INL/CON-23-72799-Revision-0

Machine Learning in Power System Operations: Training Data

July 2023

Stephen Arthur Bukowski, Christopher Roger Sticht

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/CON-23-72799-Revision-0

Machine Learning in Power System Operations: Training Data

Stephen Arthur Bukowski, Christopher Roger Sticht

July 2023

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

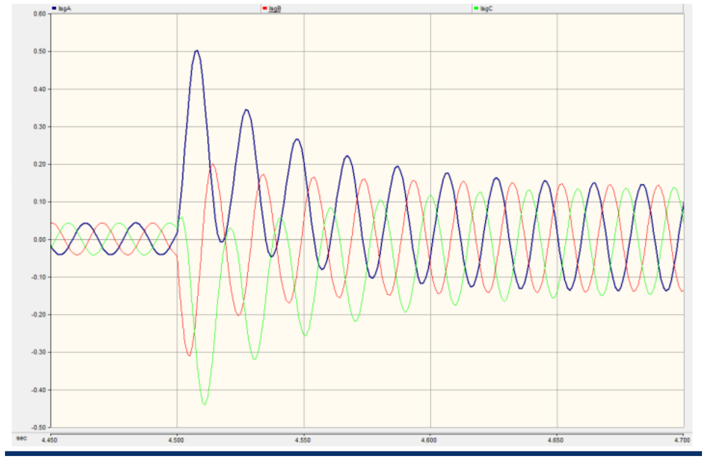
Machine Learning in Power System Operations: Training Data

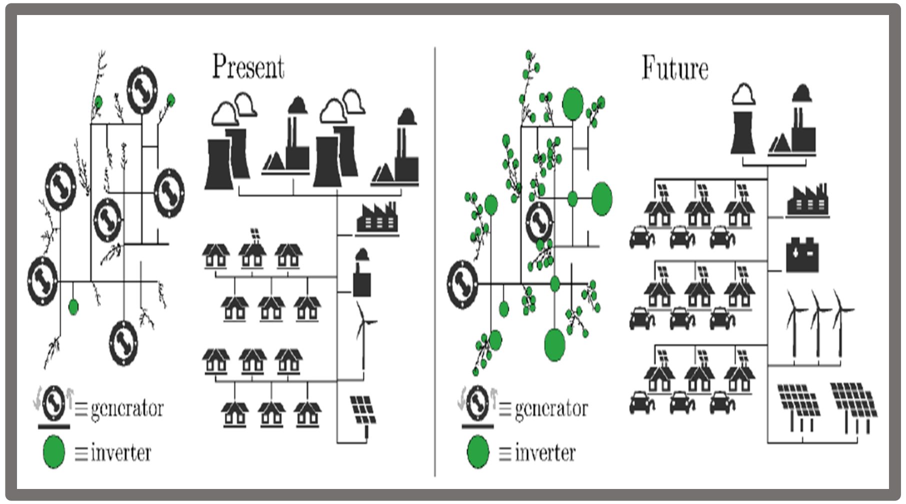
- "If we knew what it was we were doing, it would not be called research, would it?" - Albert Einstein

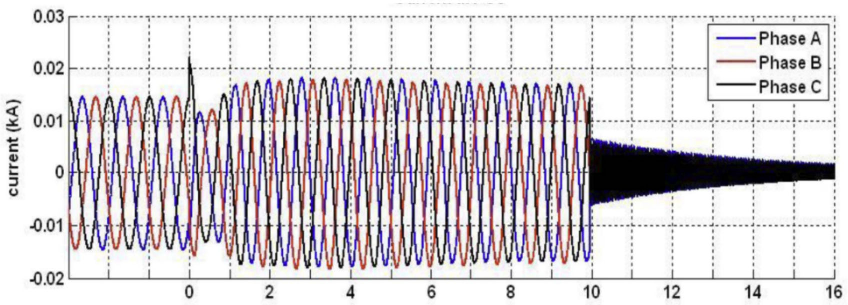
- “The important thing is not to stop questioning.”
- - Albert Einstein

