

# NSF I-DIRSE-IL Cheat Sheet

## Notation, L<sup>A</sup>T<sub>E</sub>X Commands, Terminology, etc.

Waheed Bajwa, Hagit Shatkay, and Christopher Tunnell

Last updated: June 17, 2019

### Physical Detector

#### Terminology

- Detector shall be consistently referred to as *detector*.
  - Other alternatives include *cylinder* and *time projection chamber*, which can be mentioned in the beginning, but shall be rarely used after that.
  - XENON is collaboration, xenon is element.

#### Notation

- Detector as a cylinder:  $\Omega := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 \leq 10^6, -10^3 \leq z \leq 0\}$  [units of mm]
  - $z$  points up toward the sky and is normal to the surface of the Earth.
  - $x$  and  $y$  are in the plane of the sensors.
  - Cylindrical coordinates would be parameterized by  $\phi$  and  $r$  instead of  $x$  and  $y$ , where  $r \in [0, 10^3] \subset \mathbb{R}$  is the distance in millimeters from the center of the detector and  $\phi \in [0, 2\pi) \subset \mathbb{R}$  is the angle.
- Top, bottom, and side of the detector:  $\partial\Omega_T = \{(x, y, 0) \in \Omega\}$  [top],  $\partial\Omega_B := \{(x, y, -10^3) \in \Omega\}$  [bottom], and  $\partial\Omega_S := \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 10^6, -10^3 \leq z \leq 0\}$  [side]
- Any spatial location within the detector:  $\vec{l} \in \Omega \subset \mathbb{R}^3$  [L<sup>A</sup>T<sub>E</sub>X command is `\vec{l}`]

### Physical Sensors Within the Detector

#### Terminology

- Sensors and sensor data shall be consistently referred to as *sensors* and *sensor data*.
  - Other alternatives include *photosensors*, *PMTs*, *channels*, etc., which can be mentioned in the beginning, but shall be rarely used after that.

#### Notation

- Number of sensors at the top and the bottom of the detector:  $n := 248$  [we are ignoring additional sensors that are outside (but adjacent to) the detector for the sake of the proposal]
- Number of sensors on the top of the detector:  $n_T := 127$
- Number of sensors on the bottom of the detector:  $n_B := 121$
- Indexing of sensors:  $j = 0, \dots, n - 1$ , where  $0, \dots, 126$  are the top ones
- Spatial location of  $j$ -th sensor within the detector:  $\vec{l}_j^s \in \partial\Omega_B \cup \partial\Omega_T$  [L<sup>A</sup>T<sub>E</sub>X command is `\vec{l}^s_j`]

# Raw Time-series Data Collected by the Sensors

## Terminology

- Sensors record *luminosity* incident upon them (physically, transported through photons).
- Each sensor gives rise to a *time-series data stream*, which shall always mean the *digitized data* sampled at 100 MHz (one sample every 10 ns).
- Data gathered by all sensors shall be collectively referred to as *spatiotemporal data*.
  - *Style*: We shall write *spatiotemporal*, rather than *spatio-temporal*.
- High-throughput data is another important characteristics of our spatiotemporal data streams.
  - We shall refer to this characteristic as *high-throughput data*, whose meaning shall be clearly explained in the beginning of the proposal.
  - When used generally, this term will refer to any experiment or application that accumulates more than 1 petabyte per year of data.
- Each sensor is connected to a channel. The numbering of the sensor and channel are interchangeable.

