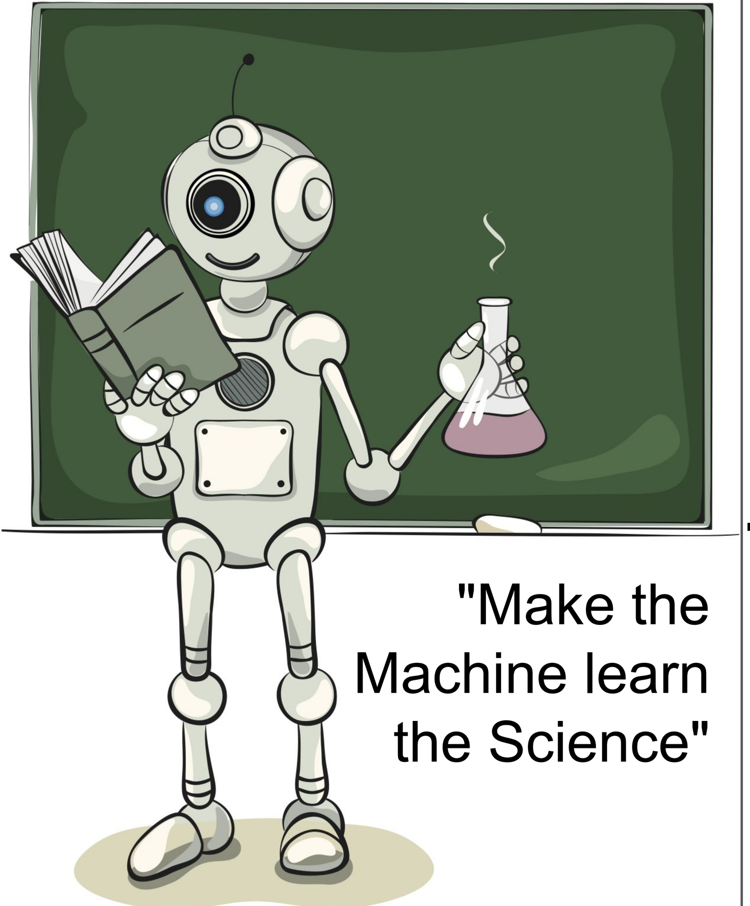
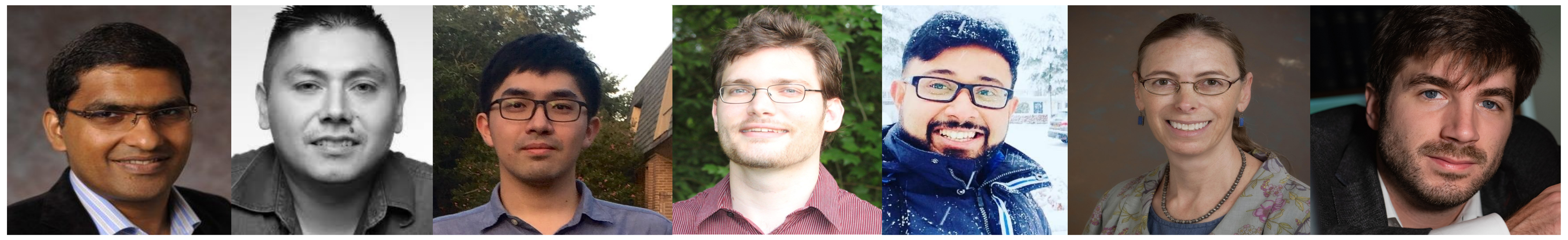


"Make the Machine learn the Science"

Data-Intensive Discovery Accelerated by Computational Techniques for Science (DIDACTS)

Bajwa, Waheed (Rutgers); Higuera, Aaron (Rice); Liang, Shixiao (Rice); Pergeorelis, Michael (U. Del); Roy, Venkat (Rutgers); Shatkay, Hagit (U. Del); Tunnell, Chris (Rice)



