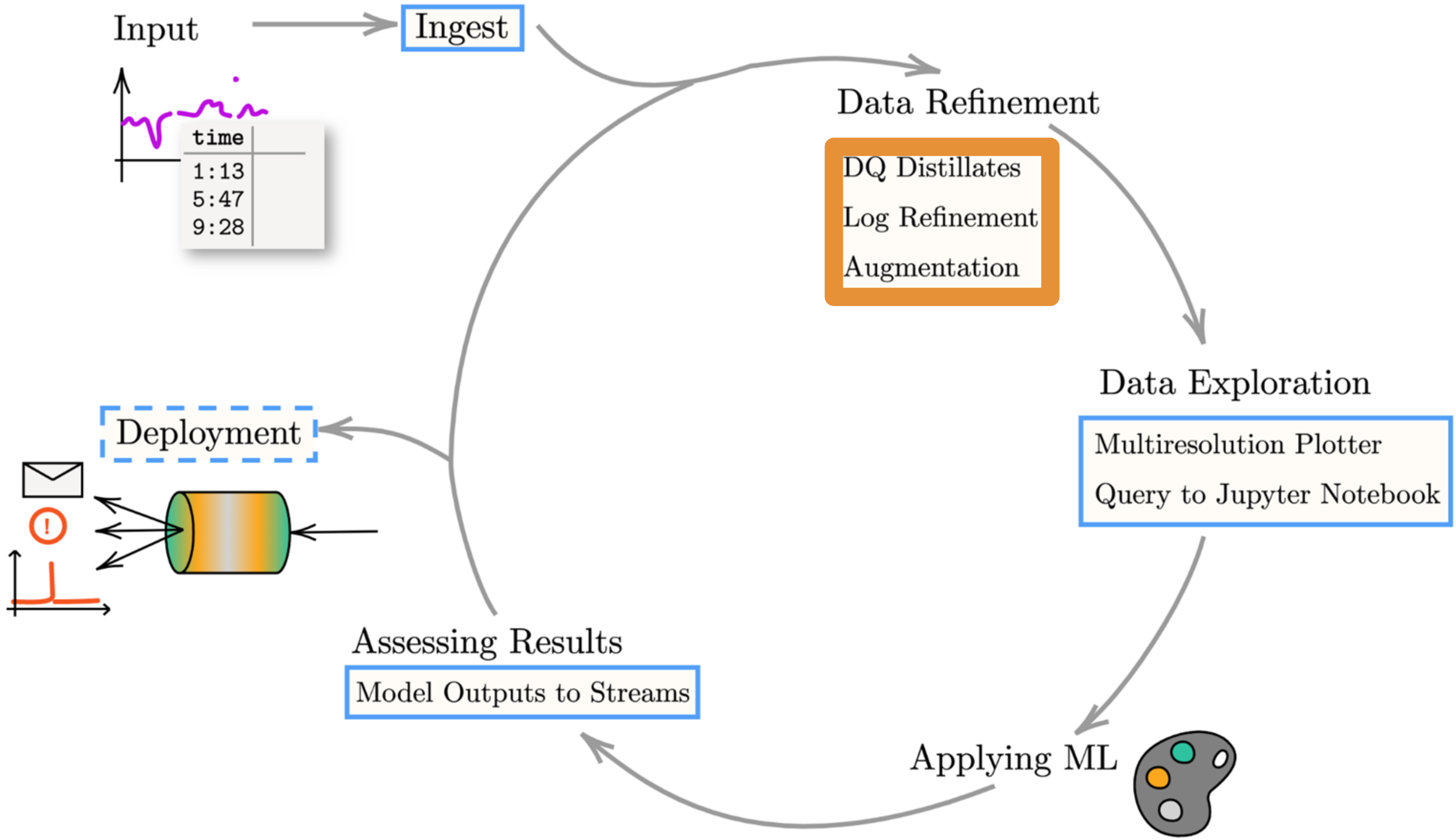
Current PlotId: 0

VPM | IPM | submit | prev

skip

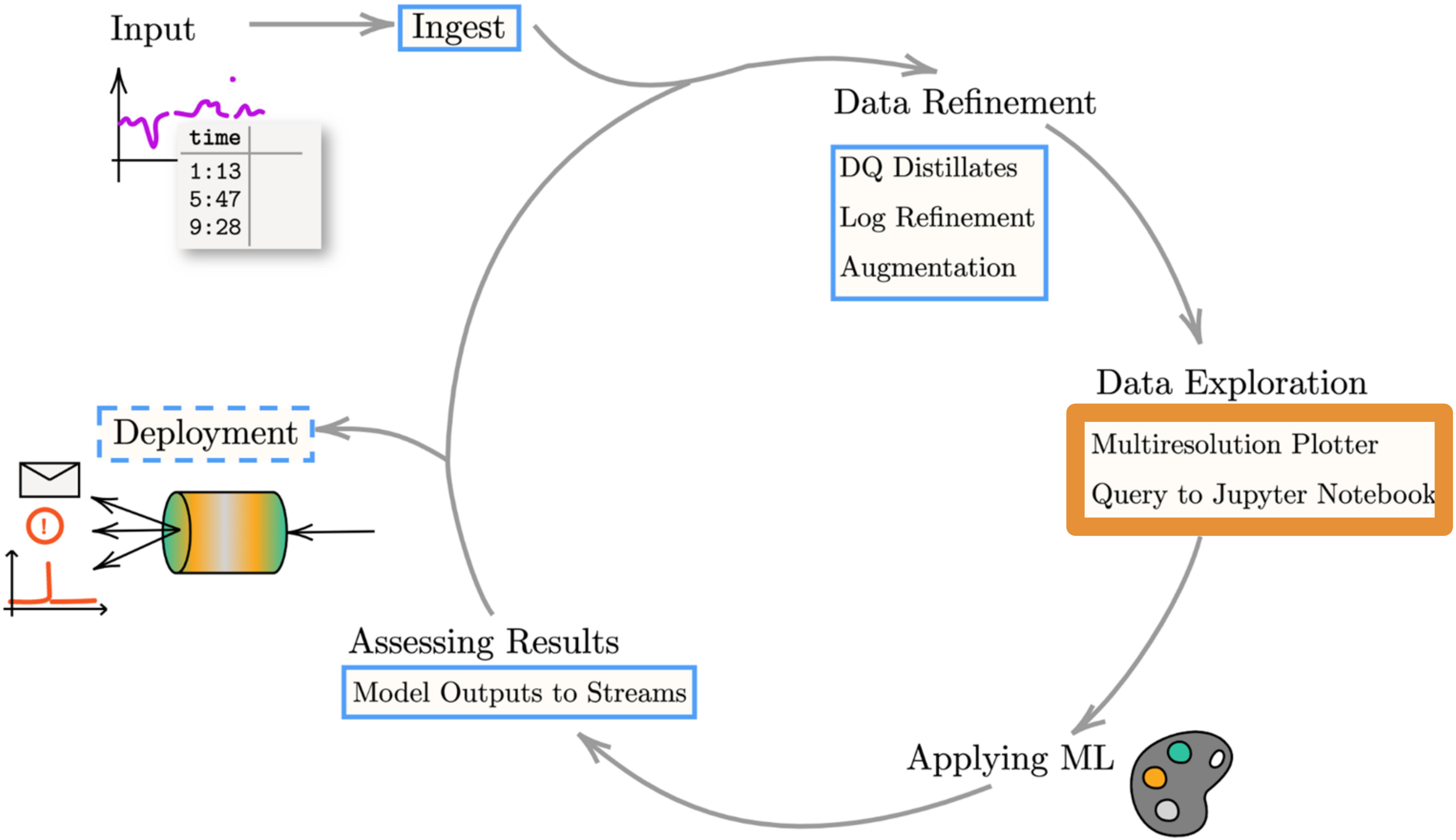
PingThings

An AI Platform for Grid Data



PingThings

An AI Platform for Grid Data



PingThings

Data Exploration

Multiresolution Plotter

PingThings

Data Exploration

Jupyter Notebooks

Power factor at PMU 4

In [142]:
```
# Get streams
uuids = db.query("SELECT collection, name FROM streams WHERE collection like 'sunshine/PMU4'")
names = [row["collection"]+"/"+row["name"] for row in uuids]
streams = db.streams(*names)
print(names)
```

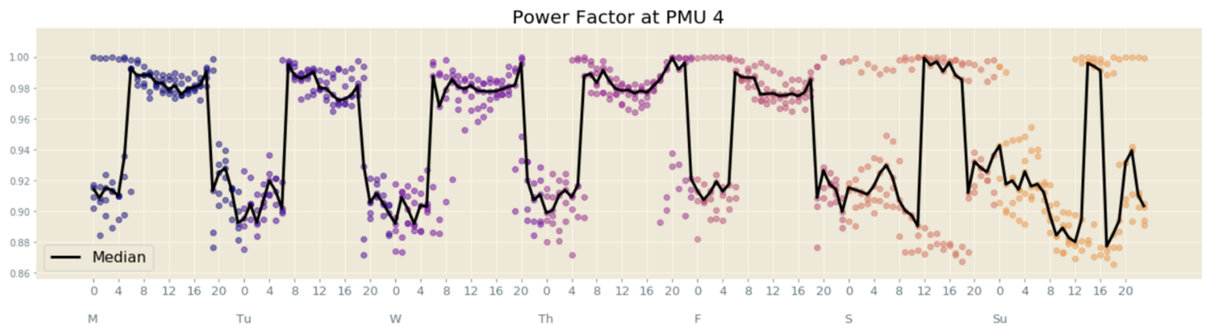
```
['sunshine/PMU4/C1MAG', 'sunshine/PMU4/L1MAG', 'sunshine/PMU4/C2ANG', 'sunshine/PMU4/C2MAG', 'sunshine/PMU4/L3MAG',
'sunshine/PMU4/L2MAG', 'sunshine/PMU4/C3MAG', 'sunshine/PMU4/L2ANG', 'sunshine/PMU4/C1ANG', 'sunshine/PMU4/L1ANG', 's
unshine/PMU4/L3ANG', 'sunshine/PMU4/C3ANG', 'sunshine/PMU4/LSTATE']
```

In [143]:
```
vang_stream = streams[9]; iang_stream = streams[8];
#t0 = ns_to_datetime(vang_stream.earliest()[0].time) + datetime.timedelta(days=10);
# start on monday, july 27, 2015
t0 = datetime.datetime(2015, 7, 27) + datetime.timedelta(hours=7);
```

In [144]:
```
pf_pmu4 = powerfactor_weeks(vang_stream = vang_stream, iang_stream = iang_stream, time = t0)
```