## Notation

- The random variable associated with measurements of the  $j$ -th sensor (which is also the  $j$ -th channel to which the sensor is connected):  $C_j \in \mathbb{R}_+$
- Data collected by  $j$ -th sensor at (discrete) time  $t$ :  $c_j^t, t = 0, 1, \dots$ , where  $t = 0$  denotes the first sample,  $t = 1$  denotes the second sample, and so forth.
- Description of sensor data:  $c_j^t = x_j^t + w_j^t$ , with  $x_j^t = 0$  in the absence of any measurable luminosity and  $w_j^t$  is sensor noise
  - Sensor noise seems to have a mixture model, with part of it being Poisson, but it also has impulsive (spikes) and sinusoidal characteristics.
- Collection of data from all sensors at time  $t$ , represented as a vector:  $\vec{C}^t \in \mathbb{R}_+^n$  [L<sup>A</sup>T<sub>E</sub>X command is `\vec{C}^t`]
- Collection of data (non-Euclidean representation) from all sensors at time  $t$ :  $\mathcal{C}^t$  [L<sup>A</sup>T<sub>E</sub>X command is `\mathcal{C}^t`]
  - Notice that linear algebra operations can be applied directly on the mathematical object  $\vec{C}^t$ , which is taken as a vector in  $\mathbb{R}^n$ , but not on the mathematical object  $\mathcal{C}^t$ .
  - The distinction between the two mathematical objects, while subtle, is important in relation to the distinction we want to make in terms of graph-based signal processing and graph-based machine learning.
- Data collected from sensors from time  $t_i$  to time  $t_k$  (both forms):  $\vec{C}^{t_i:t_k}$  [Euclidean] and  $\mathcal{C}^{t_i:t_k}$  [non-Euclidean]

# Physical Interactions Within the Detector

## Terminology

- We shall refer to any particle interacting with any matter within the detector as a *physical interaction*.
- Each physical interaction gives rise to multiple incidences of *measurable luminosity* at the top and the bottom sensors; each measureable luminous incidence at any sensor shall be referred to as a *hit* [we may want to iterate over this language a couple of times for the final version; this only needs to be accurate for the purpose of this proposal].
- Hits recorded at multiple sensors within a short time of each other are collectively referred to as a *peak*.
  - *Integrated hit pattern* of a peak is a collection of all the hits in a peak (usually referred to as *hit pattern*). Another frequently used definition of hit pattern is integrated luminosity measured by all sensors corresponding to a peak. *Hit pattern of a sample* is data collected by all sensors at a certain time. Integrated hit pattern can be acquired by integrating hit pattern of a sample from

- the start time to the end time of a peak.
- Multiple peaks recorded within the detector in a short time are collectively referred to as an *event*.
  - All evidence for a physical interaction of a particle within the detector is therefore captured in terms of an *event data frame*.
  - In the absence of any background *noise*, each event often results in two major spatiotemporal luminous signals (and thus two peaks), which shall be referred to as S1 (typically weaker and lasting for a much shorter duration) and S2 (typically stronger and having wider spread in time compared to S1) signals.
  - Each event data frame typically corresponds to around 300  $\mu\text{s}$  of data.
  - There are typically 20 to 100 events recorded by the detector per second.

#### Notation: High Level

- Particle type that interacts:  $w \in W$  [ $w$  is a variable that can be a WIMP, a neutrino, etc., and  $W$  is all possible particles; unknown] I changed  $\gamma$  to  $w$  since  $\gamma$  is the gamma ray particle. Actually, every greek symbol is a particle.
- Interaction type:  $\xi \in \Xi$  [ $\xi$  is a variable that can, e.g., indicate elastic electronic recoils, elastic nuclear recoils, or other processes such as inelastic nuclear excitation, and  $\Xi$  is all possible interactions; unknown; `\LaTeX commands are \xi for  $\xi$  and \Xi for  $\Xi$` ]
- Spatial location of particle interaction within the detector:  $\vec{l} \in \Omega \subset \mathbb{R}^3$  [unknown]
  - It is best to stick to the terminology of *location* and *localization*, rather than position.
- Energy associated with the interaction:  $\mathcal{E} > 0$  [units of keV; unknown; `\LaTeX command is \mathcal{E}`]
- (Analog) time at which interaction took place:  $T \in \mathbb{R}_+$  [units of ns; can be assumed known; while it can be absolute or relative, we treat it as the Unix Epoch time without loss of generality for computational and practical reasons (i.e.,  $T = 0$  is the start of 1970 in UTC)]
- Distinct interactions to be enumerated using an index like  $i$  on top of any of the above quantities:  $i = 1, 2, \dots$ , with  $i = 1$  being the first interaction recorded by the detector and so forth.