RICE

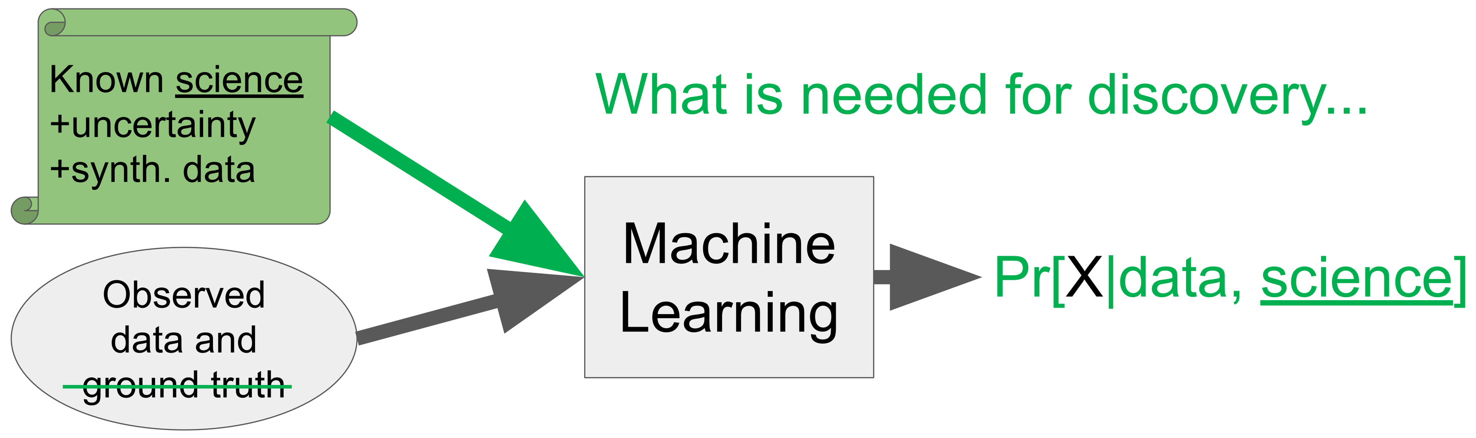
RUTGERS

UNIVERSITY OF DELAWARE

Data Science Problem Faced by Physical-Sciences:

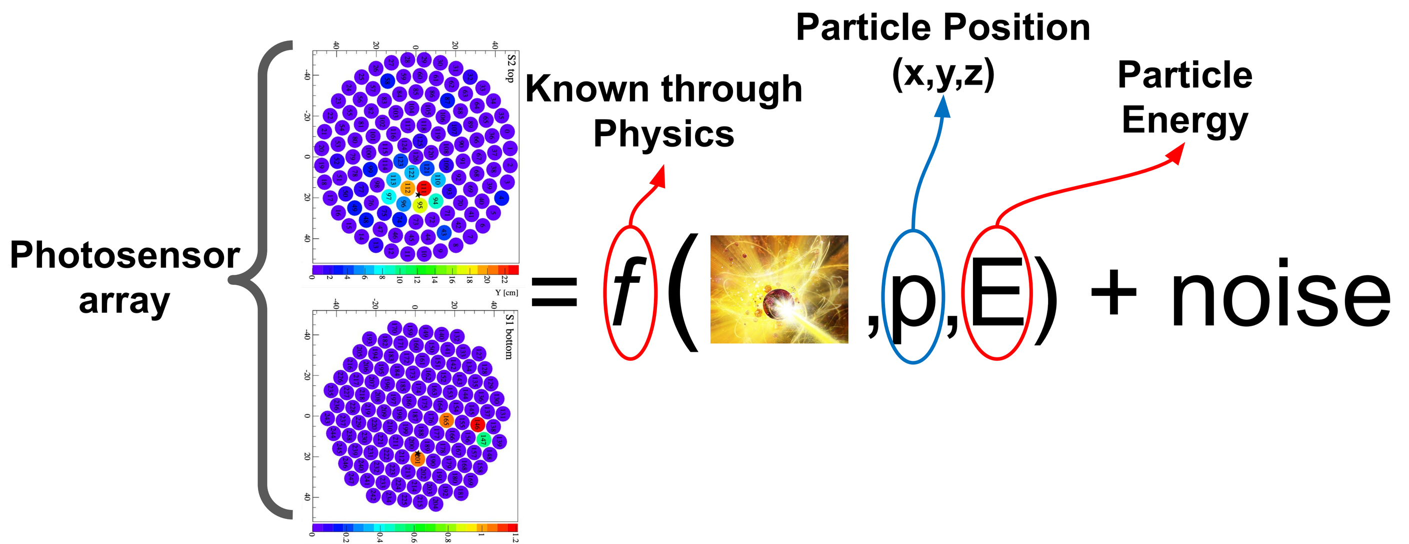
Physical sciences involve complex measurements necessitating methods that:

