Dark Matter Direct Detection Experiments & Machine Learning
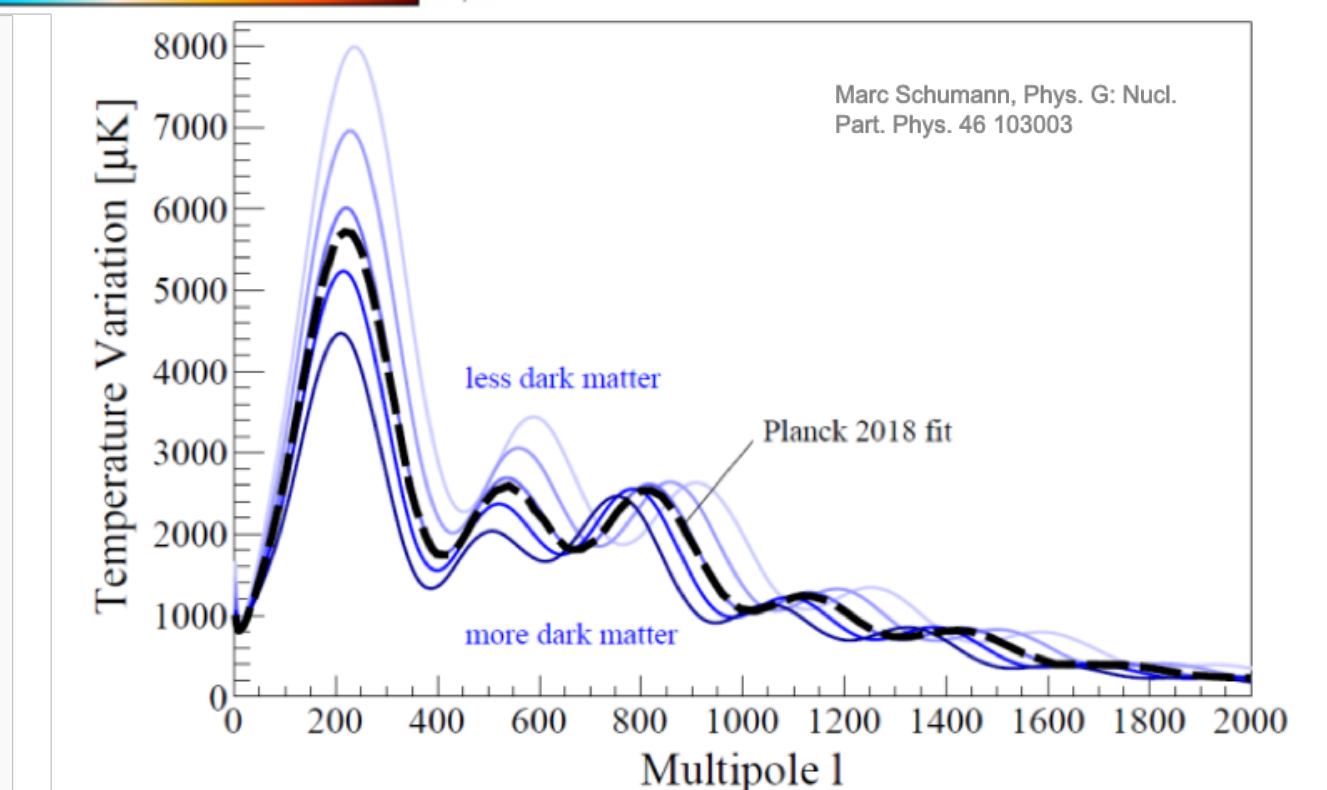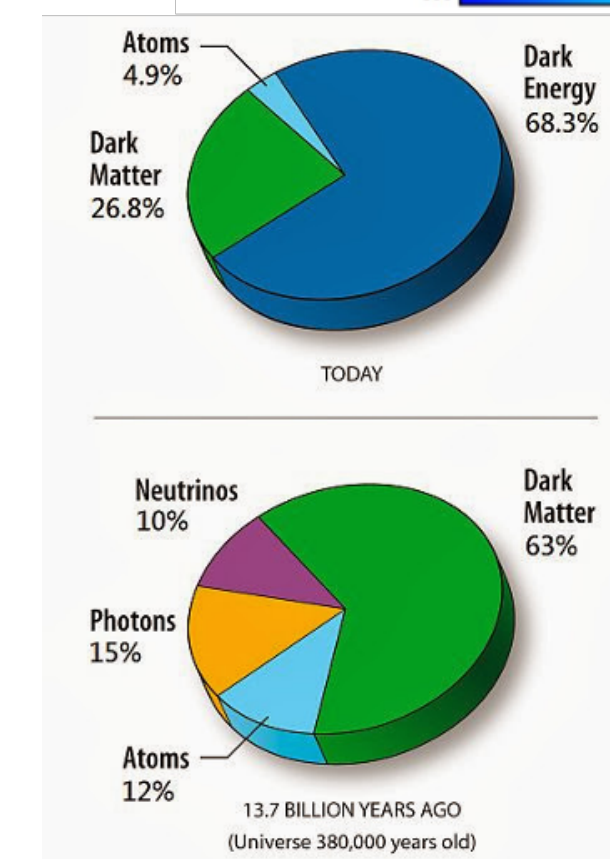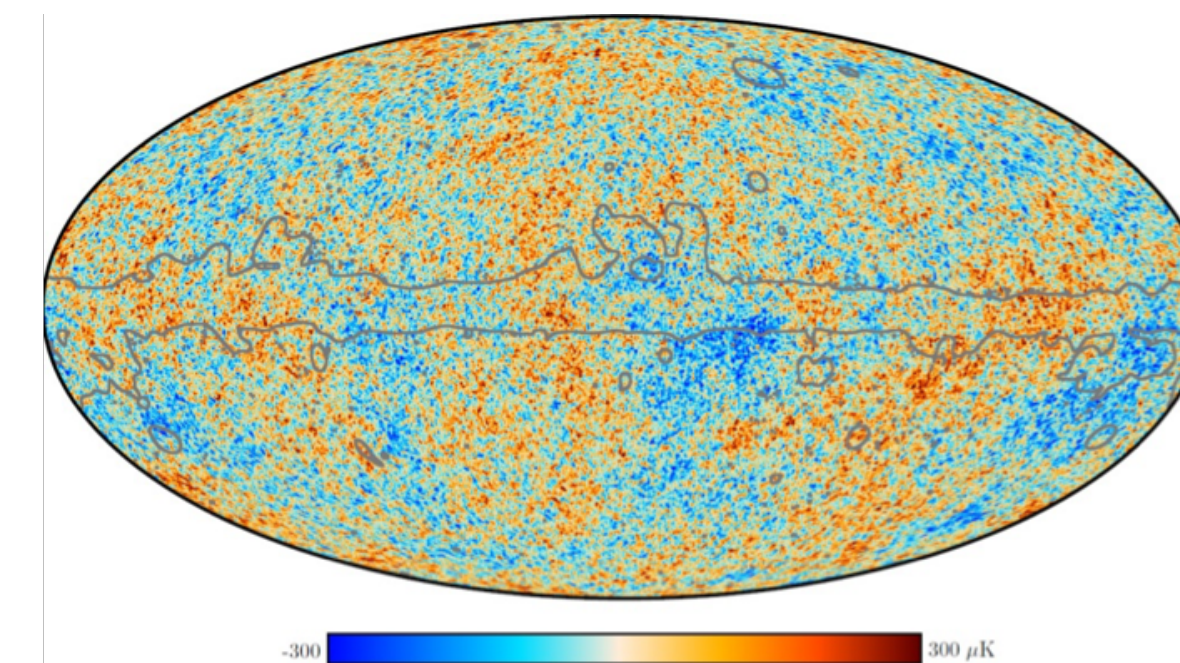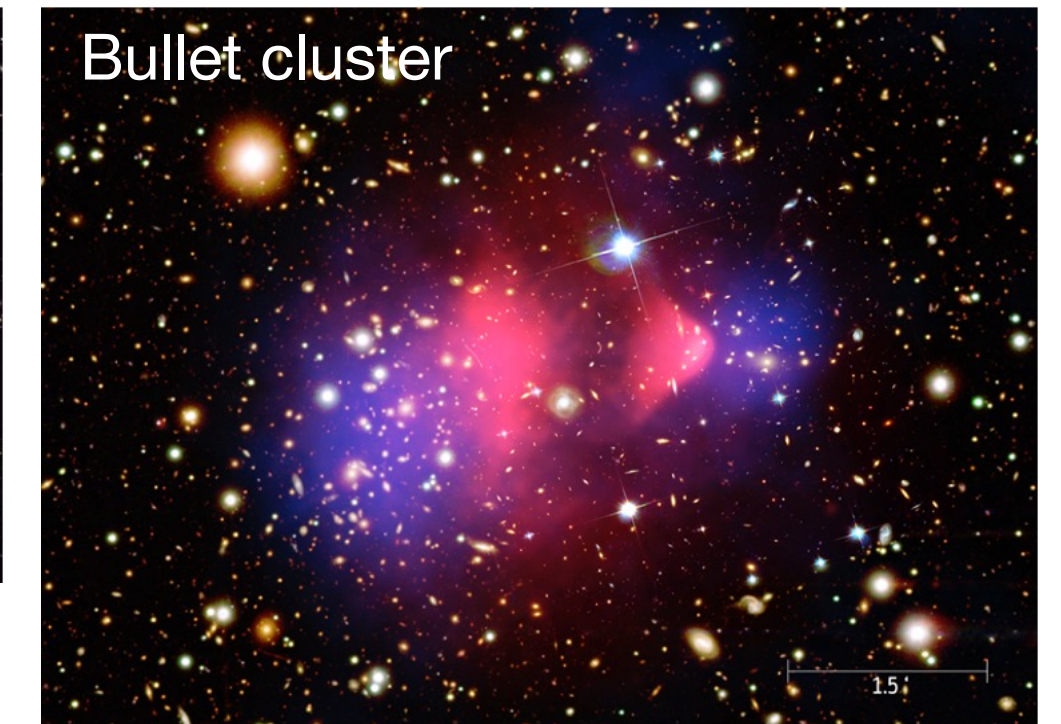
Aaron Higuera

RICE UNIVERSITY

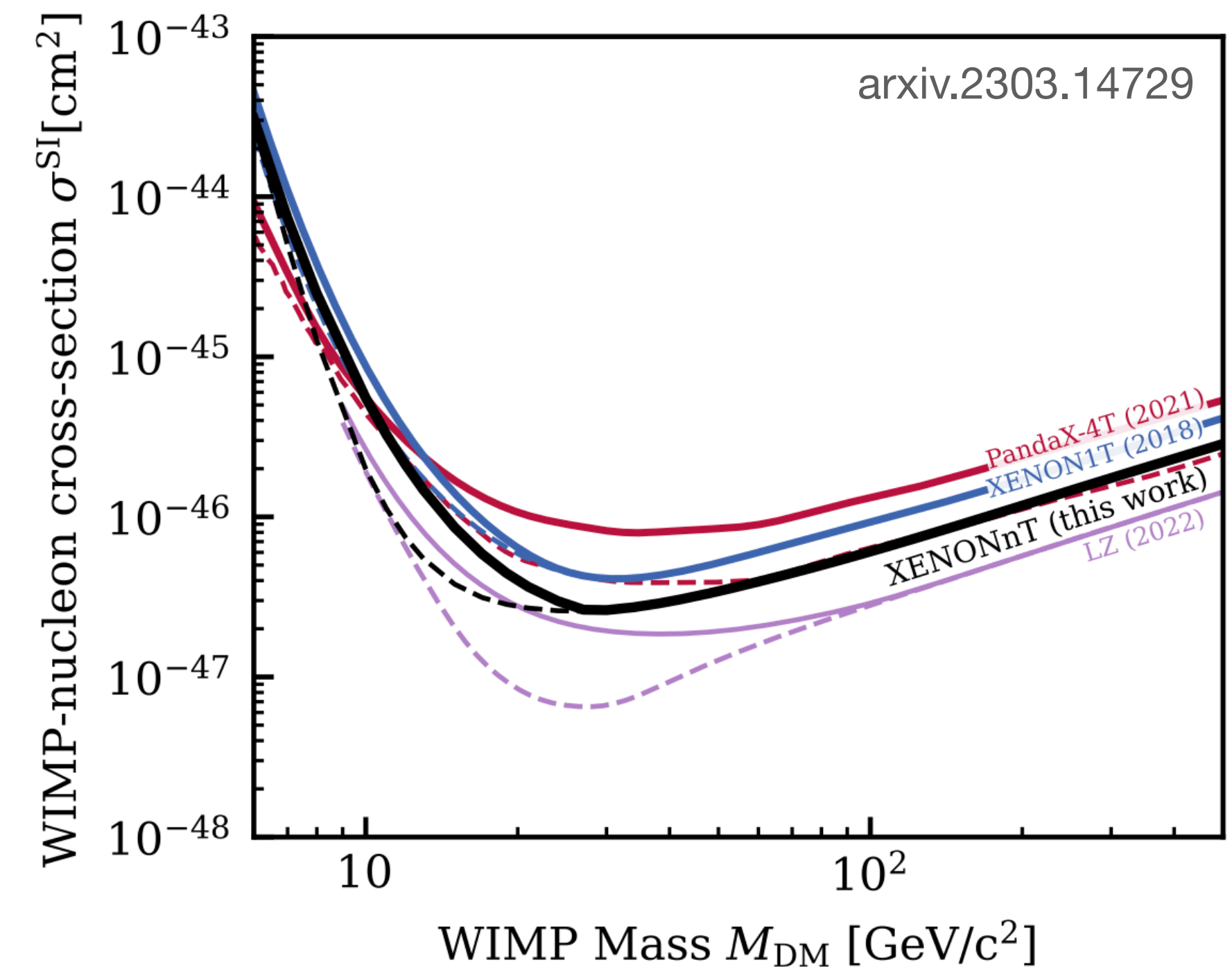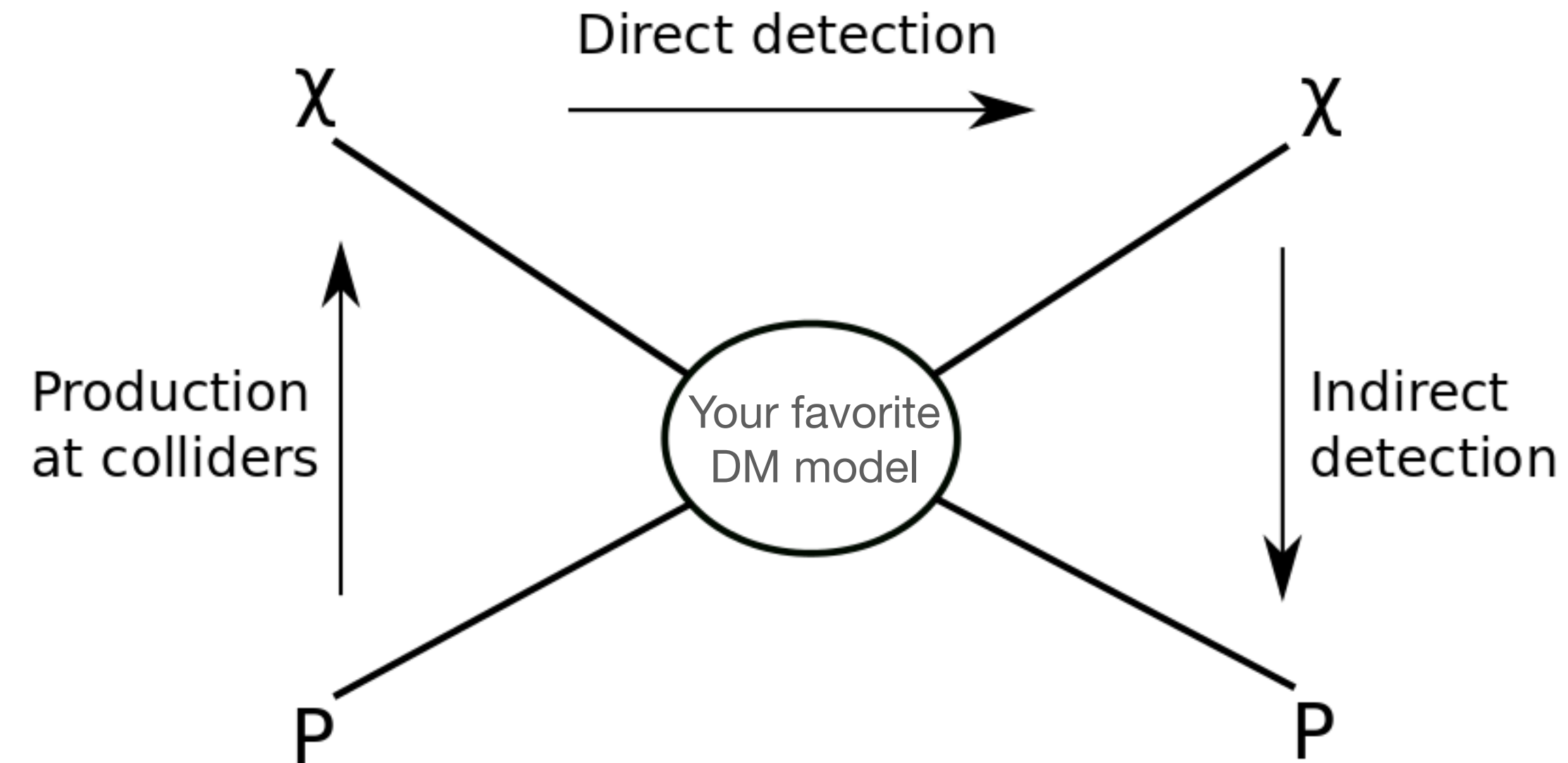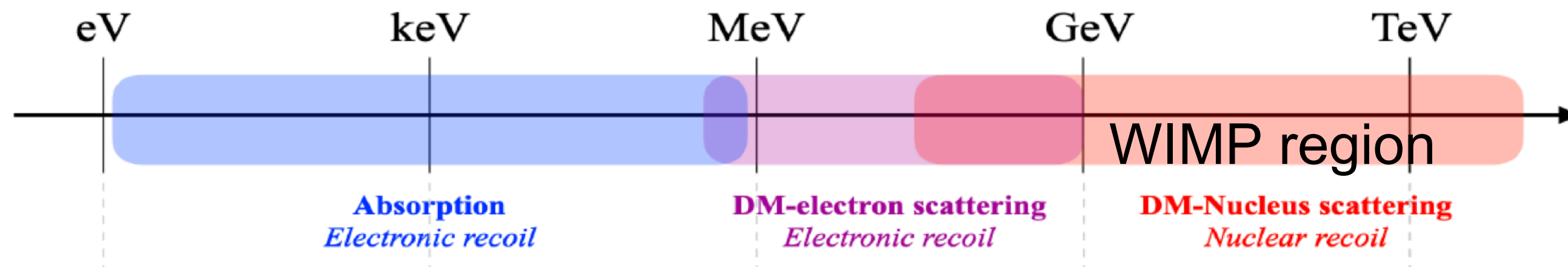**MODE workshop, Princeton University, July 2023**