#### Notation: Low Level (Events, Peaks, and Hits)

- The  $i$ -th event:  $E_i$  [ $i$  can be dropped when referring to a particular event]
  - Start and end times of  $i$ -th event:  $t_0^i$  and  $t_1^i$  [ $i$  can be dropped when referring to a particular event]
  - Data collected from all sensors corresponding to  $i$ -th event:  $\mathcal{C}^i := \mathcal{C}^{t_0^i:t_1^i}$  [non-Euclidean] and  $\vec{\mathcal{C}}^i := \vec{\mathcal{C}}^{t_0^i:t_1^i}$  [Euclidean] [ $i$  can be dropped when referring to a particular event]
    - \* Here, non-Euclidean means that the 248-dimensional data at any time  $t$  (corresponding to all sensors) is treated as lying on a graph of 248 vertices, where Euclidean means that the data is treated as lying in  $\mathbb{R}^{248}$ .
- Peaks associated with the  $i$ -th event:  $\pi_1^i, \dots, \pi_k^i$  [ $i$  can be dropped when referring to a particular event]
  - Start and end times of  $k$ -th peak within  $i$ -th event:  $t_{0,k}^i$  and  $t_{1,k}^i$  [ $i$  can be dropped when referring to a particular event]
  - Data collected from all sensors corresponding to  $k$ -th peak within  $i$ -th event:  $\mathcal{C}^{\pi_k^i} := \mathcal{C}^{t_{0,k}^i:t_{1,k}^i}$  [non-Euclidean] and  $\vec{\mathcal{C}}^{\pi_k^i} := \vec{\mathcal{C}}^{t_{0,k}^i:t_{1,k}^i}$  [Euclidean] [ $i$  can be dropped when referring to a particular event]
  - Integrated hit pattern of  $k$ -th peak within  $i$ -th event is  $\sum_{t=t_{0,k}^i}^{t_{1,k}^i} \mathcal{C}^t$ . Top hit pattern is integrated hit pattern where  $j = 0, \dots, 126$ .
- Hits associated with  $j$ -th sensor,  $k$ -th peak, and  $i$ -th event:  $h_1^{j,\pi_k^i}, \dots, h_m^{j,\pi_k^i}$  [ $i$  can be dropped when referring to a particular event]

## Notation (Other Variables)

- There are a number of other derived quantitative data available to us, which we cannot possibly discuss in the proposal. Any variable that is not explicitly defined will be lumped into auxiliary variables  $\theta$ .

## Physical Forward Model

### Terminology

- The term *forward model* refers to the mathematical process that relates the physical process (in this case a particle interacting within the detector) to the output data (in this case, event spatiotemporal data corresponding to that interaction).
  - This forward model (up to a modeling error; see below) is known to us, but is too complicated to mathematically express in an analytical form. However, it can be generated precisely using numerical simulations.

### Notation

- The relationship between a particle interacting and the event data frame is expressed as follows:

$$(\vec{C}, t_0, t_1) := \mathcal{F}(\vec{l}, \mathcal{E}, T, \gamma, \xi) + \mathcal{W} + \Delta.$$

- Note that  $t_0$  and  $t_1$  depend on underlying physical properties of the interaction, and hence explicit mention in the above equation (but it can be dropped later).
- We shall sometimes refer to  $\mathcal{F}(\vec{l}, \mathcal{E}, T, \gamma, \xi)$  as the *noiseless* luminous spatiotemporal data  $\vec{X}$  [LaTeX notation: `\vec{X}`].
- $\mathcal{F}(\vec{l}, \mathcal{E}, T, \gamma, \xi)$  is completely known to us through numerical simulations (formed from an analytical expression, as noted above).
- Just like in all statistical modeling problems, we cannot model all aspects of the detector's hardware. We capture this model uncertainty in the object  $\Delta$  (of appropriate dimensionality) above.
  - \* Notice the difference between  $\mathcal{W}$ , which only models sensor noise, and  $\Delta$ , which models other uncertain aspects of our detector.