- Thrust 1: Incorporate domain knowledge into data models and inference methods:
  a. Robust solutions adhering to physical knowledge
  b. Faster inference using limited (or absence of) labeled data
- Thrust 2: Directly model uncertainty in both domain knowledge and data
  a. Readily handles noisy, missing and unexpected input
  b. Presents results in terms of (joint) probability distributions



Thrust 1: Domain-enhanced, Graph-regularized Inverse Problems

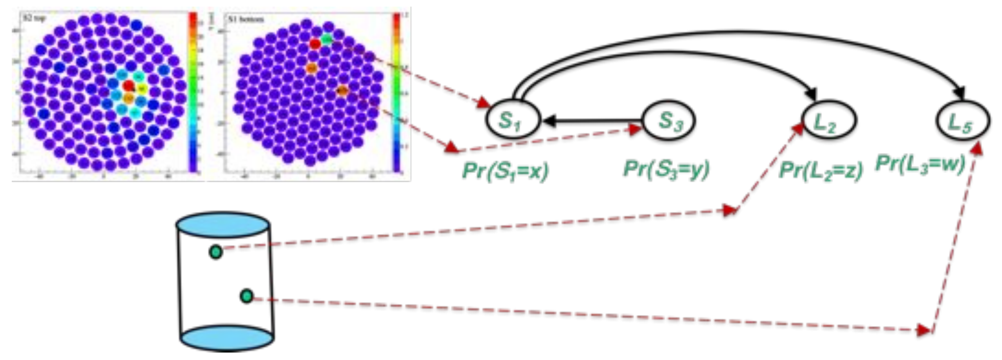
1. Development of a forward ***measurement model*** incorporating the ***underlying physics*** of the particle interactions as observed by the photo-multiplier (PMT) sensors.
2. Development of fast graph-regularized inverse problem solvers for the forward model.



Thrust 2: Probabilistic Graphical Models

Capturing inherent ***physical constraints*** and ***domain uncertainty*** via sparse probabilistic graphical models

Sparse Constrained Graphical Models



Nodes: Random Vars: Sensors & Tank Regions
Edges: Interdependence among Sensors/Regions/Events
Introducing Domain Knowledge via: *Priors; Distributions; Interdependency constraints*
Inference in Graphical Models ➡ ***Pr[X | Data, Model]***

Example of a *Hard* Physical-Science Problem:

Science:

1. What is Dark Matter, and why is it that 85% of the matter in the Universe? (link)
2. Why is the Universe made of matter and not anti-matter, does this relate to nature of neutrino mass? (link)

Experimental challenge, in vat of xenon observed by 248 sensors, perform:

1. Extreme rare-event search: needle in a haystack the size of Texas
2. Extreme precision energy measurement
3. Petabytes of data, can never have false positive

With the most sensitive dark-matter detector that also measured longest half life ever, the science is still limited by the ability to infer precisely and accurately positions and energies of faint energy depositions (intensity and location of S1 and S2).