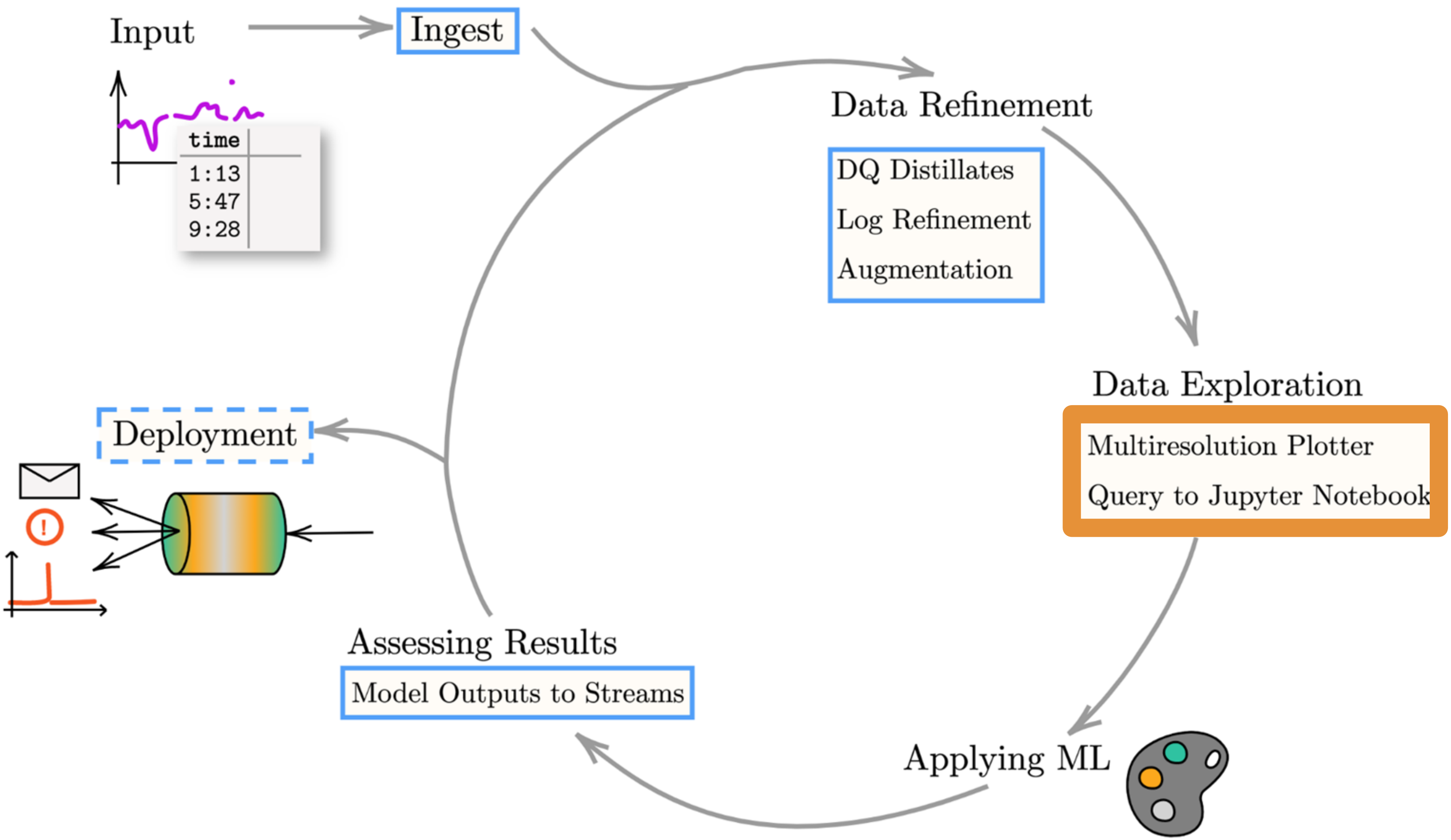
A Jupyter widget could not be displayed because the widget state could not be found. This could happen if the kernel storing the widget is no longer available, or if the widget state was not saved in the notebook. You may be able to create the widget by running the appropriate cells.

In [145]:
```
pfmed_pmu4 = plot_hourly_pf(pf_pmu4);
plt.title('Power Factor at PMU 4', fontsize=20);
plt.tight_layout();
plt.savefig('pf_pmu4', dpi=200);
```