Dark Matter

- Decades of astronomical evidence
  - From galaxies and clusters
    - Rotation curves
    - Cluster movement, collisions and gravitational lensing
  - Cosmic scale
    - Structure of the universe
    - Cosmic microwave background

- What do we know about dark matter?
  - Electrically neutral
  - Weakly interacting (weaker than the SM weak interaction)
  - Either long-lived/stabled or produced
  - Should be among us

A. Higuera, MODE workshop, Princeton University, July 2023

Coma cluster

Bullet cluster

Atoms 4.9%
Dark Matter 26.8%
Dark Energy 68.3%
TODAY

Neutrinos 10%
Photons 15%
Atoms 12%
Dark Matter 63%
13.7 BILLION YEARS AGO (Universe 380,000 years old)

Marc Schumann, Phys. G: Nucl. Part. Phys. 46 103003

less dark matter
Planck 2018 fit
more dark matter

Temperature Variation [μK]

Multipole l

Dark Matter

- How do we search for dark matter?
  - In direct detection
    - Dark matter particle-nucleus scattering
    - Ultra-low background experiments
    - Low thresholds

WIMP region

arxiv.2303.14729

A. Higuera, MODE workshop, Princeton University, July 2023
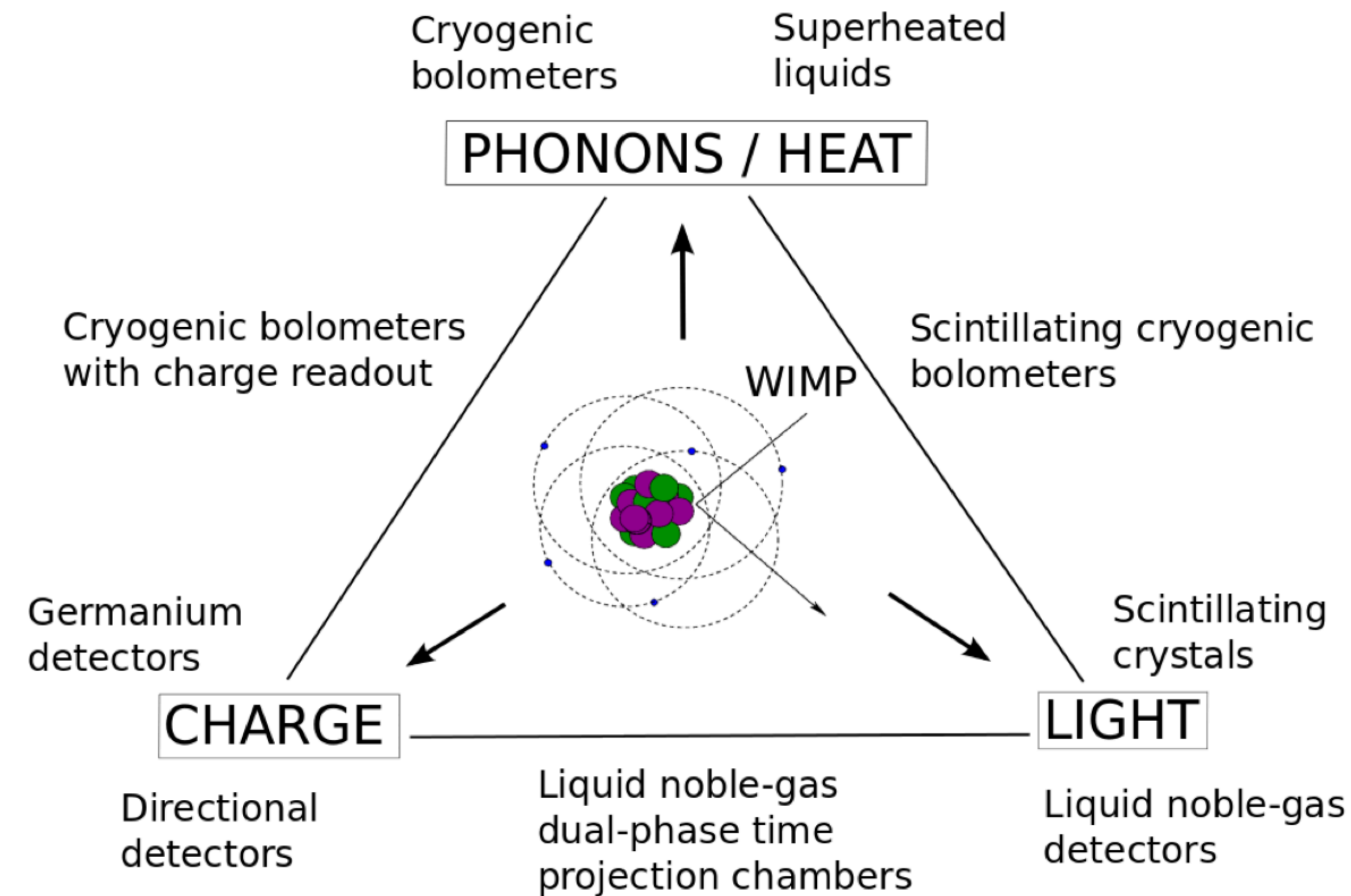
Dark Matter

- How do we search for dark matter?
  - In direct detection
    - Dark matter particle-nucleus scattering
    - Ultra-low background experiments
    - Low thresholds
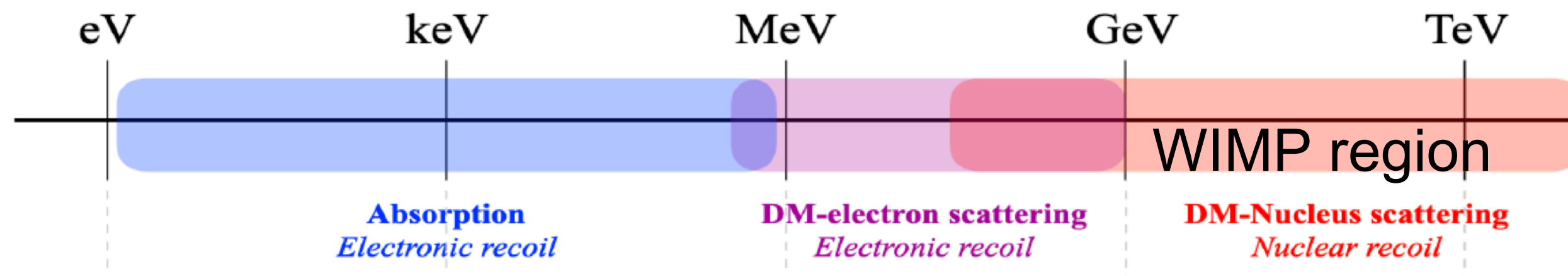

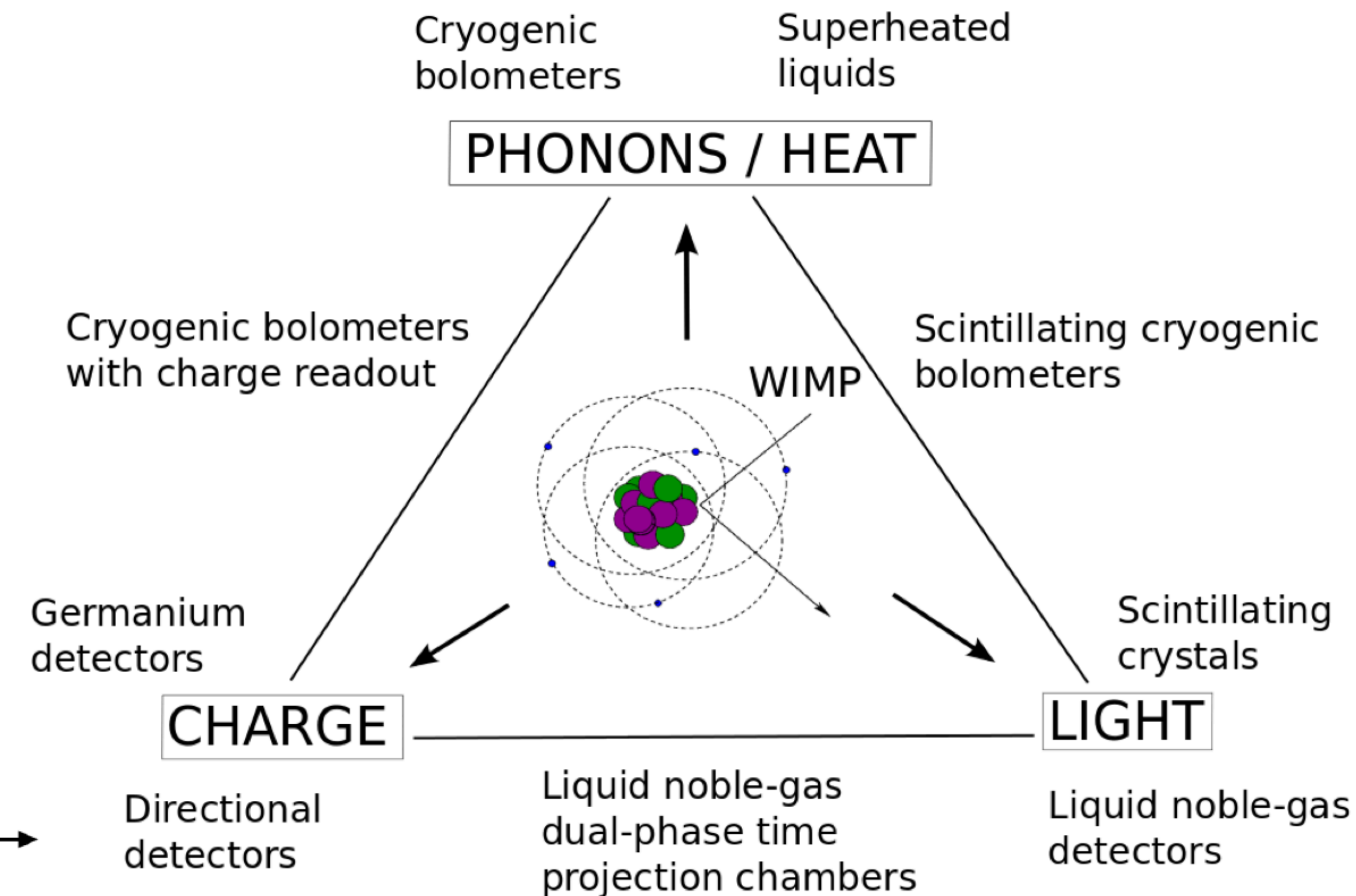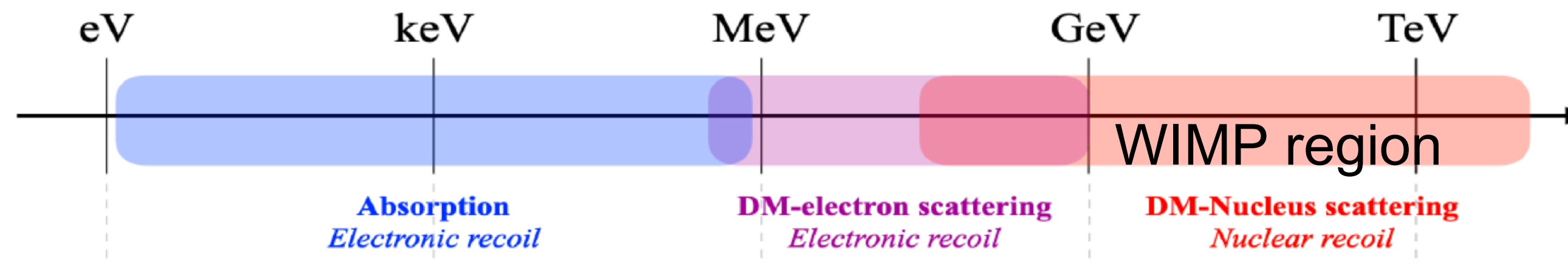WIMP region

Dark Matter

- How do we search for dark matter?
  - In direct detection
    - Dark matter particle-nucleus scattering
    - Ultra-low background experiments
    - Low thresholds