•Transition to a Renewable Grid:

- **Grid or Power Systems operations today are based upon solution and methodologies tied the physics and electrical characteristics of synchronous generators, as we transition to a carbon free renewable grid, the physics and electrical characteristics of Inverter Based Resource (IBR) are not the same as synchronous generators, ultimately eroding effectiveness of existing methodologies.**







Fault current contribution by generators can be 6 to 10x

Challenge

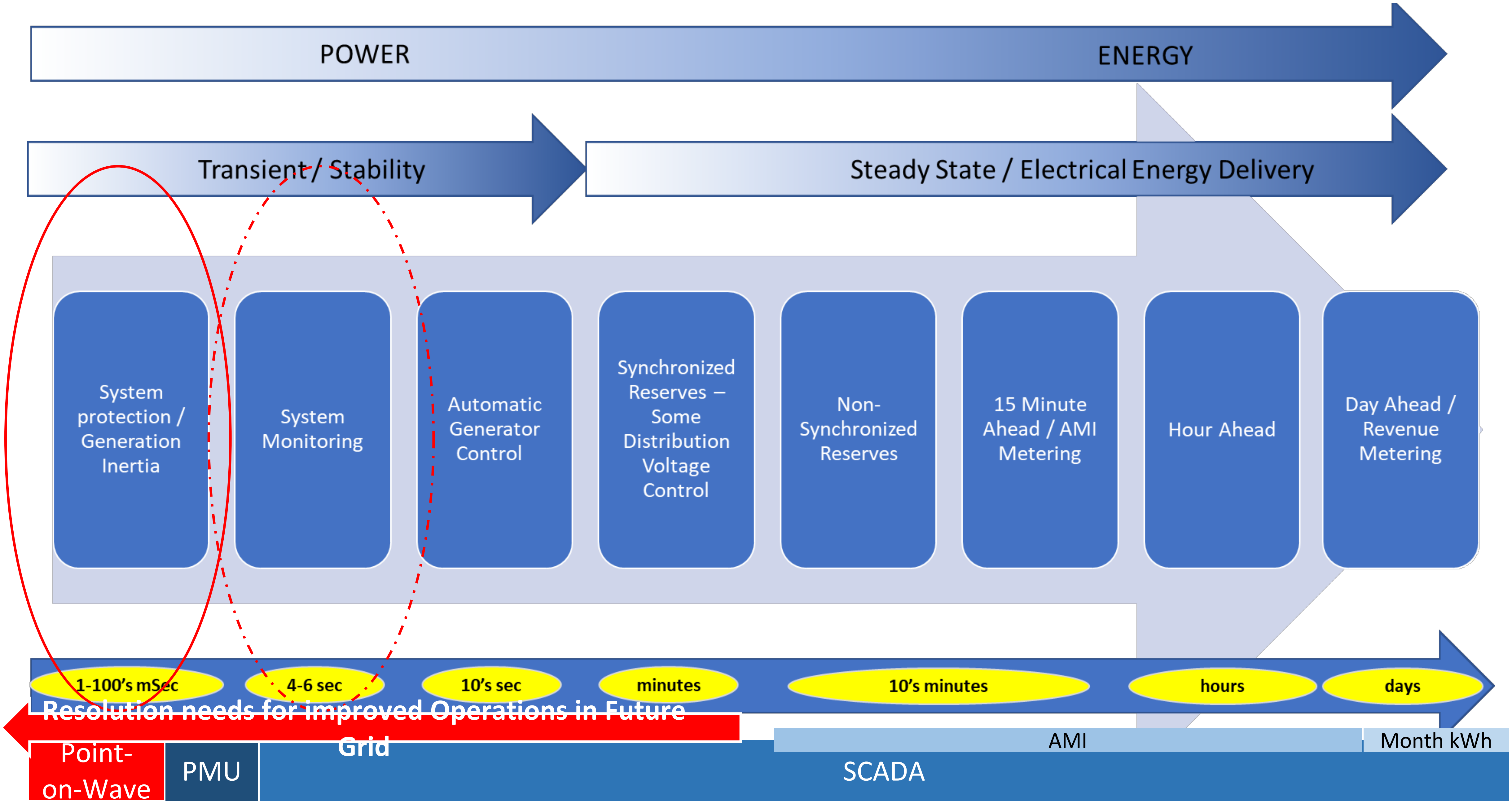
Opportunity

Fault current contribution by IBR generators are typically1 to 1.2x

Sensing local and Act local
– today's protective relaying

Sensing Everywhere and Act local
– tomorrow's protective relaying

Remove assumptions and shift traditional approaches, can real-time ML provide a lens to see the challenge differently



POWER

ENERGY

Transient / Stability

Steady State / Electrical Energy Delivery

System protection / Generation Inertia

System Monitoring

Automatic Generator Control