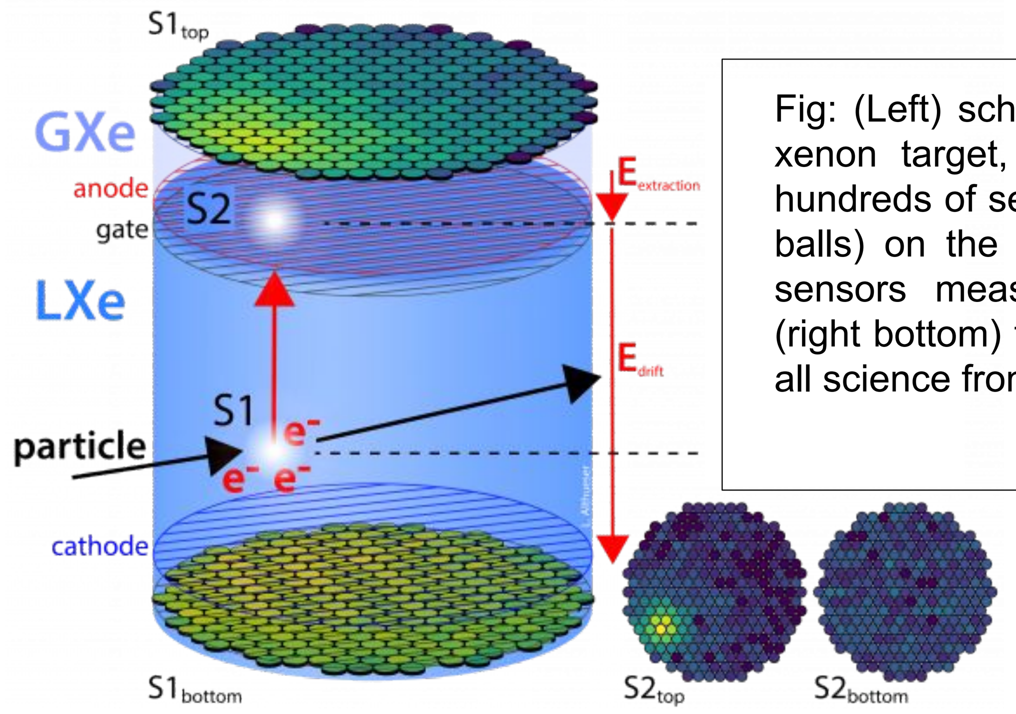


Fig: (Left) schematic of 1.5 m x 1.5 m xenon target, which is instrument by hundreds of sensors (visually shown as balls) on the top and bottom. These sensors measure different intensities (right bottom) that are the foundation of all science from these instruments.

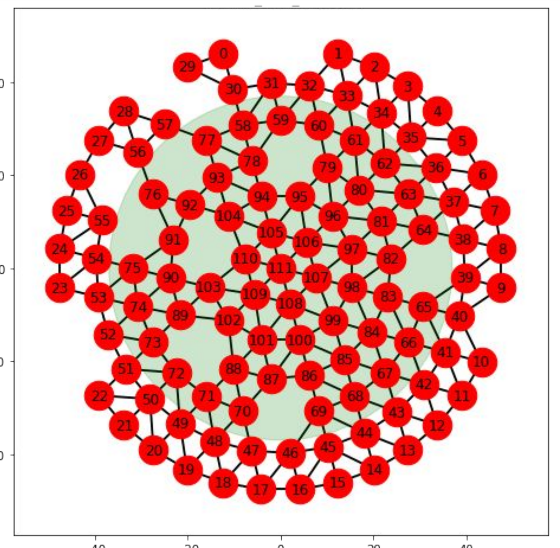


Fig: PMT sensor nodes viewed as a set of **Graph nodes**.

- Formulation and solution of an ***Inverse problem*** for the reconstruction of location and energy of an event.
- Sensor data contains a **"*graph structure*"** where sensors are nodes and data (luminosity) from them are graph signals, i.e., $\{\mathbf{x}_m\}_{m=1}^{M_0}$, where M_0 :number of nodes.
- ***Estimation of graph structure*** (Laplacian: $\mathbf{L}$) from observed signals via a linear program.

$$\min_{\mathbf{L}} \frac{1}{M_0}\sum_{m=1}^{M_0} \mathbf{x}_m^T \mathbf{L} \mathbf{x}_m \quad \text{s.t. } \mathbf{L1} = \mathbf{0},\ \text{trace}(\mathbf{L}) = n,\ (\mathbf{L})_{ij} = (\mathbf{L})_{ji} \le 0.$$