WIMP region

Absorption
Electronic recoil

DM-electron scattering
Electronic recoil

DM-Nucleus scattering
Nuclear recoil

Cryogenic bolometers

Superheated liquids

PHONONS / HEAT

Cryogenic bolometers with charge readout

Scintillating cryogenic bolometers

WIMP

Germanium detectors

CHARGE

LIGHT

Directional detectors

Liquid noble-gas dual-phase time projection chambers

Scintillating crystals

Liquid noble-gas detectors

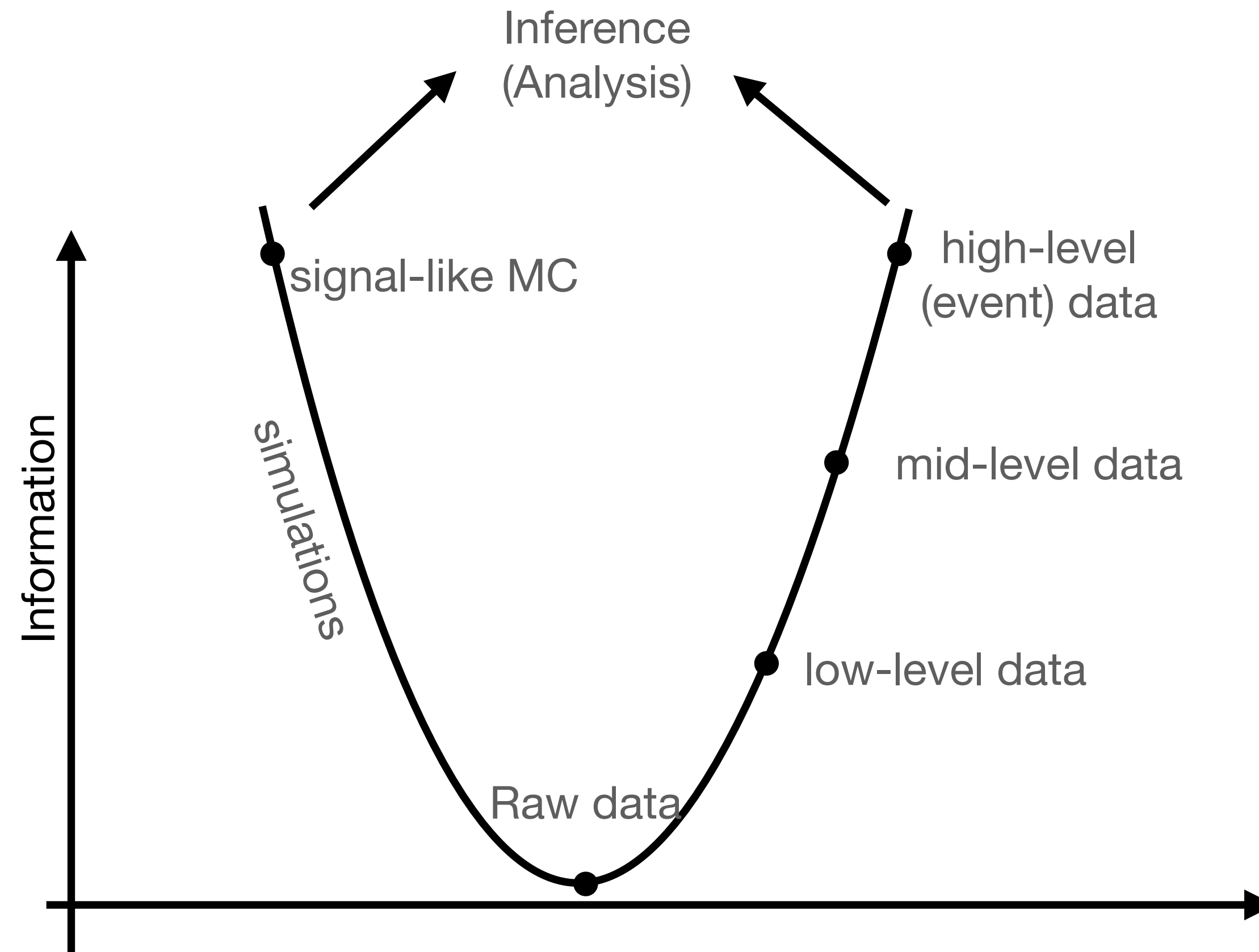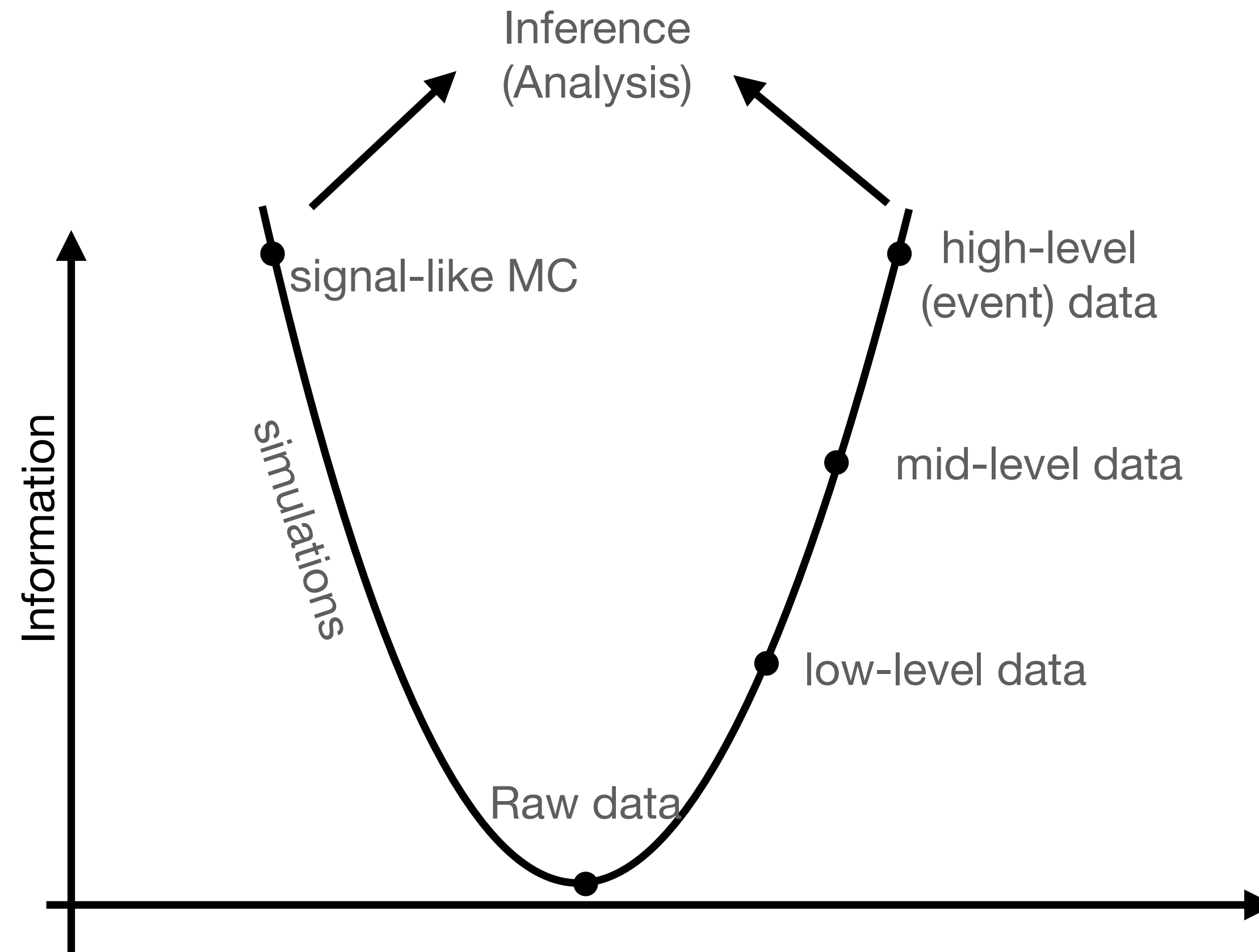**This talk will discuss liquid-noble TPC detectors**

What are the challenges in DM direct detection?
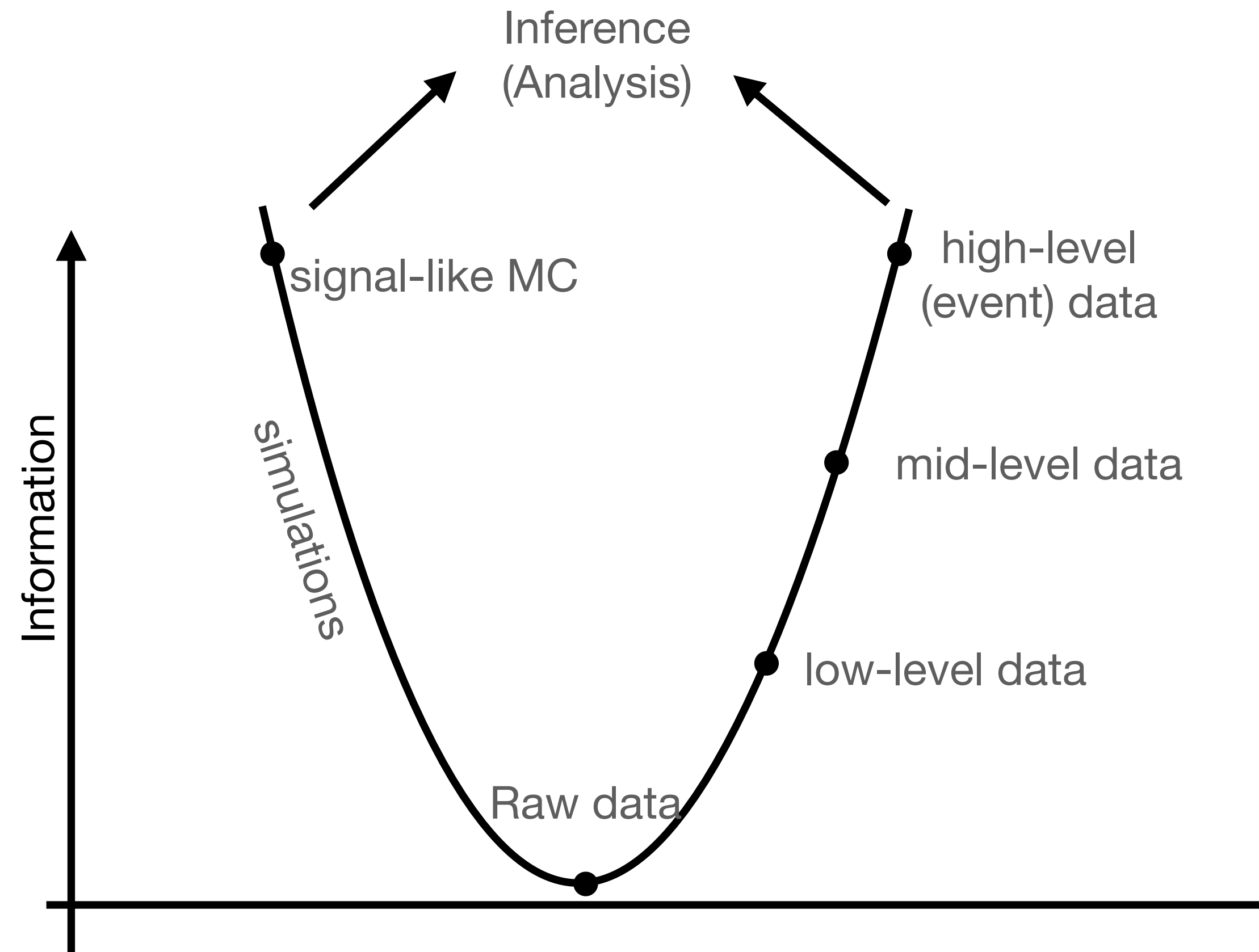
"The standard HEP workflow"
- Forward model and solving the inverse problem
- Detailed knowledge of the physical processes are encoded in the forward model

A. Higuera, MODE workshop, Princeton University, July 2023

What are the challenges in DM direct detection?

Information

simulations

Inference (Analysis)

signal-like MC

high-level (event) data

mid-level data

low-level data

Raw data

"The standard HEP workflow"
- Forward model and solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~
  - Model dependency
  - Pushing the limits, low-thresholds
    - Hardware
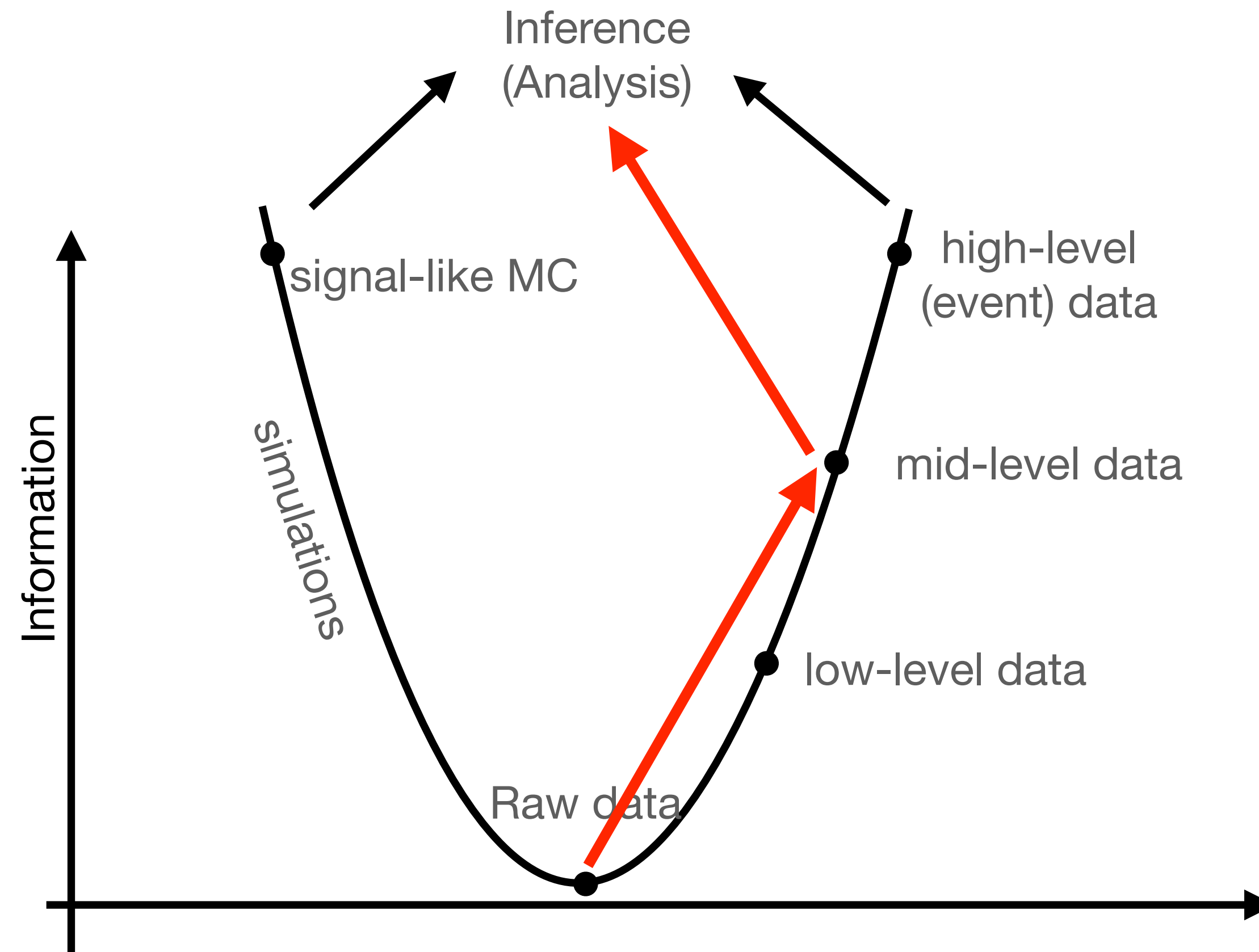   - Software (high-rates ~50 MB/s in search mode and up to 500 MB/s in calibration)

What are the challenges in DM direct detection?

"~~The standard HEP workflow~~"

- ~~Forward modeling and~~ solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~
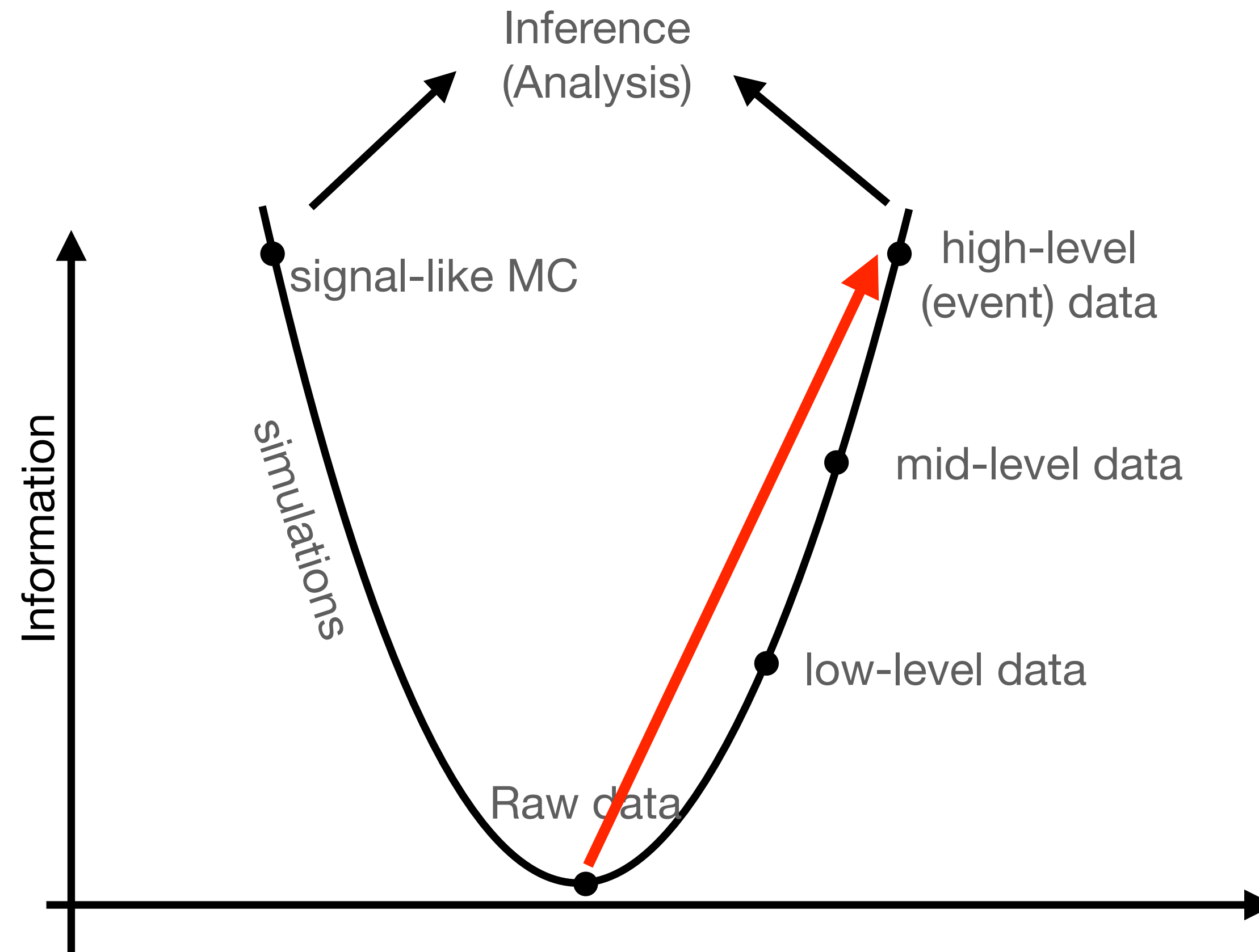- **How can we optimize analysis such that we can improve the science?**

What are the challenges in DM direct detection?