Synchronized Reserves – Some Distribution Voltage Control

Non-Synchronized Reserves

15 Minute Ahead / AMI Metering

Hour Ahead

Day Ahead / Revenue Metering

1-100's mSec

4-6 sec

10's sec

minutes

10's minutes

hours

days

Resolution needs for improved Operations in Future Grid

Point-on-Wave

PMU

SCADA

AMI

Month kWh

• Areas of Data Application

• Data Landscape

- Monthly – Main stay of billing and growth predictions and system planning
- SCADA – Distribution Management, Transmission situational awareness, widely used and accepted data source
- AMI – Primarily for Monthly Billing, Starting to be used for Distribution Planning & Operations
- PMU –Situational Awareness (Phase Angle) – improving stability tools
- Point on Wave (POW) – leveraging today's technology to improve and advance system operations and protection via novel approaches, machine learning, and artificial intelligence.

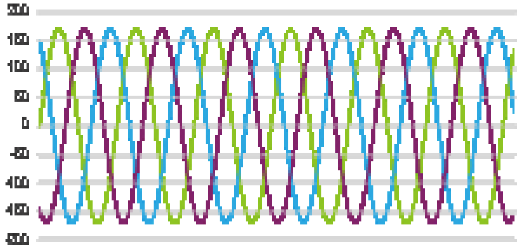
- **Machine Learning Opportunity for Power Systems Operation**