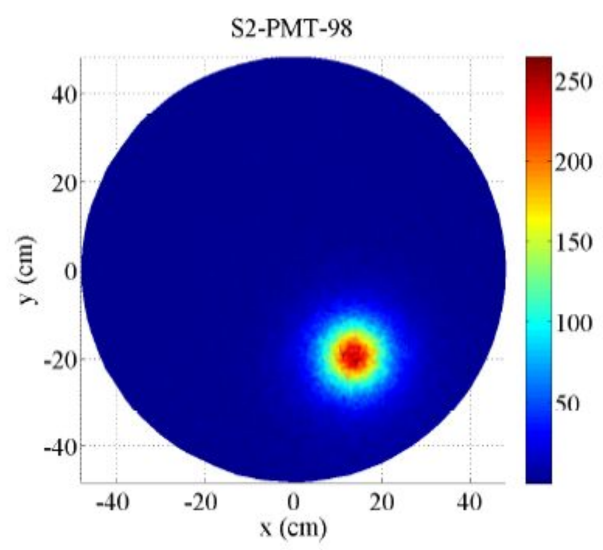


Fig: Spatial luminosity of one sensor (can be viewed as **Graph signal** of one node)

- ***Scalable approach*** for learning graphs: representation of large graphs as ***products of smaller factor graphs***.
- ***Computational*** and sample ***complexity*** gains over traditional graph learning approaches.

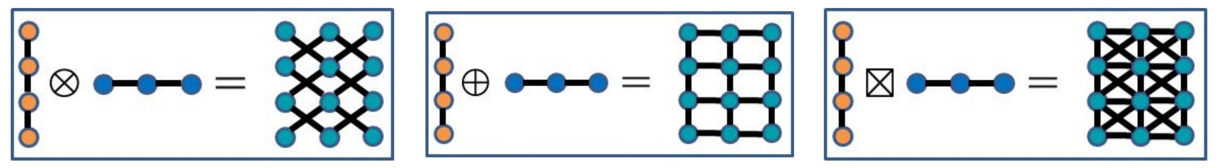
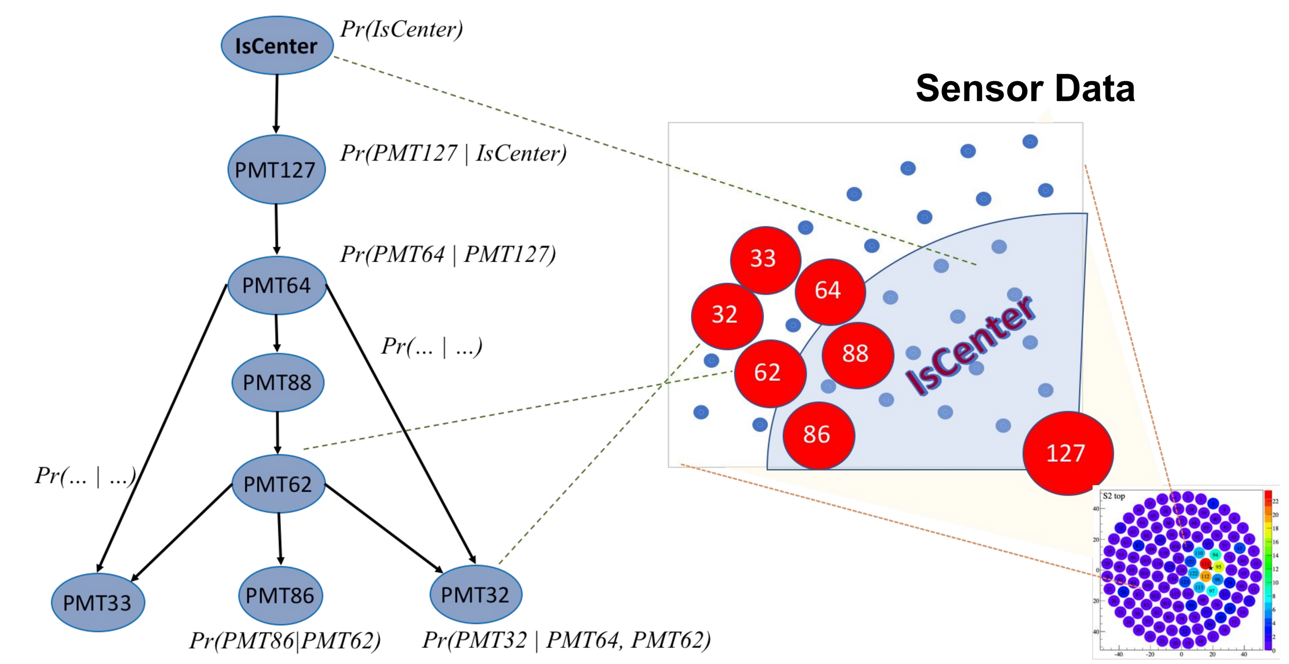