"~~The standard HEP workflow~~"

- ~~Forward modeling and~~ solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~

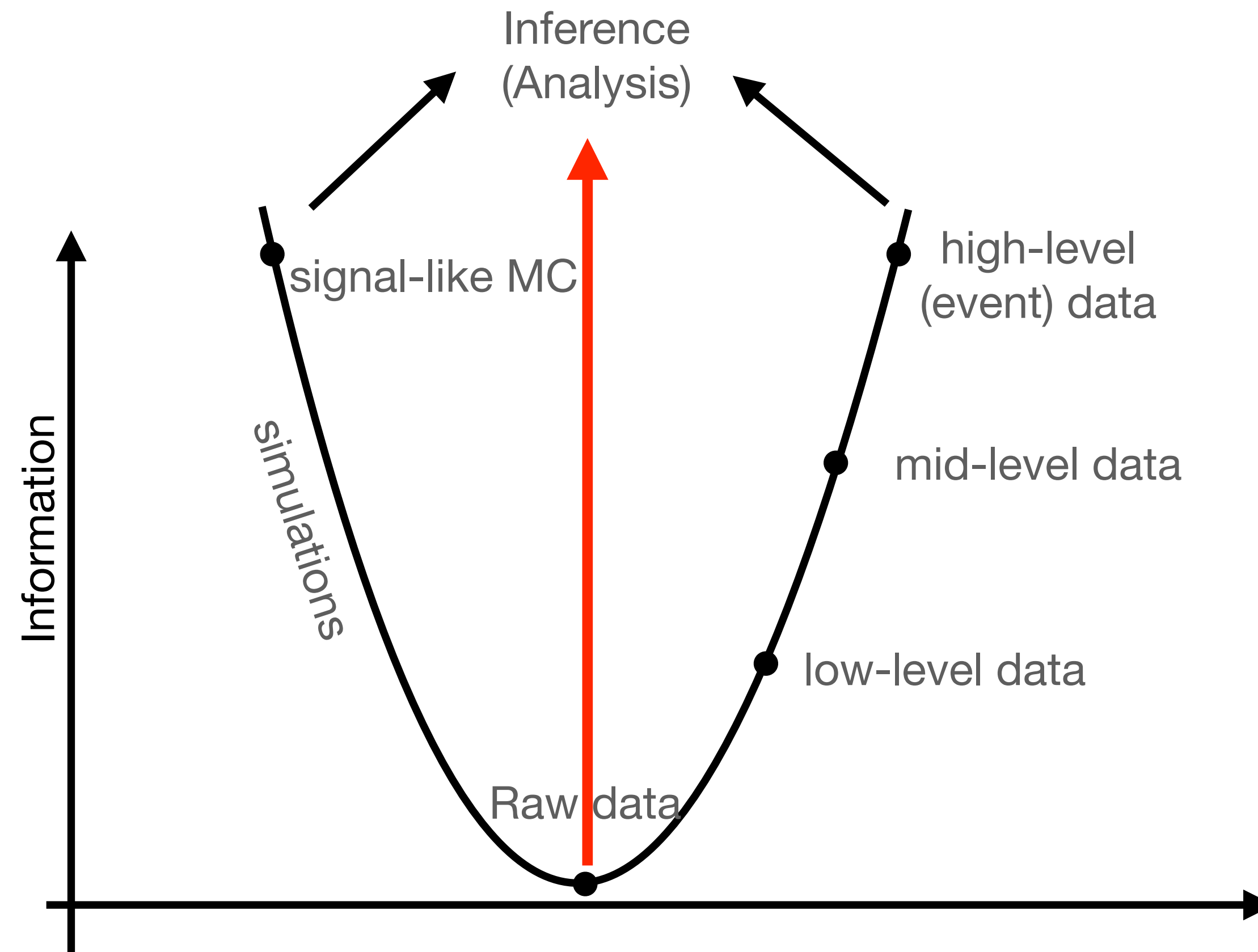- **How can we improve "reconstruction" to advance the science?**

A. Higuera, MODE workshop, Princeton University, July 2023

What are the challenges in DM direct detection?

Inference
(Analysis)

signal-like MC

high-level
(event) data

simulations

Information

mid-level data

low-level data

Raw data

"~~The standard HEP workflow~~"

- ~~Forward modeling and~~ solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~

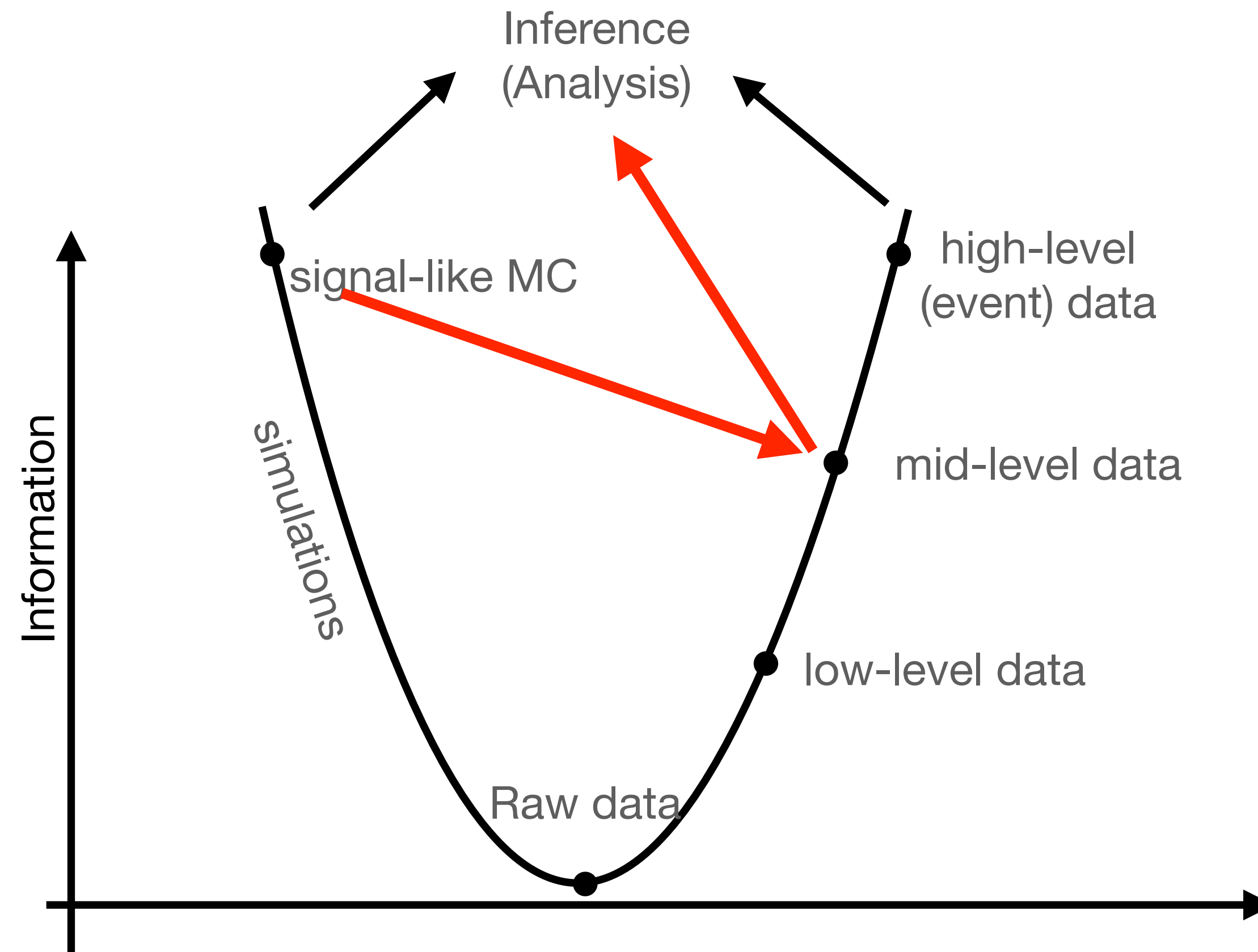- **How can we improve "reconstruction" to advance the science?**

What are the challenges in DM direct detection?

"~~The standard HEP workflow~~"
- ~~Forward modeling and~~ solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~
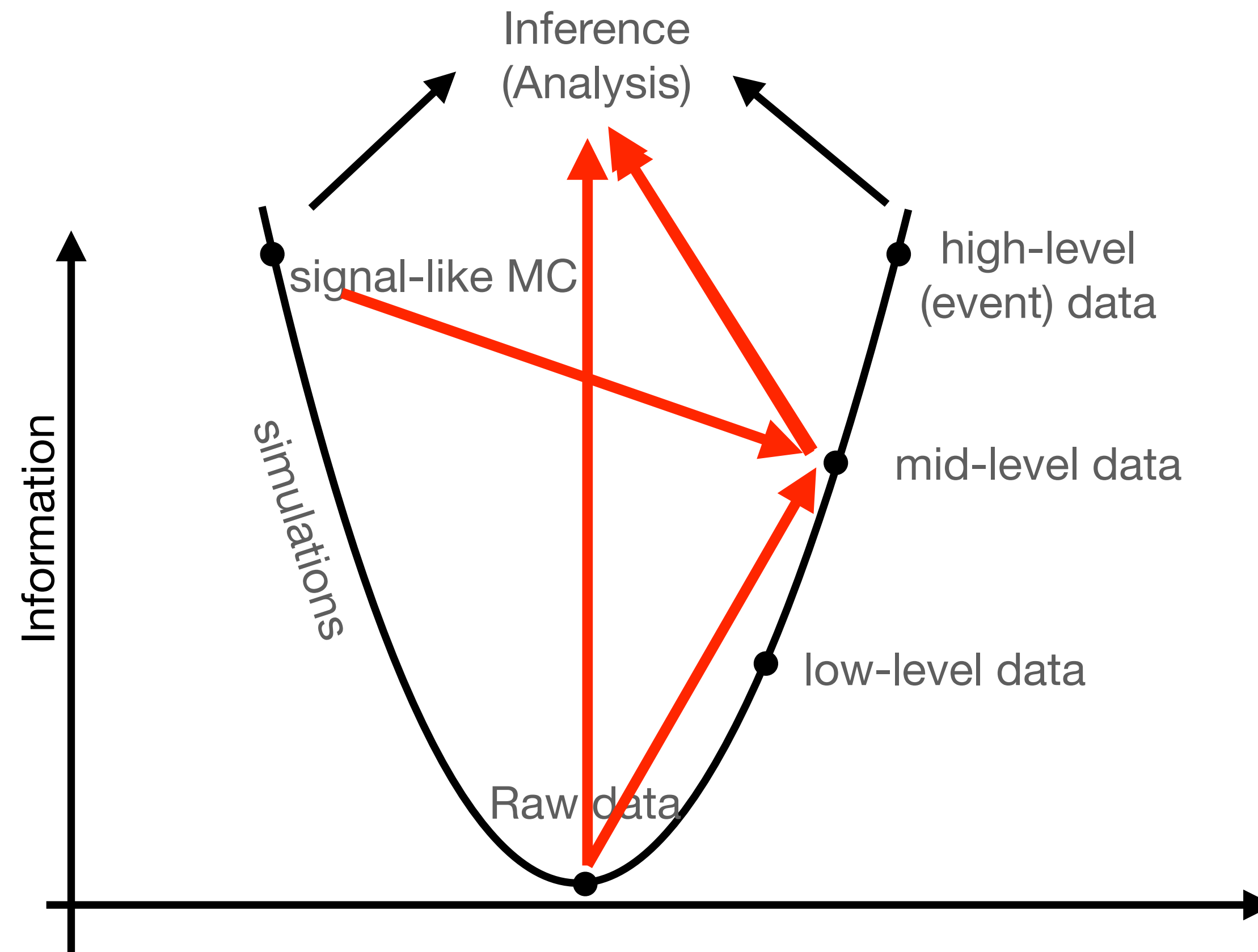- **How can we improve "reconstruction" to advance the science?**

What are the challenges in DM direct detection?

"~~The standard HEP workflow~~"
- ~~Forward modeling and~~ solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~
- **How can we improve "reconstruction" to advance the science?**

What are the challenges in DM direct detection?

"The standard HEP workflow"
- ~~Forward modeling and~~ solving the inverse problem
- ~~Detailed knowledge of the physical processes are encoded in the forward model~~
- **How can we improve "reconstruction" to advance the science?**
  - Problems
    - **Regression (energy, position)**
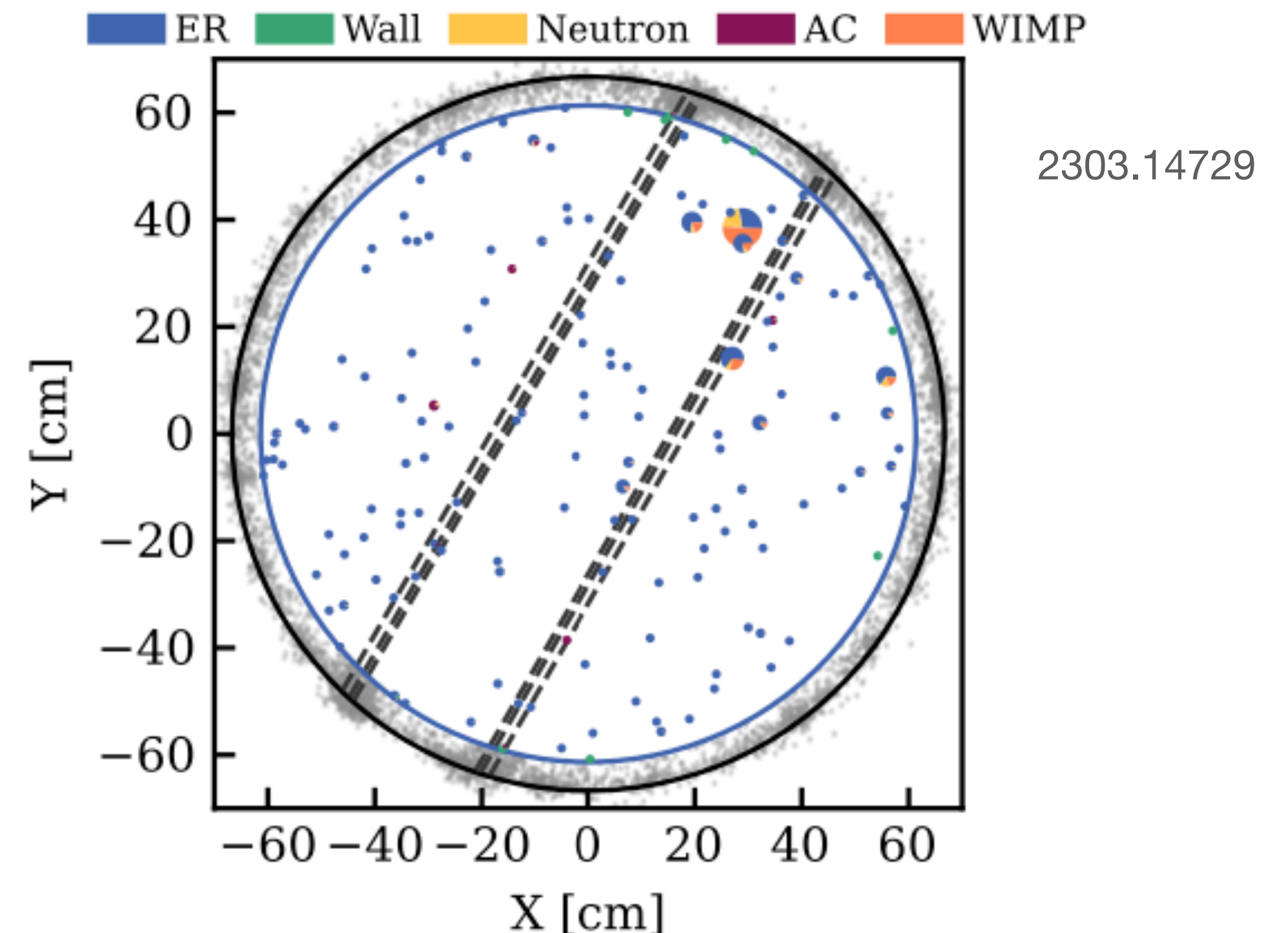   - **Classification (PID, anomaly detection)**
    - **Inference**