- # High Level Approach – Opening the Door to ML in PSO

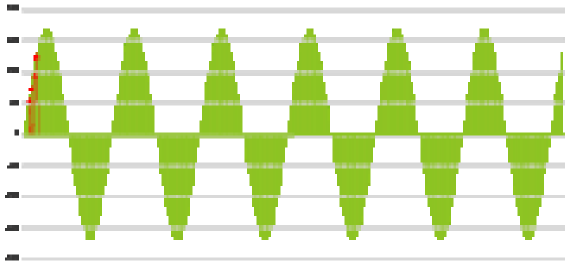
Transformation

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \frac{\cos(n\pi t)}{L}$$

Time Varying analog signal from a sensor

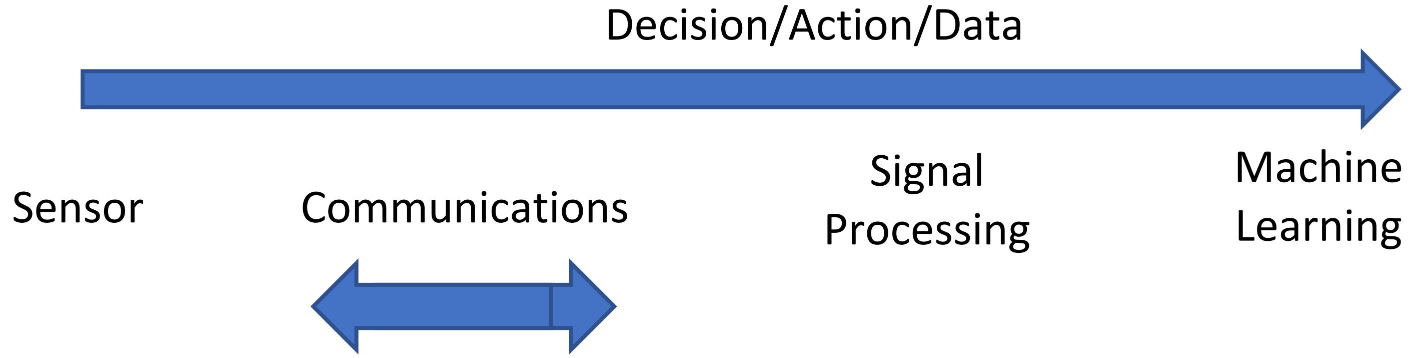


Digitally sampled time varying signal



Machine Learning Techniques

M-dimensional image


Decision/Action/Data

Sensor

Communications

Signal Processing

Machine Learning

• No machine learning without good data sets:

- All MLTs require training data and the *quality of the data* is critical to its performance, accuracy, and complexity of the algorithm.
- Where have we seen success of ML? *Google voice, Amazon Alexa, Facebook – chatbot Army, Twitter – Curated Timelines, IBM –HealthCare*
- What do they all have in common? Ability to offer products and services to the world either free or at low cost. Huge data sets are being used to develop machine learning that make our lives easier. => *Siri, Alexa*, etc....
- Large training and testing data set for Power Systems Operations will facilitate the development of algorithms to support the power system of the future

Training and Testing Data for Power Systems Operations: Today

- **<u>Data for ML is challenging in power systems</u>**: Challenges include limited availability and tend to be unstructured, multi-modal, heterogeneous, decentralized, and highly nonlinear in nature based upon the owner of the data and decisions around timing, selection of data type, and locational limits.
- Data sets needed for transients and stability are challenged by *variety, volume, velocity, and low veracity.*

1) Data variety – multiple sources of unstructured and semi-structured data that typically require preprocessing required to derive useful meaning and supported metadata
2) Data volume – high volume, number of devices x sample rate x duration
3) Data velocity – rate of sample for the data
4) Data veracity – ratio of meaningful data to non-meaning full data represented for example by the seldom occurring faults versus duration of normal operations

Training and Testing Data for Power Systems Operations: Today

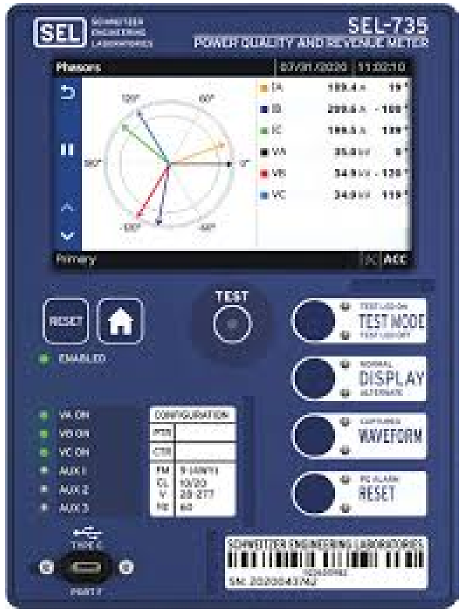
- In other words:

We have little public data and what data is available:

1. No standards around data needs at this level
2. Very few devices record data at high sampling rates.
3. Realtime – does not exist.
    - Problem was thought of as too expensive in storage (memory) and communication (bandwidth)
4. RMS and phasors have been the backbone of PS, but ignore significant amounts of information
5. Lack of labeled datasets
6. General lack of data collection standardization
7. Lack of ground-truth labels.

•Challenges and barriers

- Commercial equipment is not readily available for this application. (60Hz, RMS, Phasors)
- Equipment in the utility has not been designed to sample, transmit, and collect this type of data.
  - Meters, relays, and digital fault recorders have access to the critical analog data or "ground truth" information - the secondaries of Voltage and Current Transformers which as scaled versions of the voltage and current on the system.