## Training Data

### Terminology

- In the case of supervised learning, labeled training data will be generated using numerical simulations.
- In the case of unsupervised learning, unlabeled training data will be generated through both numerical simulations and real experiments.

### Notation

- Supervised learning: We will have access to  $N$  labeled data (at the physical interaction/event level) generated through simulations, expressed as  $\{\vec{C}^i, (\vec{l}^i, \mathcal{E}^i, T^i, \gamma^i, \xi^i)\}_{i=1}^N$  [note that the start and end times of each event  $i$  are being implicitly encoded in the size of  $\vec{C}^i$ ; also, use LaTeX code `\vec{1}^{\sim\{,i\}}` for  $\vec{l}^i$ , as it appears as  $\vec{l}^i$  without the `\`, space]
  - We will use the shorthand notation  $\mathcal{L}^i := (\vec{l}^i, \mathcal{E}^i, T^i, \gamma^i, \xi^i)$  to capture the entire *labeled tuple*  $(\vec{l}^i, \mathcal{E}^i, T^i, \gamma^i, \xi^i)$  into one quantity [LaTeX command is `\mathcal{L}`]
- Unsupervised learning: We will have access to  $N$  unlabeled data (at the event level) generated through both simulations and real experiments, expressed as  $\{\vec{C}^i\}_{i=1}^N$

- We will use  $\widehat{\cdot}$  to distinguish between data and labels from numerical simulations versus real experiments; e.g.,  $(\vec{\widehat{C}}^i, \widehat{\mathcal{L}}^i)$  [data/labels from numerical simulations] versus  $(\vec{\widetilde{C}}^i, \widetilde{\mathcal{L}}^i)$  [data from real experiments]
- Experimental data has  $(\vec{l}^i, \mathcal{E}^i, T^i, \gamma^i, \xi^i)$  unknown or partially known (e.g., just  $\mathcal{E}^i$ ), but we may know the statistical properties of the data. For example, the probability density function  $f(\vec{l}^i)$  may be known. The probability  $\Pr[a \leq X \leq b] = \int_a^b f_X(x) dx$ . For example, if  $f$  is normally distributed then  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ .

## Other Terminology

- Our algorithms are both *data science and machine learning algorithms*. In the interest of pithiness, we will sometimes only use the term *data science*, which will subsume *machine learning* within it [we shall say something like this explicitly in the proposal].
- Our approach can be called many different things and we should stick to **Science-aware data science and machine learning**.

## Other Notation

- Our sensor data have a graph structure, given by:  $\mathcal{G} = (\mathcal{V}, \mathbf{A})$ , with  $\mathcal{V} = \{0, \dots, n-1\}$  representing the sensors and  $\mathbf{A}$  representing a weighted adjacency matrix.

## Other Auxiliary Information

- Digitizers generate 14-bit unsigned data (positive valued data).
- Location reconstruction accuracy should be  $\pm 5$  mm or less; another important goal is very restrictive confidence intervals (methods with very small standard deviation). Said differently, as this is a rare event search, it is better to have a resolution of 1 cm and no mismeasurement of 10 cm than a resolution of 1 mm with occasional 10 cm misreconstruction. Current state of the art is a few mm to 1 cm, but these are hard to verify since people quote average L1 loss.
- Energy reconstruction accuracy should be  $\pm 0.5\%$  or less (ideally smaller than this). The current state of the art is 1.2% from EXO. NEXO aims for 0.5%. The statistical limit is 0.3%.
- **Define Kr83m and other sources.**