A. Higuera, MODE workshop, Princeton University, July 2023
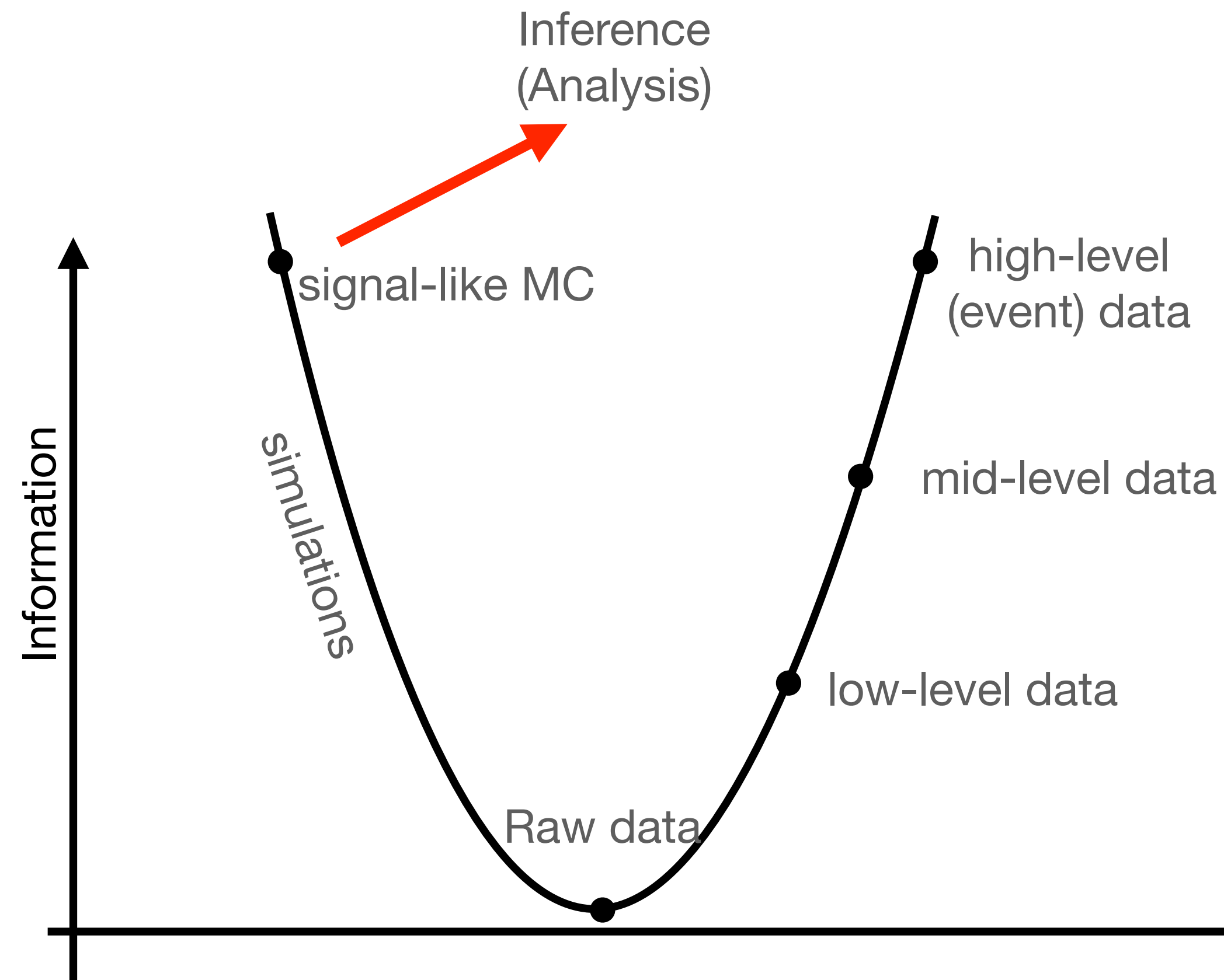
Regression in DM experiments
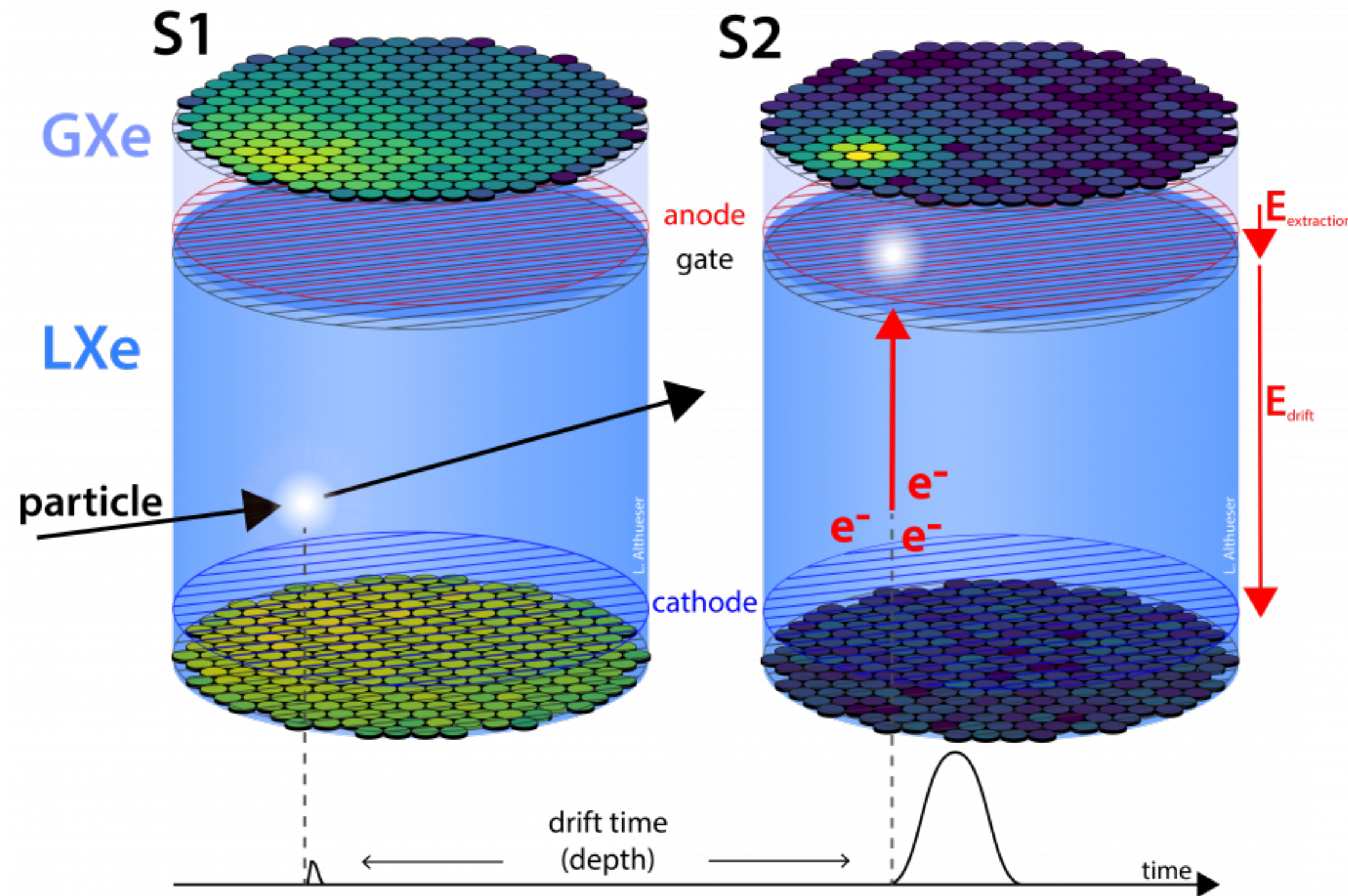
$$E = W \left(\frac{cS_1}{g_1} + \frac{cS_2}{g_2} \right)$$

Interaction vertex

- **2D position given by hit pattern (bottom array)**
- **XENONnT, LZ, Panda-X (dual-phase LXe-TPC)**
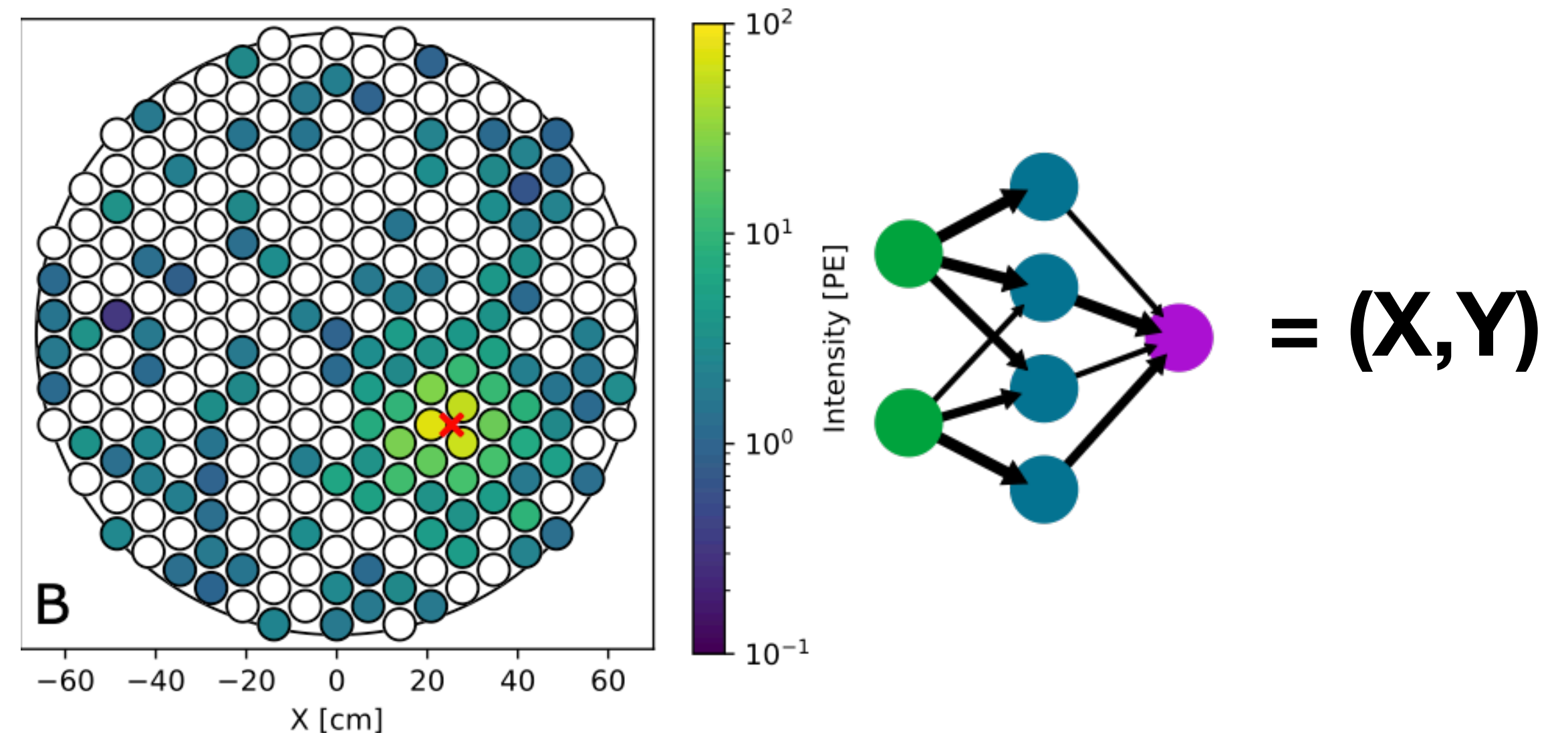- **Darkside -20k (dual-phase LAr-TPC)**

2303.14729

A. Higuera, MODE workshop, Princeton University, July 2023

Regression in DM experiments

Inference
(Analysis)

Information

simulations

signal-like MC

high-level
(event) data

mid-level data

low-level data

Raw data

Interaction vertex
- **2D position given by hit pattern (bottom array)**
- **Train on simulation (MC)**

= (X,Y)

Classic approach: MLP
GCN *Front. Artif. Intell.* 5 (2022) 832909
arXiv:2112.07995

Regression in DM experiments

Inference
(Analysis)