SEL
SEL-735
POWER QUALITY AND REVENUE METER
Phasors
TEST
RESET
TEST MODE
DISPLAY
WAVEFORM
RESET
SCHWEITZER ENGINEERING LABORATORIES

•Challenges and barriers

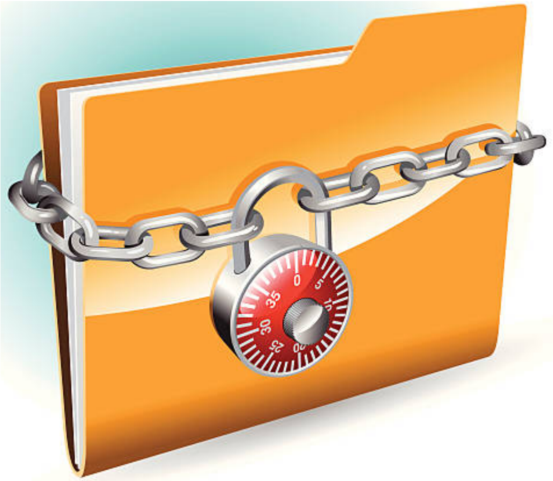
- High sampling rates needed are not supported in current standards for transmitting, writing, and reading files.
- State of the Art: Currently C37.118 defines synchrophasors at a maximum rate of 1 message per cycle of RMS phasor data

$$RMS = \sqrt{\frac{1}{n}\sum_i x_i^2}$$

RMS – Root Mean Square (average over time)

Phasor – Magnitude and Angle based upon FFT at an assumed 60 Hz (digital signal processing filtered)

Relays and Digital Fault Recorders use higher speed sampling for generating fault files, but often written over and very short. All complicated by having too much meta-data by the owners making them difficult to exchange without NDA's



• Challenges and barriers

Methodologies to integrate to existing systems must be addressed

- Want to make a protection engineer “squirm” in their chair? Tell them you want to put another device in their CT circuits.
- Substation communications have just been evolving from serial data (sub-DS0) data rates to peer-to-peer communications and are in many cases not ready for the needs of high-fidelity point on wave data being transmitted
- Time synchronization is foundational
- Integration must be seamless, passive, and non-disruptive
- Must have methods for writing locally if communications are limited as well as transmitting

•Challenges and barriers

Machine Learning to date has not taken the step to real-time data usage for ML application and environments.

- Training on small data sets against other static training data sets
- Ignoring “normal” conditions or conditions that are similar but not necessarily faulted conditions
- We need to think how to implement ML in PR, our initial goal is to AUGMENT not replace.

- **Existing Data Repositories**
- **Current Work**

- EPRI and Grid Event Signature Library
  - Standardized data formats would decrease ambiguity and guesswork, guarantee quality, boost productivity, and increase industry confidence
    - sampling rates
    - labeling
    - Within file and between sources
  - The industry needs more samples of each event type in order to train machine learning algorithms
  - Normalization (Per-unit values) will tend to improve machine learning accuracy and performance

•Alternative Grid Operations: Addressing the challenges

☑ Development of hardware – Pico Point On Wave (POW)