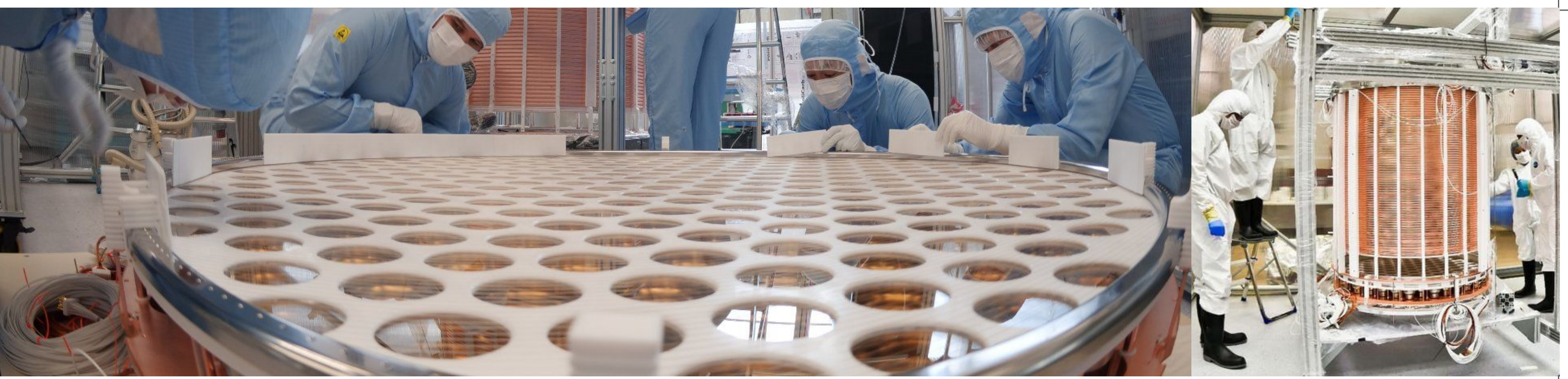
Fig: Graphs resulting from Kronecker (left), Cartesian (middle) and strong (right) **products** of two line graphs.

Example of a Simplified and Partial Learned Model



Learnt Graphical Model

- ***Training Data:*** Sensor (PMTs) readings over time. Currently, coarsely quantized at levels 1 / 2 / 3.
  +
  Interaction Location within the tank. Currently, coarsely quantized - *IsCenter* = 1 / 0.
- **Model Learning:**
  1. **Graph Structure Learning**: Currently, using BANJO, Greedy search + Bayesian-Dirichlet scoring.
  2. **Parameter Estimation:** *Pr[PMTx | Parents(PMTx)]*
- **Next Steps:** Refine granularity of readings and locations; Increase # of nodes & parameters



Literature produced by DIDACTS *so far*:

1. M. A. Lodhi, and W. U. Bajwa, "Learning Product Graphs Underlying Smooth Graph Signals". arXiv preprint. *arXiv:2002.11277* (2020).
2. DARWIN Collaboration: F. Agostini, ..., C. D. Tunnell, ..., K. Zuber. "Sensitivity of the DARWIN observatory to the neutrinoless double beta decay of 136Xe". Submit to EPJ C. arXiv:2003.13407 (2020)

More info about DIDACTS:

- Collaborate: Please reach out if you do physical-science measurement or domain-enhanced data science!
- Website: https://DIDACTS.org
- Open Science: https://github.com/DidactsOrg
- Photos: Credit Twitter:@XENONexperiment
- Award info: NSF awards 1940074, 1940209, and 1940080.