signal-like MC

high-level
(event) data

simulations

mid-level data

low-level data

Raw data

Information
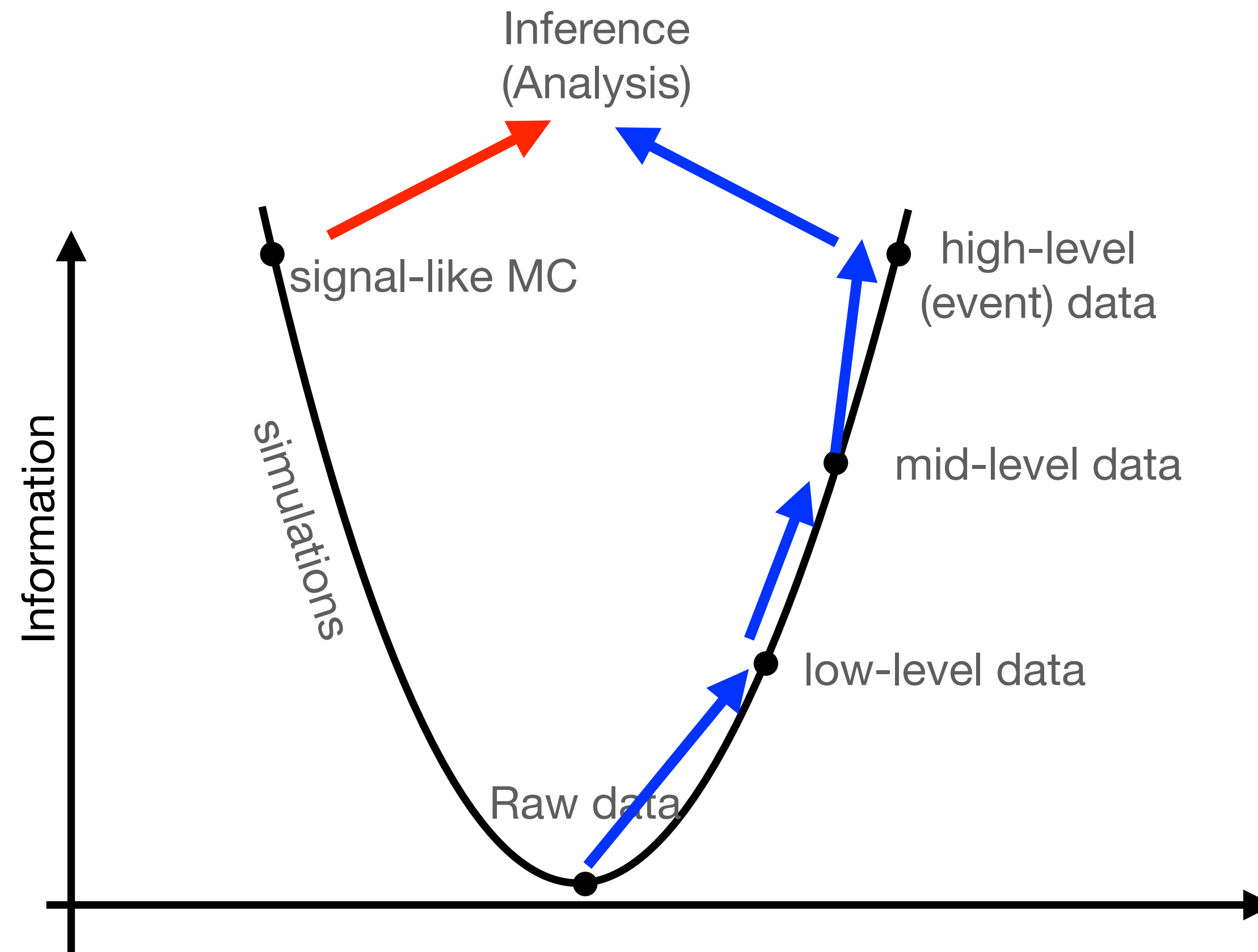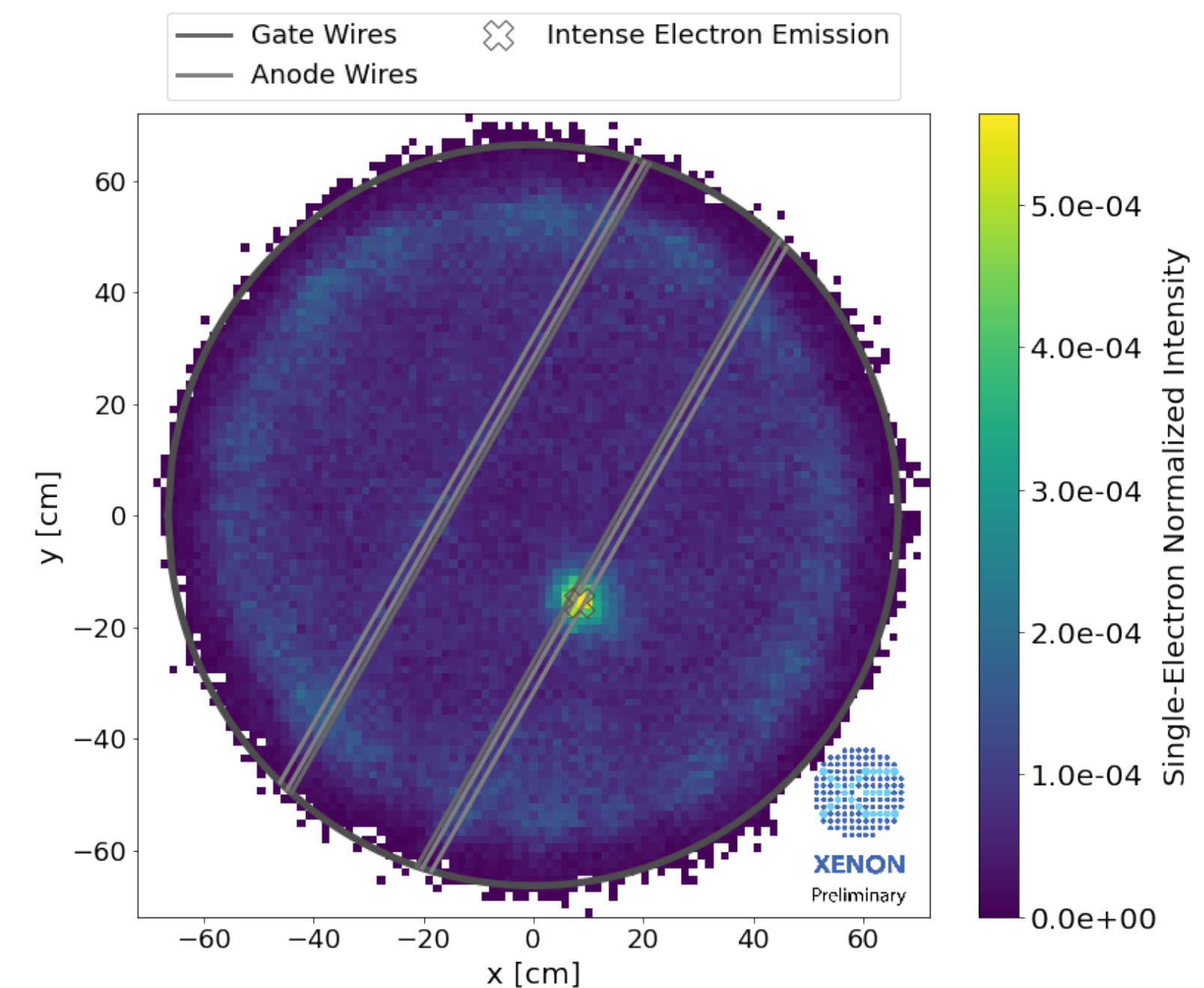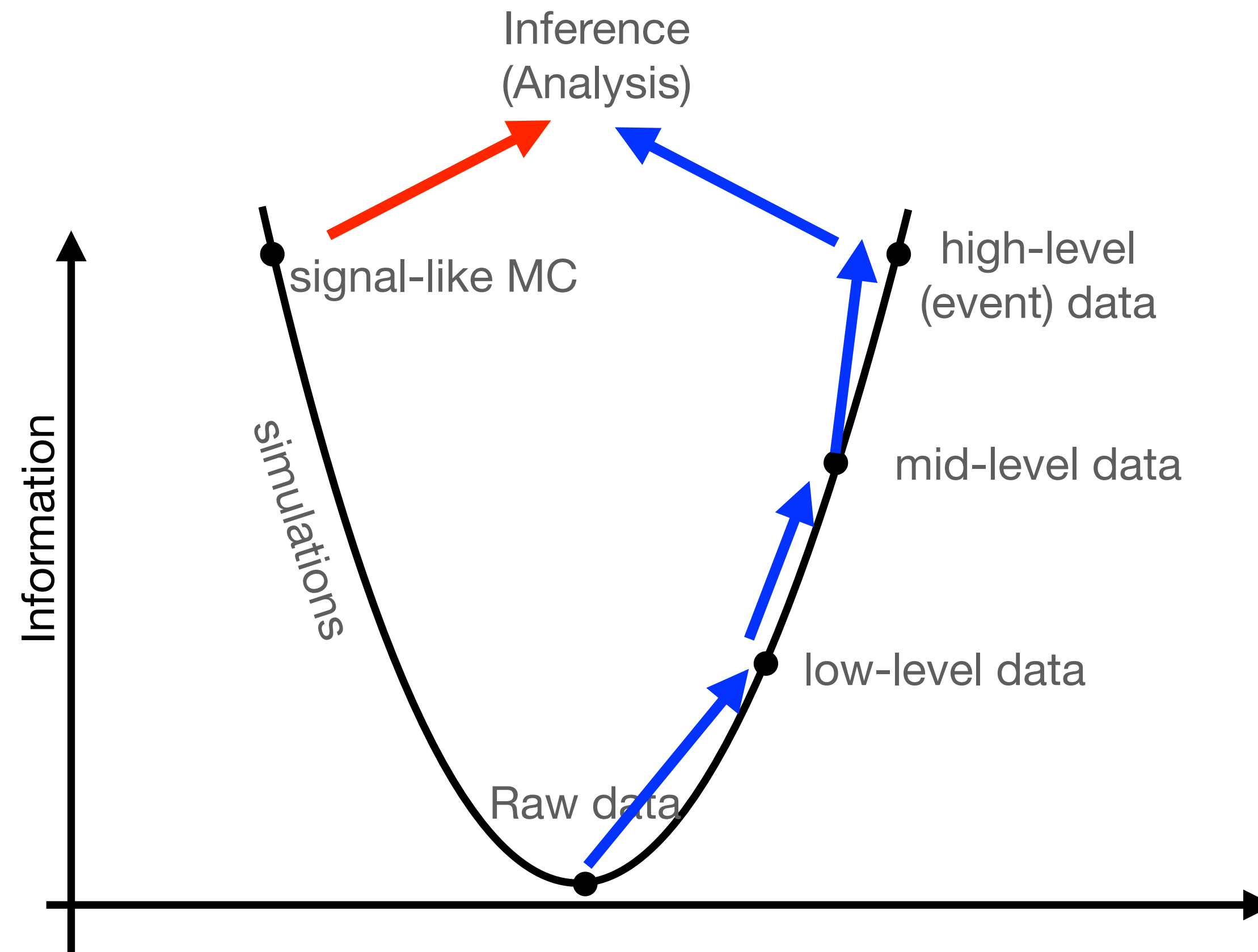
A. Higuera, MODE workshop, Princeton University, July 2023

Interaction vertex

- **2D position given by hit pattern (bottom array)**
- **Train on simulation (MC)**
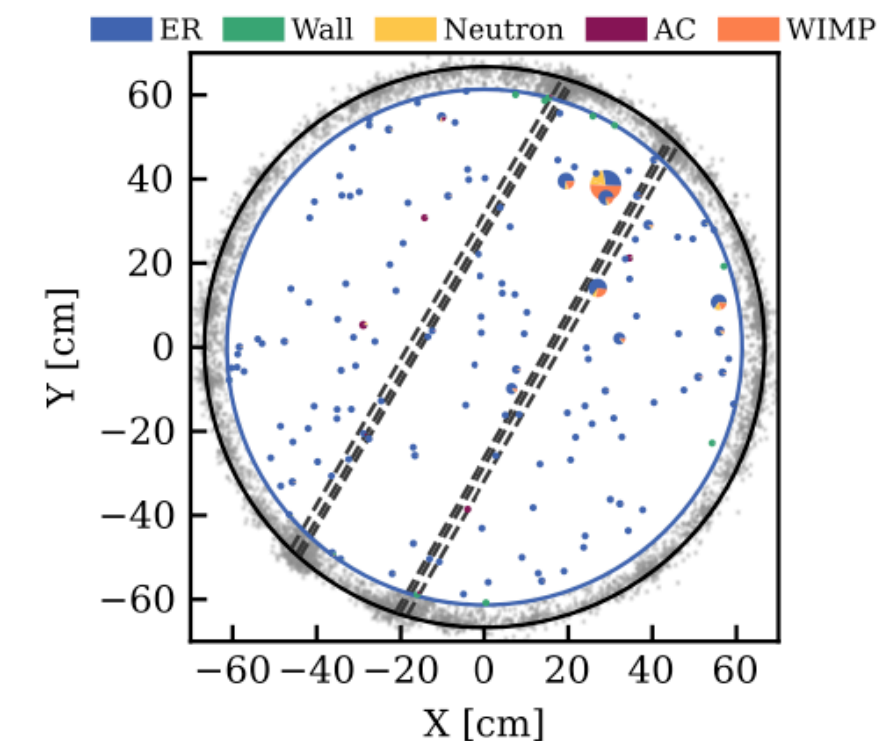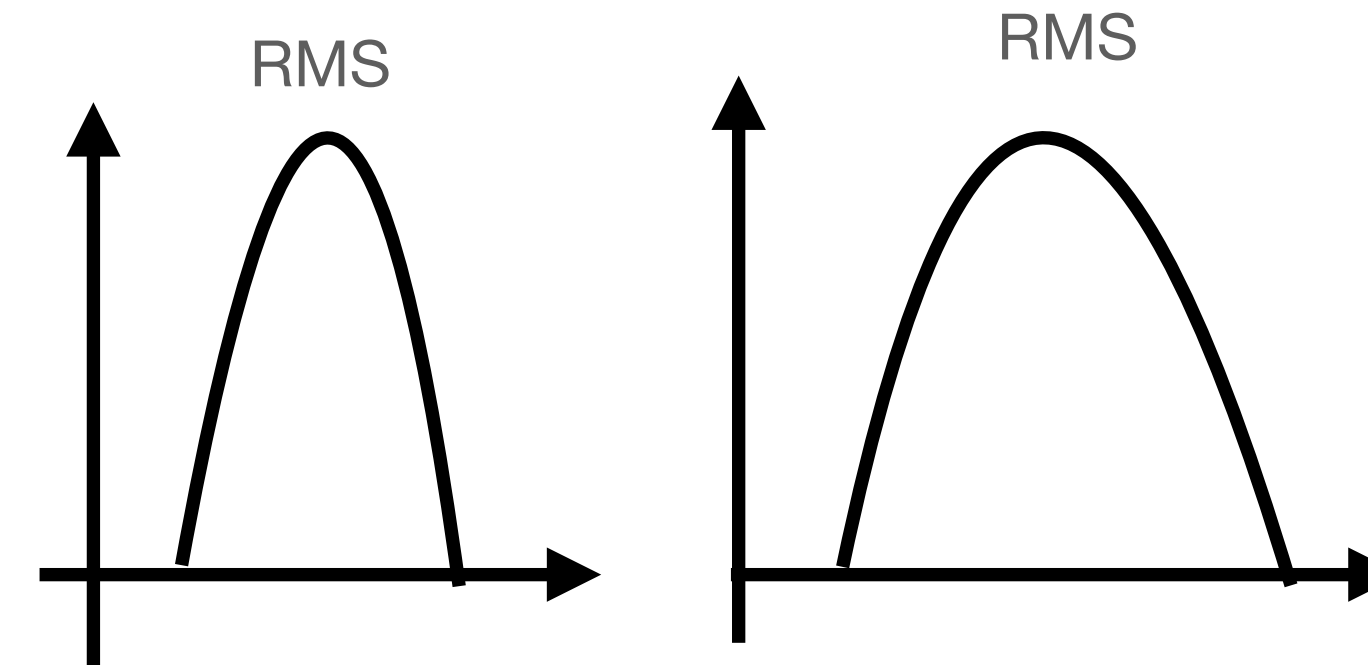- ~~Detailed knowledge of the physical processes are encoded in the forward model~~

Regression in DM experiments

Inference
(Analysis)

signal-like MC

high-level
(event) data

Information

simulations

mid-level data

low-level data

Raw data

Interaction vertex
- **2D position given by hit pattern**
- **What is the uncertainty on the inference?**

RMS

RMS

ER   Wall   Neutron   AC   WIMP

$$x \pm \sigma_x$$

$$y \pm \sigma_y$$

A. Higuera, MODE workshop, Princeton University, July 2023

Regression in DM experiments

Interaction vertex
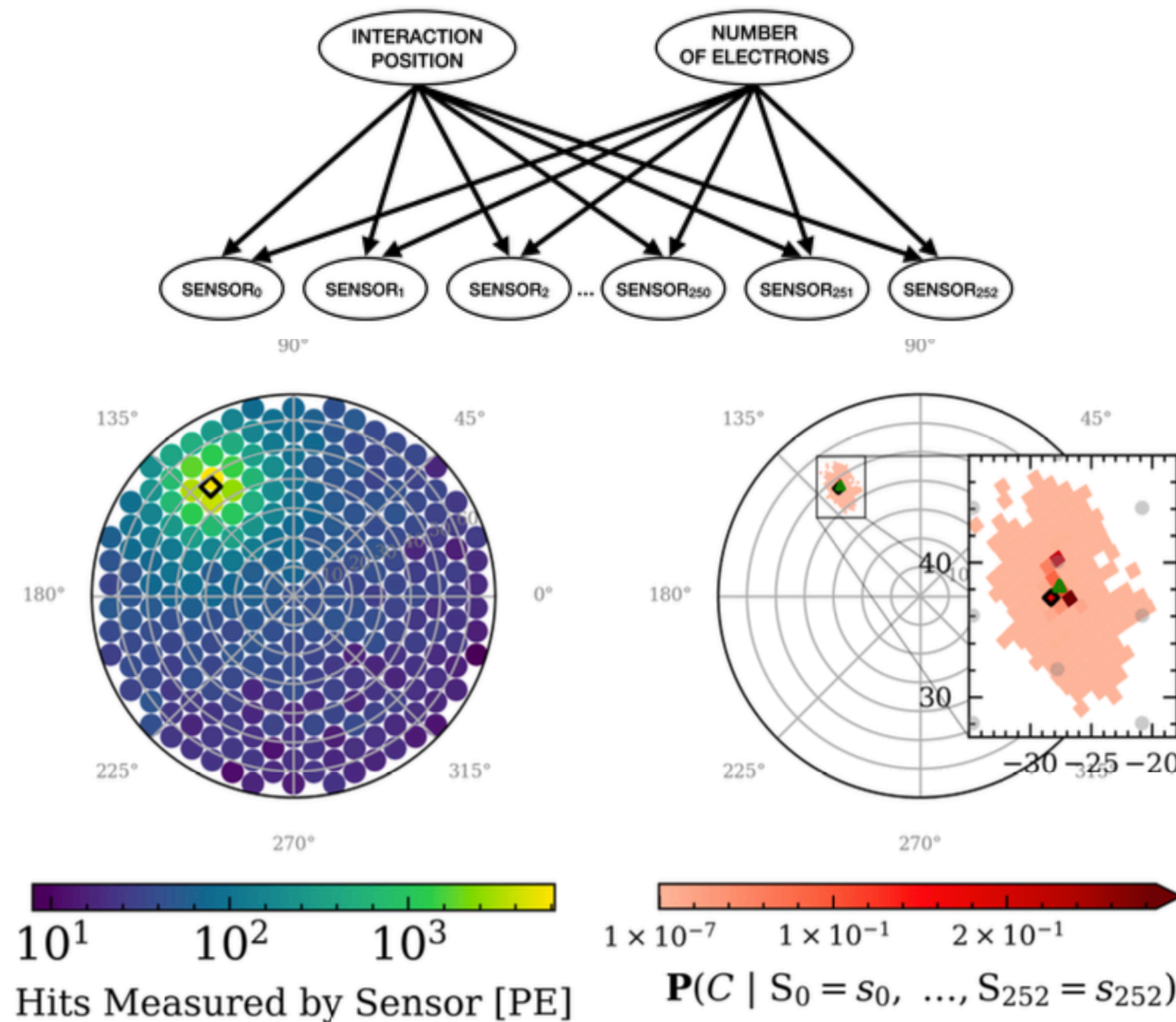
- **2D position given by hit pattern**
- **What is the uncertainty on the inference?**

**Graphical Models are All You Need: Per-interaction reconstruction uncertainties in a dark matter detection experiment**

NeurIPS 2022 by C. Peters arxiv:2205.10305

- **Provides per-event uncertainty on the position**

$$\mathbf{P}(C \mid S_0 = s_0, \; ..., \; S_{252} = s_{252})$$

A. Higuera, MODE workshop, Princeton University, July 2023

Regression in DM experiments

Hits Measured by Sensor [PE]

$\mathbf{P}(C \mid S_0 = s_0, \ldots, S_{252} = s_{252})$

Graphical Models are All You Need: Per-interaction reconstruction uncertainties in a dark matter detection experiment