☑ Established approach for data rates – variable approach

☑ Established transmission capabilities – variable approach

☑ Extended C37.118 and C37.111 to adapt to higher sampling rates

☑ Established formatting: Per Unit approach, minimal meta data

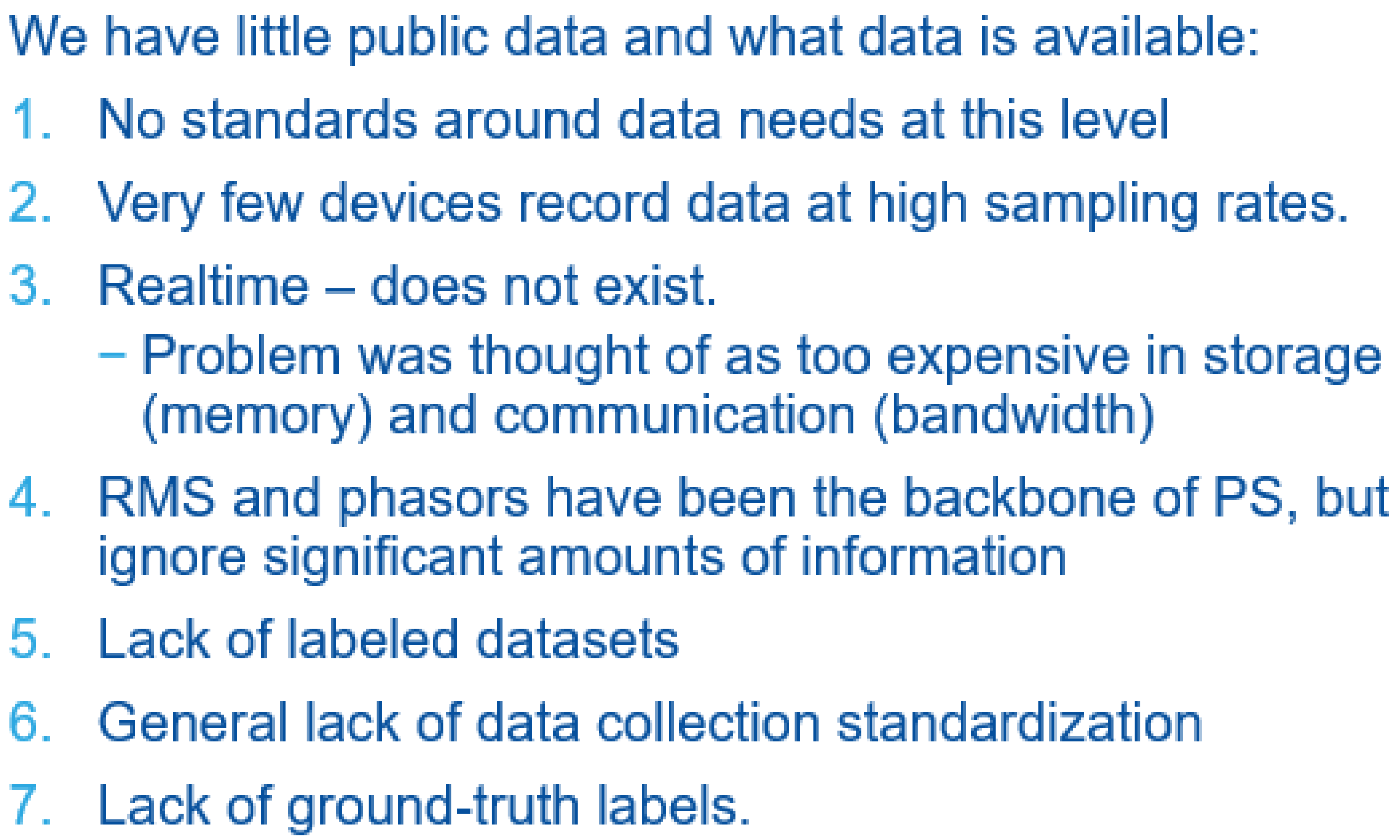
☑ Establish approach to passively adapt to current CT/PT environments

☑ Establish real-time environment – Parity and Failure approach

☑ Developed 100's of thousands of training data sets – Public approach

☑ Established utility feeds with utility partners

• Alternative Grid Operations: Addressing the challenges



We have little public data and what data is available:

1. No standards around data needs at this level
2. Very few devices record data at high sampling rates.
3. Realtime – does not exist.
   - Problem was thought of as too expensive in storage (memory) and communication (bandwidth)
4. RMS and phasors have been the backbone of PS, but ignore significant amounts of information
5. Lack of labeled datasets
6. General lack of data collection standardization
7. Lack of ground-truth labels.