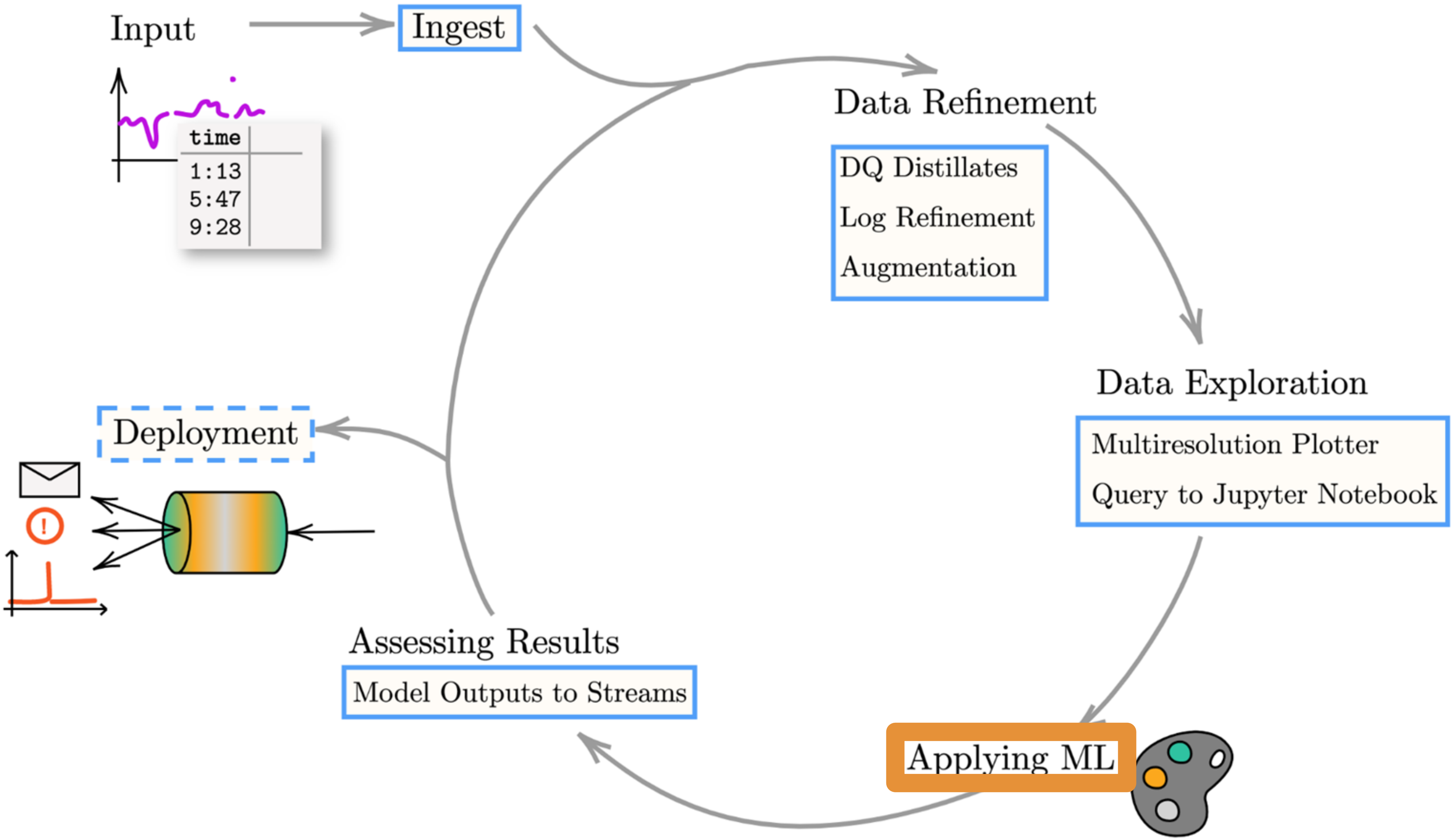
PingThings

An AI Platform for Grid Data



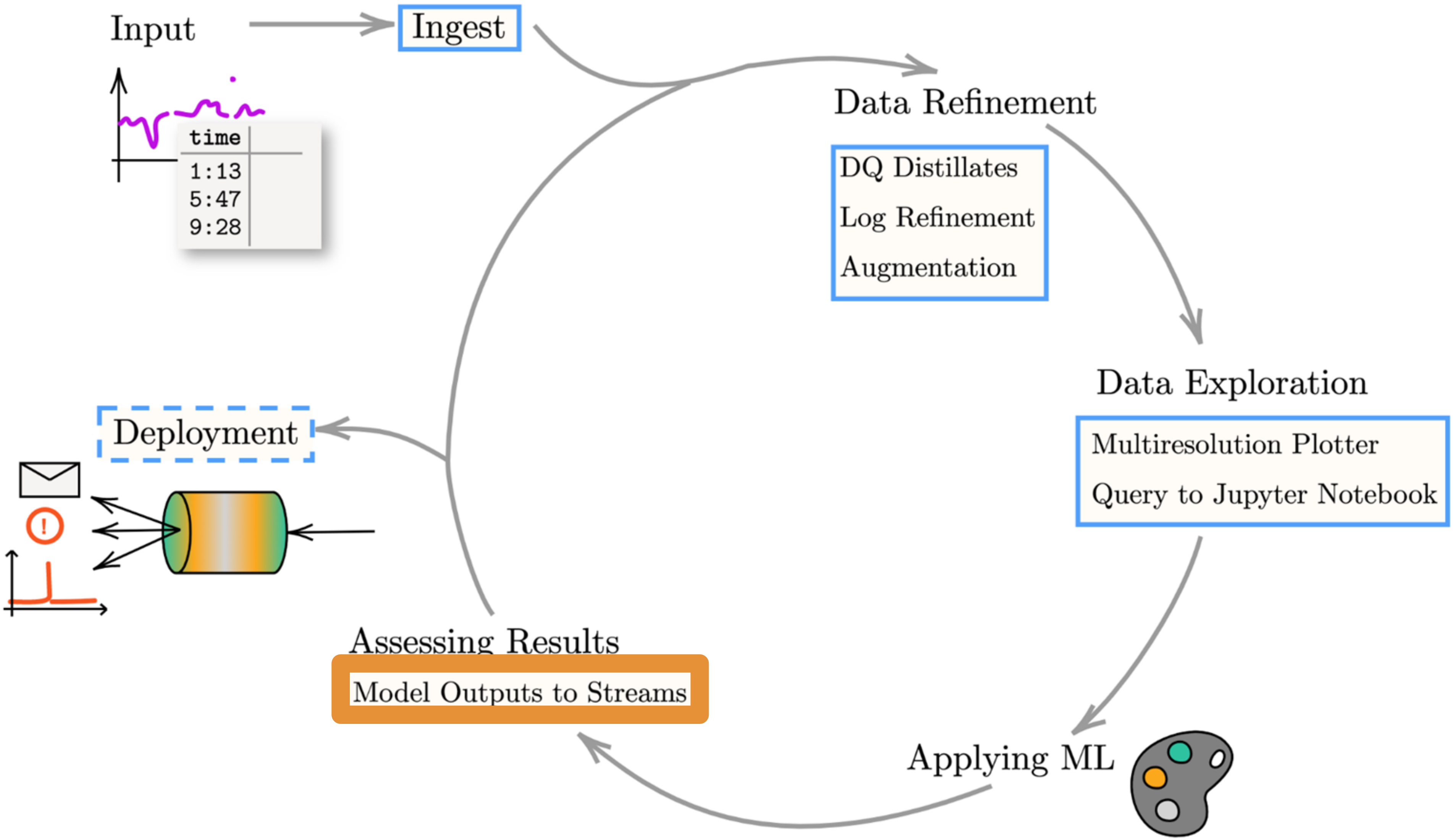
PingThings

An AI Platform for Grid Data



PingThings

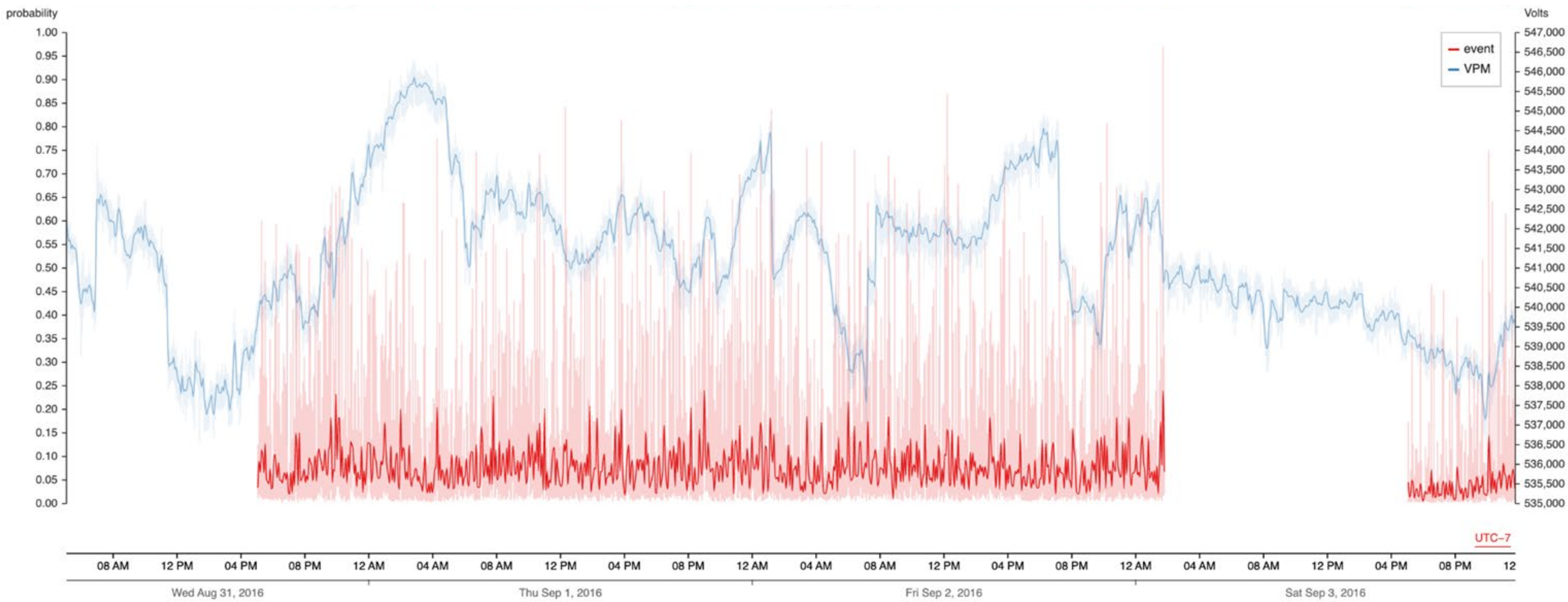
An AI Platform for Grid Data



PingThings

Assessing Results

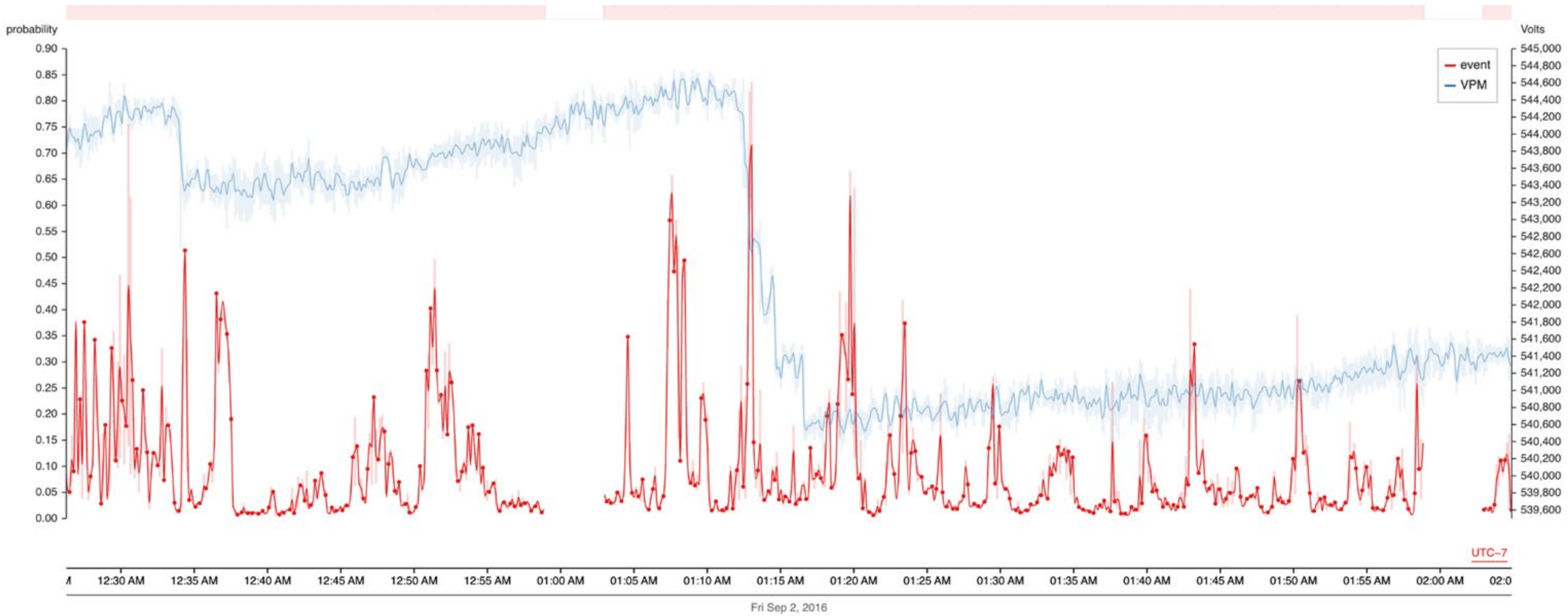
Model Outputs to Streams



PingThings

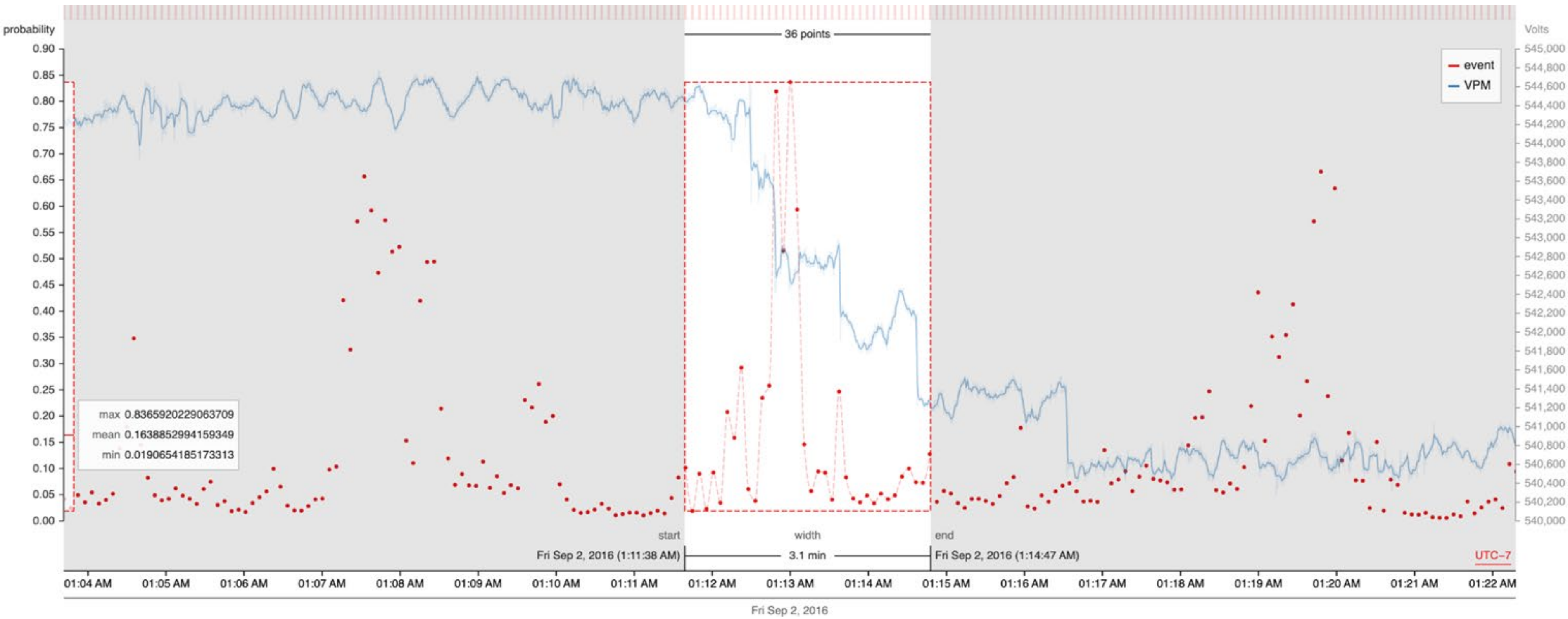
Assessing Results

Model Outputs to Streams

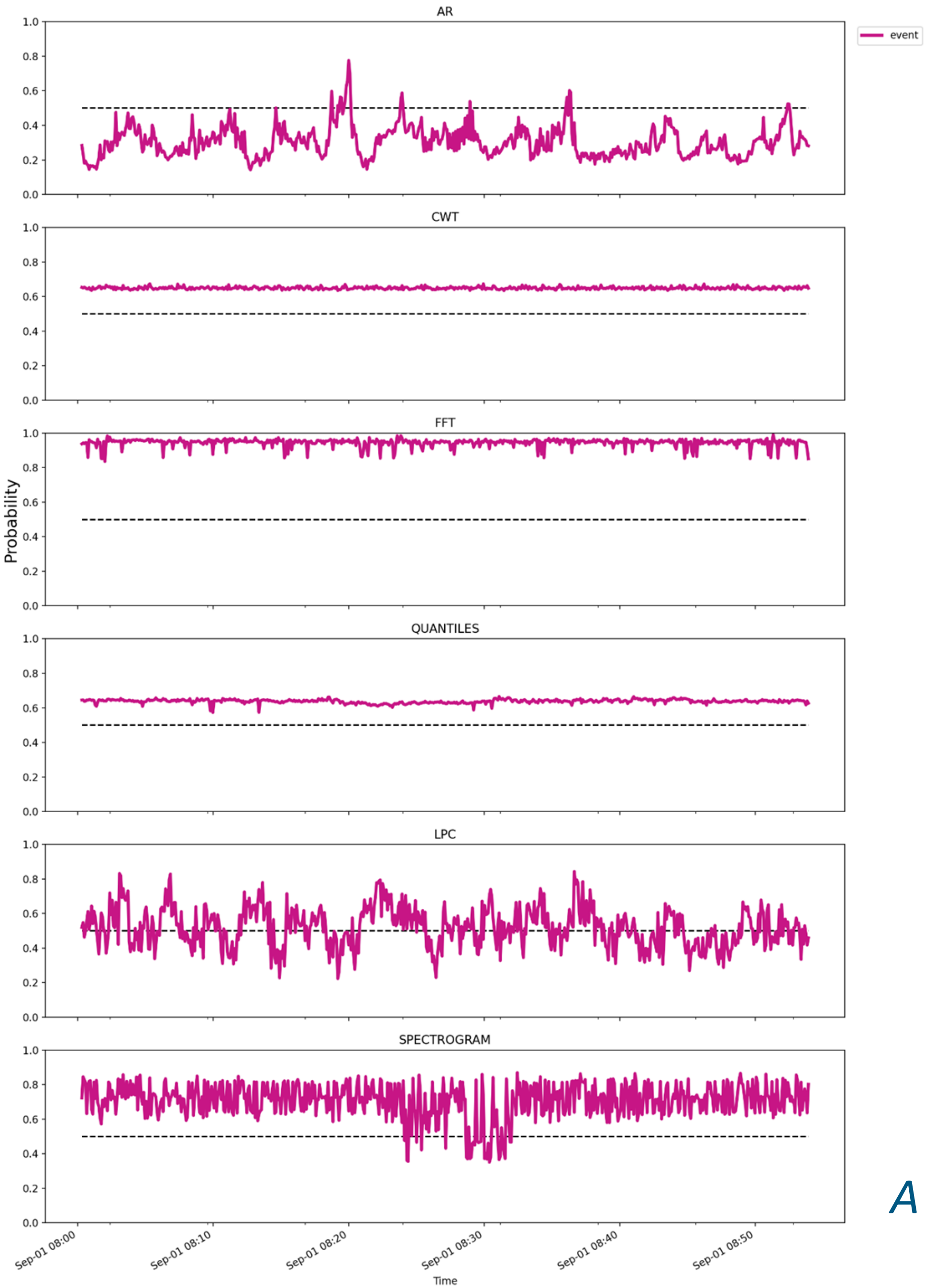


PingThings

Model Outputs to Streams

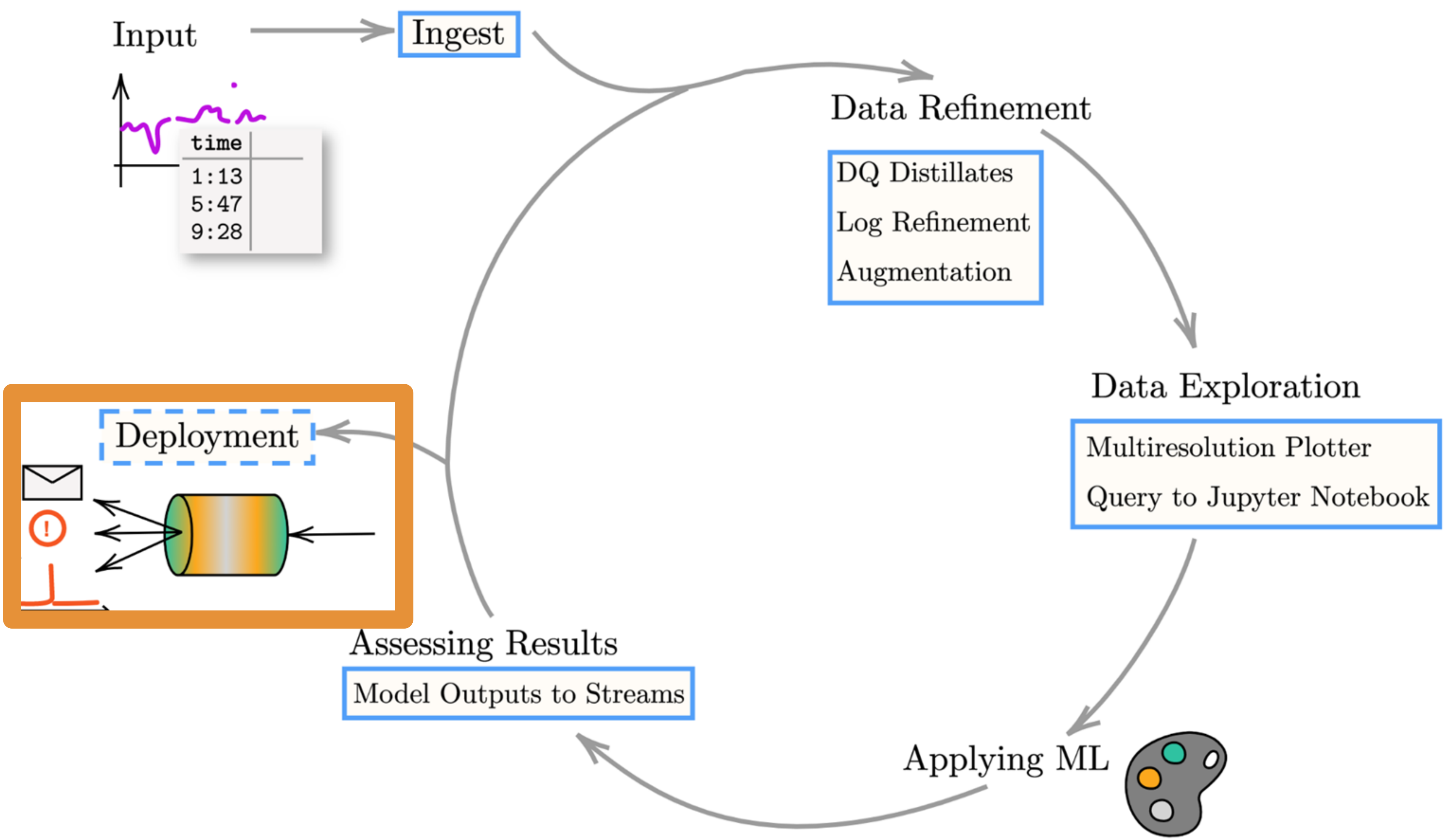


Assessing Results



A enormous need for results exploration

An AI Platform for Grid Data



PingThings

Value of Work

Tools