A. Higuera, MODE workshop, Princeton University, July 2023
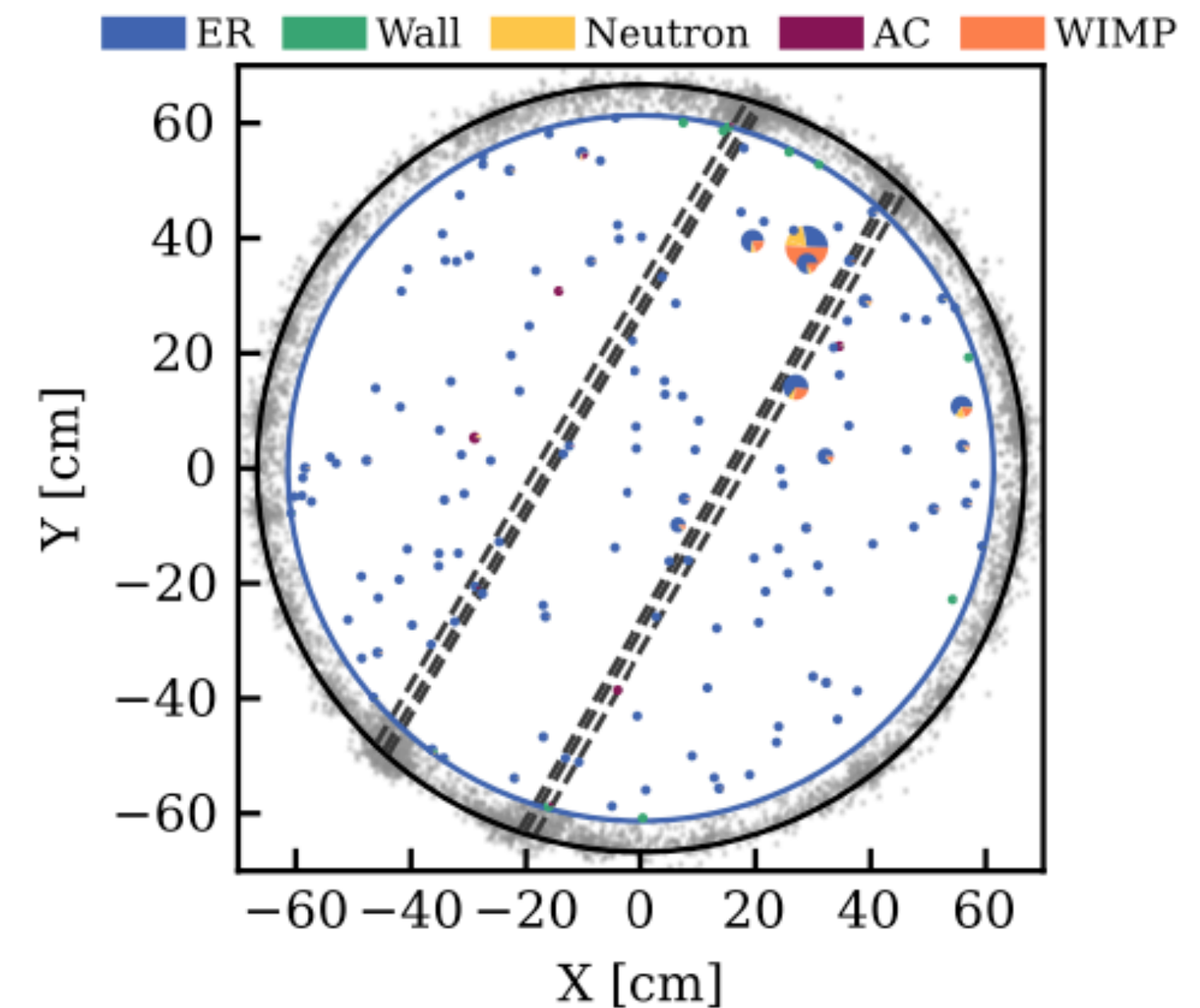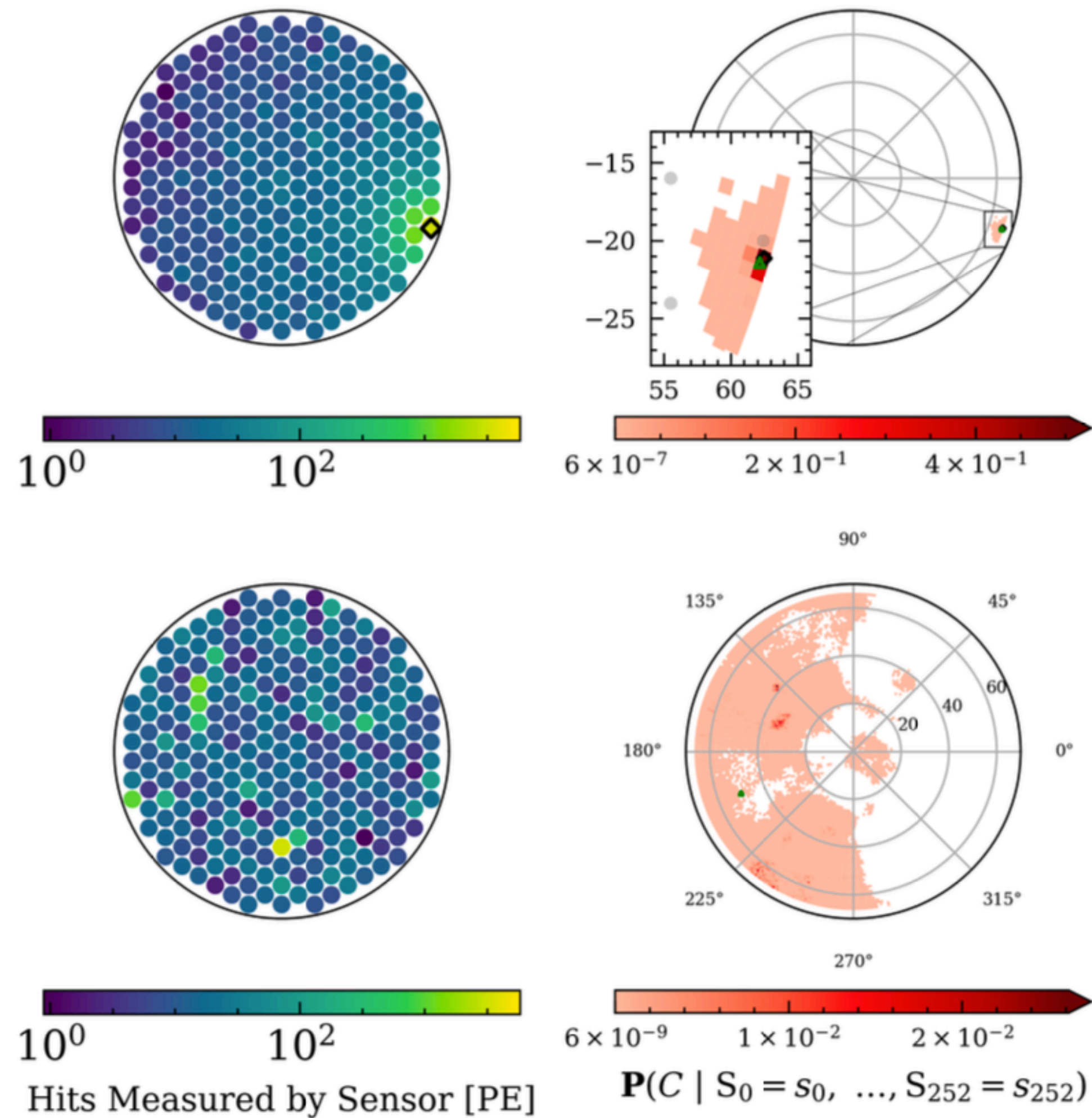
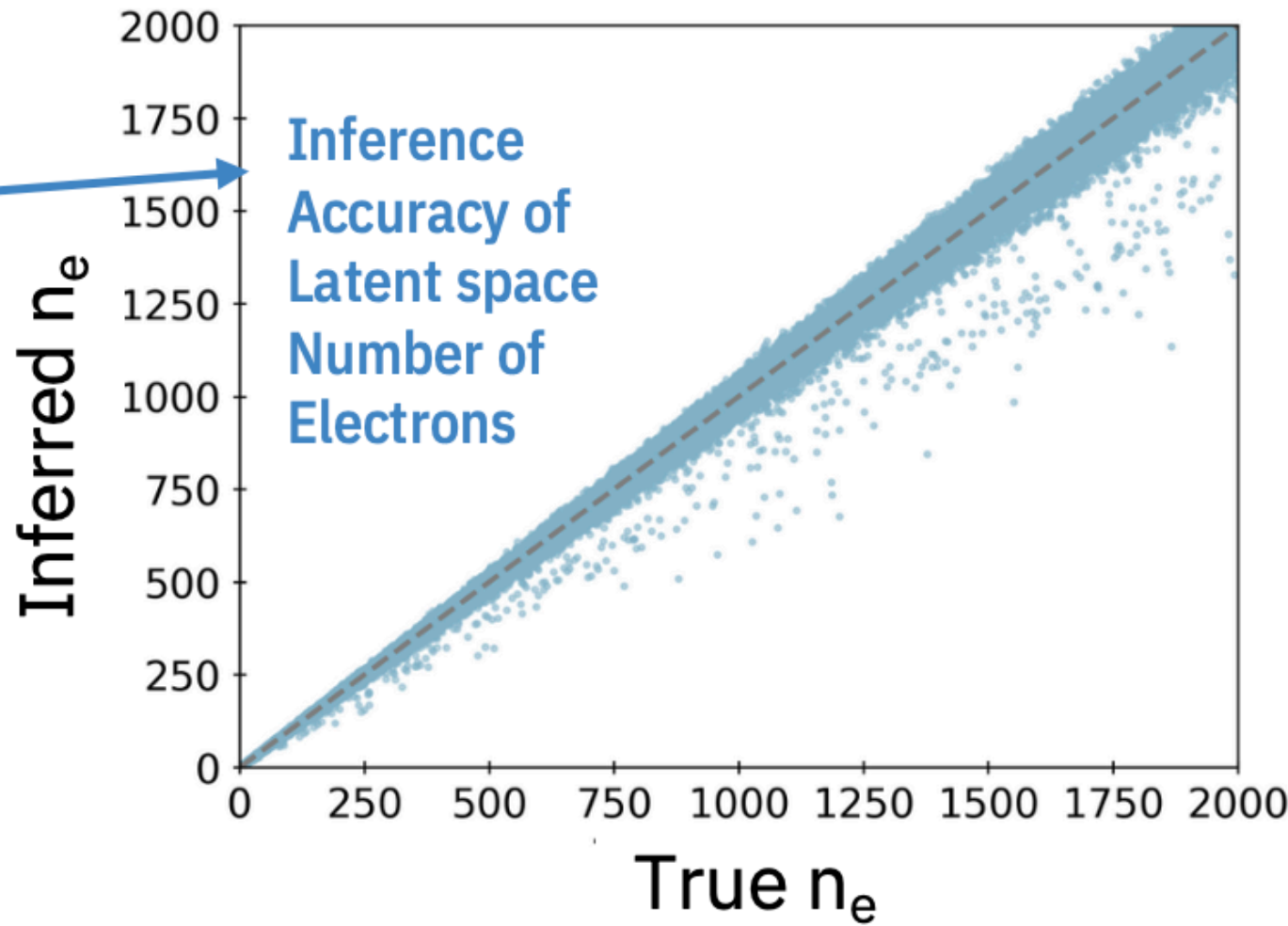# Regression in DM experiments

Ivy Li, Rice
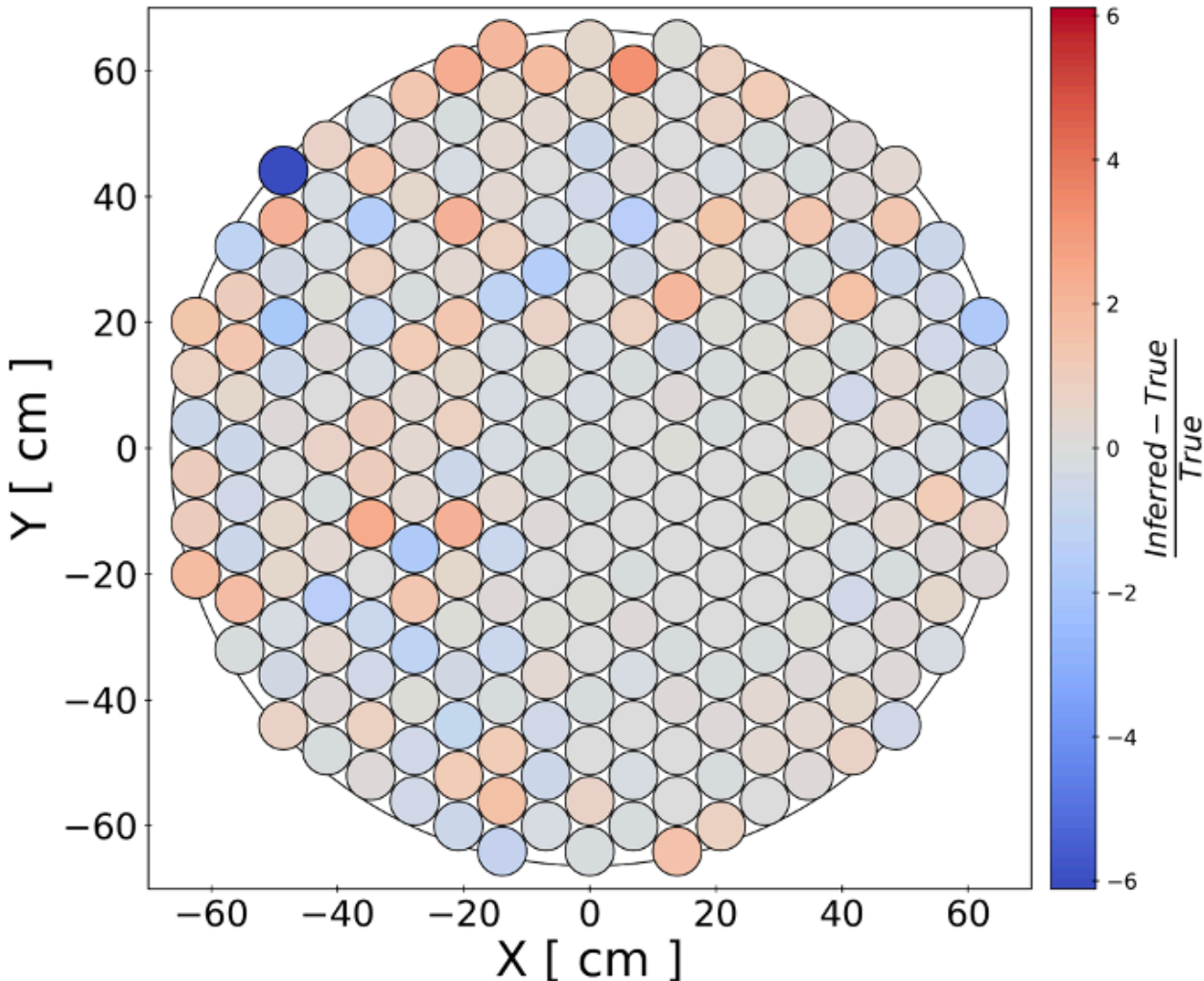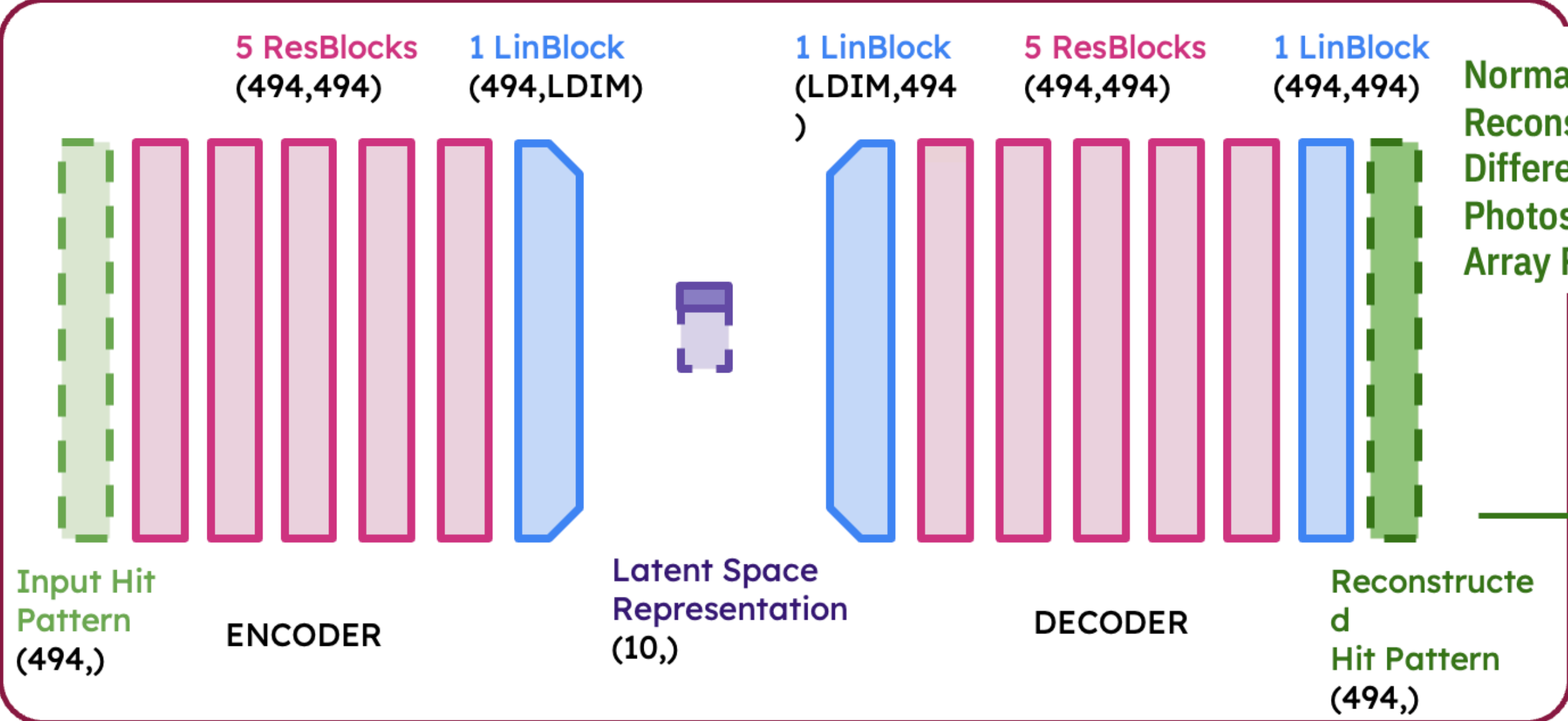
DATASET: simulations from a given $[0, 2000]$ n_e