Datasets for Machine-Learning

- **Datasets are essential to the field of machine learning (ML). Meaningful datasets can and have result(ed) in advances in ML. High-quality labeled training datasets for machine learning algorithms are usually difficult and expensive to produce because of the large amount of time needed to label the data. Examples include:**

- MNIST database
  - https://en.wikipedia.org/wiki/MNIST_database
  - http://yann.lecun.com/exdb/mnist/
  - https://github.com/mbornet-hl/MNIST/tree/master/IMAGES/GROUPS
- ImageNet
  - https://en.wikipedia.org/wiki/ImageNet
  - https://image-net.org/
- Common Voice
  - https://en.wikipedia.org/wiki/Common_Voice
  - https://commonvoice.mozilla.org/en/datasets

AudioSet

- https://research.google.com/audioset/

LibriSpeech

- http://www.openslr.org/12

Speech Commands Dataset

- https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

Dataset Labeling

- Dataset labelling, or data annotation, is the process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels that provides context to it, so that machine learning algorithms can use the data to learn.
- Labelling is a critical step in machine learning as it provides context to datasets.
- Labels are used to indicate objects in photos such as a bird or plane, spoken words in an audio recordings, or various tumors in x-ray images.
- Data labeling is required for a variety of use cases including computer vision, natural language processing, and speech recognition.

• Dataset Labeling

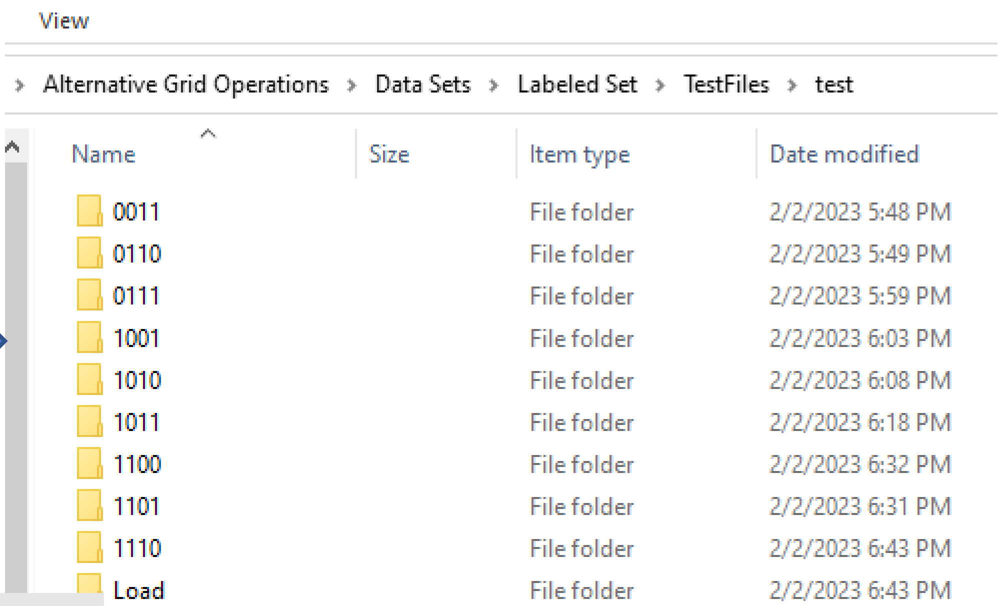
- Labeled data is used to train the machine learning speech algorithms in ***Siri*** and ***Alexa***.
- Labeled datasets were used to train the algorithms that the post office uses to automatically sort hand addressed envelopes in the mail.
- Labeled datasets were used to train algorithms to automatically spot tumors in x-rays.

- **We need labeled waveform data to develop and train power system machine learning algorithms to improve protection and control systems on the power grid.**