- In 5 years: Operators & engineers collaborate with ever stronger ML models.
- Many use cases! Broadly useful tools that can be tuned & applied to various applications & datasets.

Data

- Sharing anonymized data is critical! This dataset makes the impossible, possible.
- Real-world data from multiple contexts enables generalizable, efficacious algorithms.

Readiness for Commercialization

0 10

Ready for use:

- Tools for most stages of the analysis & design cycle.
- Refinement is always possible, but may best be done on the job.

Work with *you* on application to specific use cases.

PingThings

Being Ready for ML & BD Analytics

Difficulties

- More domain expertise can help with feature selection.
- Need more standardized assessment.
- Labels have poor temporal specificity & no spatial information.
- What is normal? More work to distinguish significant from inconsequential.

Recommendations

- Prepare data for ML: More measurements, define & save <u>standardized records.</u>
  - Records on how problem was discovered (measurements, call, manual)
  - Which streams revealed an issue?

Lessons Learned and Next Steps

Dataset should be made open access for further work and *accessible.*
For example, in the NI4AI project: https://ni4ai.org/

Next steps

- Greater focus on algorithmic transparency and visualization.
- Enabling feedback for learning on the job.

PingThings

Extra Slides

Extra slides after this point.

Diversity of Model Types

Model Type	Description