• Types of Labeling

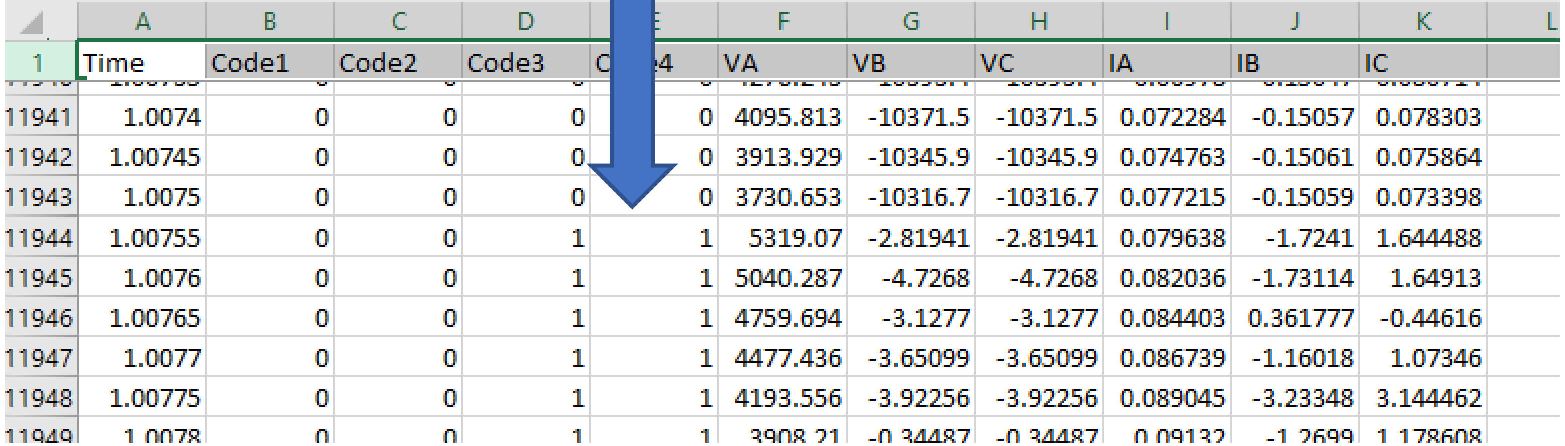
• Take away

- **Labeling can take different forms including:**
- Having a separate file with notable event times within each file as is done in COMTRADE files.
- Organizing event recordings of the same type into common file folders (i.e. all recordings representing load)
- Labeling each timestep with an event recording with relevant information (i.e. when a breaker opens or closes in response to an event)



› Alternative Grid Operations › Data Sets › Labeled Set › TestFiles › test

Name	Size	Item type	Date modified
0011		File folder	2/2/2023 5:48 PM
0110		File folder	2/2/2023 5:49 PM
0111		File folder	2/2/2023 5:59 PM
1001		File folder	2/2/2023 6:03 PM
1010		File folder	2/2/2023 6:08 PM
1011		File folder	2/2/2023 6:18 PM
1100		File folder	2/2/2023 6:32 PM
1101		File folder	2/2/2023 6:31 PM
1110		File folder	2/2/2023 6:43 PM
Load		File folder	2/2/2023 6:43 PM



	A	B	C	D	E	F	G	H	I	J	K
1	Time	Code1	Code2	Code3	C[illegible]4	VA	VB	VC	IA	IB	IC
11941	1.0074	0	0	0	0	4095.813	-10371.5	-10371.5	0.072284	-0.15057	0.078303
11942	1.00745	0	0	0	0	3913.929	-10345.9	-10345.9	0.074763	-0.15061	0.075864
11943	1.0075	0	0	0	0	3730.653	-10316.7	-10316.7	0.077215	-0.15059	0.073398
11944	1.00755	0	0	1	1	5319.07	-2.81941	-2.81941	0.079638	-1.7241	1.644488
11945	1.0076	0	0	1	1	5040.287	-4.7268	-4.7268	0.082036	-1.73114	1.64913
11946	1.00765	0	0	1	1	4759.694	-3.1277	-3.1277	0.084403	0.361777	-0.44616
11947	1.0077	0	0	1	1	4477.436	-3.65099	-3.65099	0.086739	-1.16018	1.07346
11948	1.00775	0	0	1	1	4193.556	-3.92256	-3.92256	0.089045	-3.23348	3.144462
11949	1.0078	0	0	1	1	3908.21	-0.34487	-0.34487	0.09132	-1.2699	1.178608