CatBoost	Model consisting of ensemble of weaker classifiers, usually decision trees.
Decision Tree	Non-parametric model consisting of layered, simple decision rules.
Extra Trees	Fits a number of randomized decision trees on various sub-samples of the dataset. Uses averaging to improve the predictive accuracy and control over-fitting.
Gaussian Naive Bayes	Fits Gaussian distribution to data using Bayesian methods.
K Neighbors	Non-parametric method classifies new sample based on k nearest training samples.
LGBM	Model consisting of ensemble of weaker classifiers, usually decision trees.
MLP	Multi-layer perceptron, a type of neural network.
MLP Deep	A deep multi-layer perceptron.
Random Forest	Fits an ensemble of decision trees, which vote to produce a single classification.
SGD with hinge	Fits a linear classifier using stochastic gradient descent. The optimized loss function is the hinge loss.
SGD with log	Above, but the optimized loss function is the log loss.
SGD with modified Huber	Above, but the optimized loss function is the modified Huber loss.

Diversity of Features

Feature Family	Feature Types	Number of Features
Auto-regressive	Aggregated Autocorrelation	186
	AR Coefficient	186
	Aggregated Linear Trend	186
Quantiles	Quantiles	207
	Change Quantiles	207
	Coefficient of Variation	207
Continuous Wavelet Transform (CWT)	CWT using the Ricker aka Mexican Hat Wavelet	180
Fast Fourier Transform (FFT)	FFT	1200
Spectrogram	Spectrogram	405

Publications

-

INSERT ORG LOGO (Optional)