TRAIN/VALIDATION/TEST SPLIT:
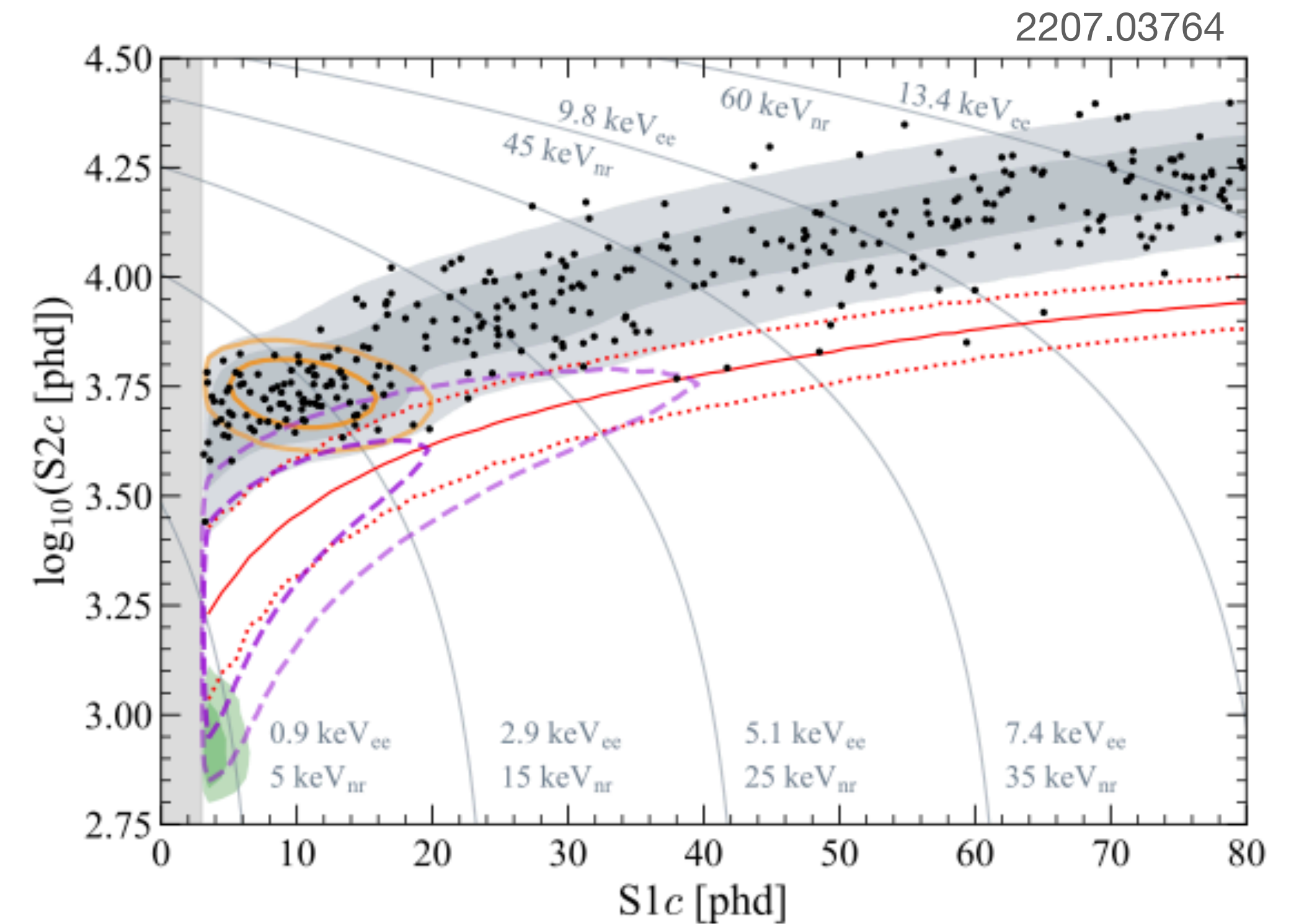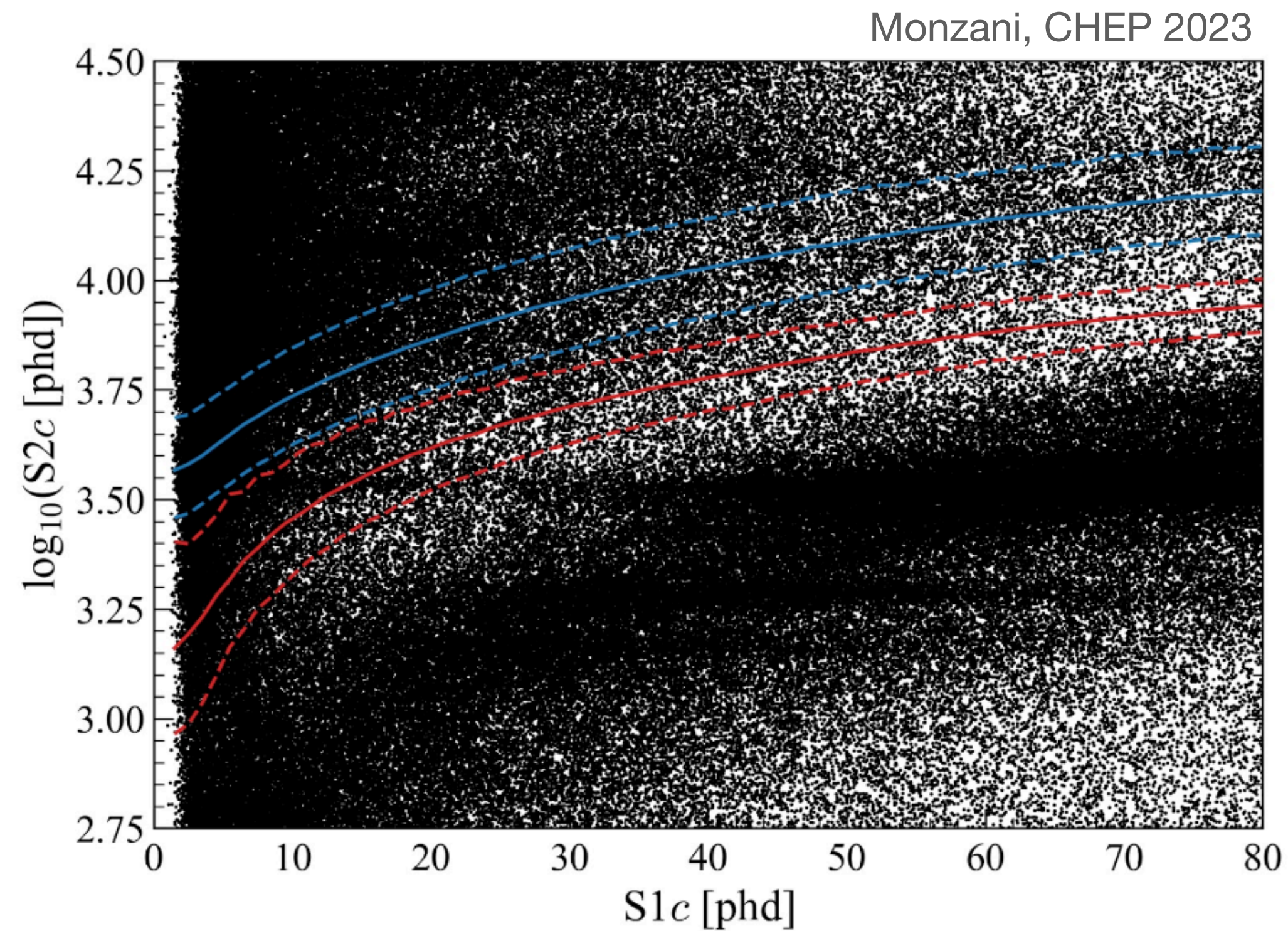447500/447500/100000 hit patterns

LATENT SPACE REPRESENTATION: constrain one value **as number of electrons** but allow the others to evolve freely

LOSS FUNCTION: weighted sum of the MSE loss of reconstruction of hit pattern and MSE loss of inferred number of electrons, trained with cyclic annealing
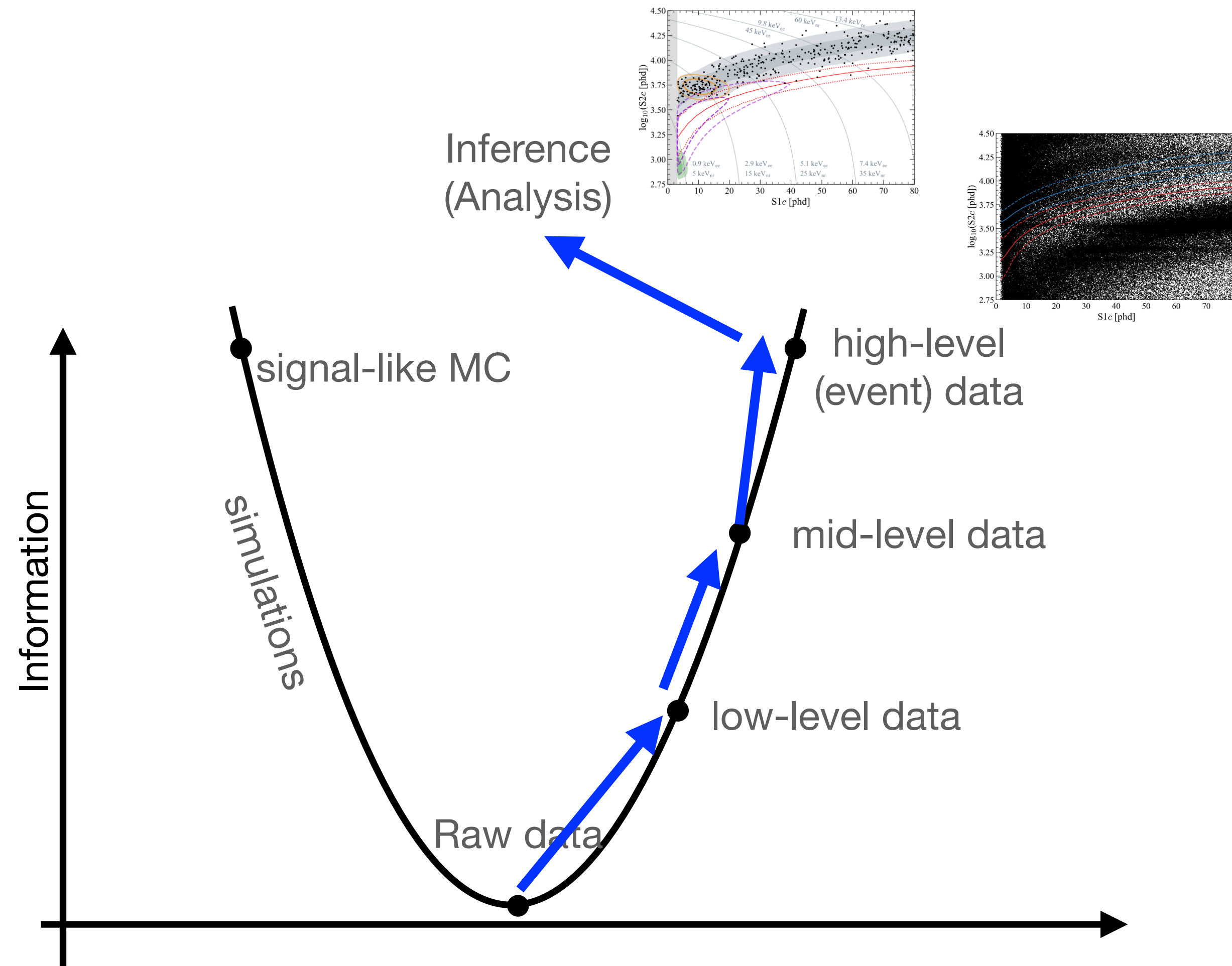
A. Higuera, MODE workshop, Princeton University, July 2023
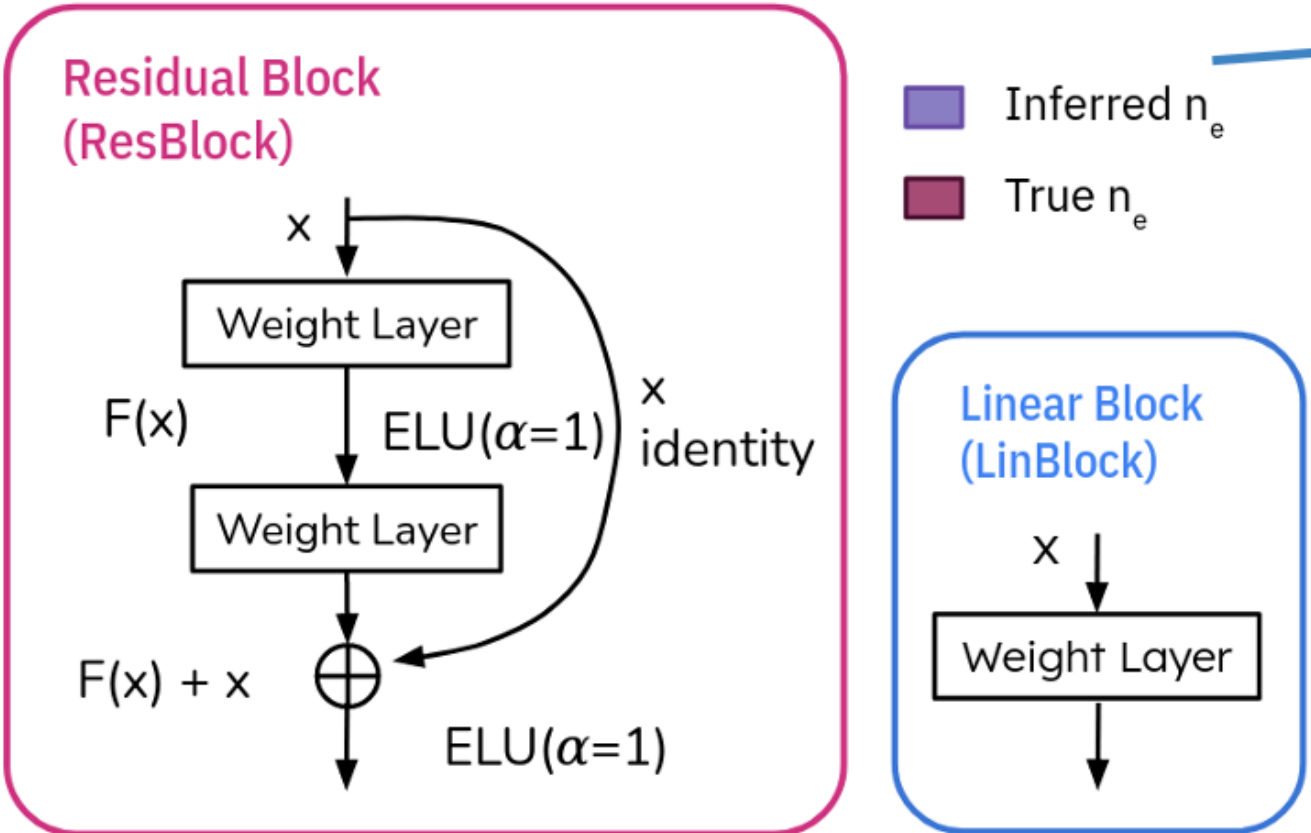
Classification in DM experiments

Everything is background!
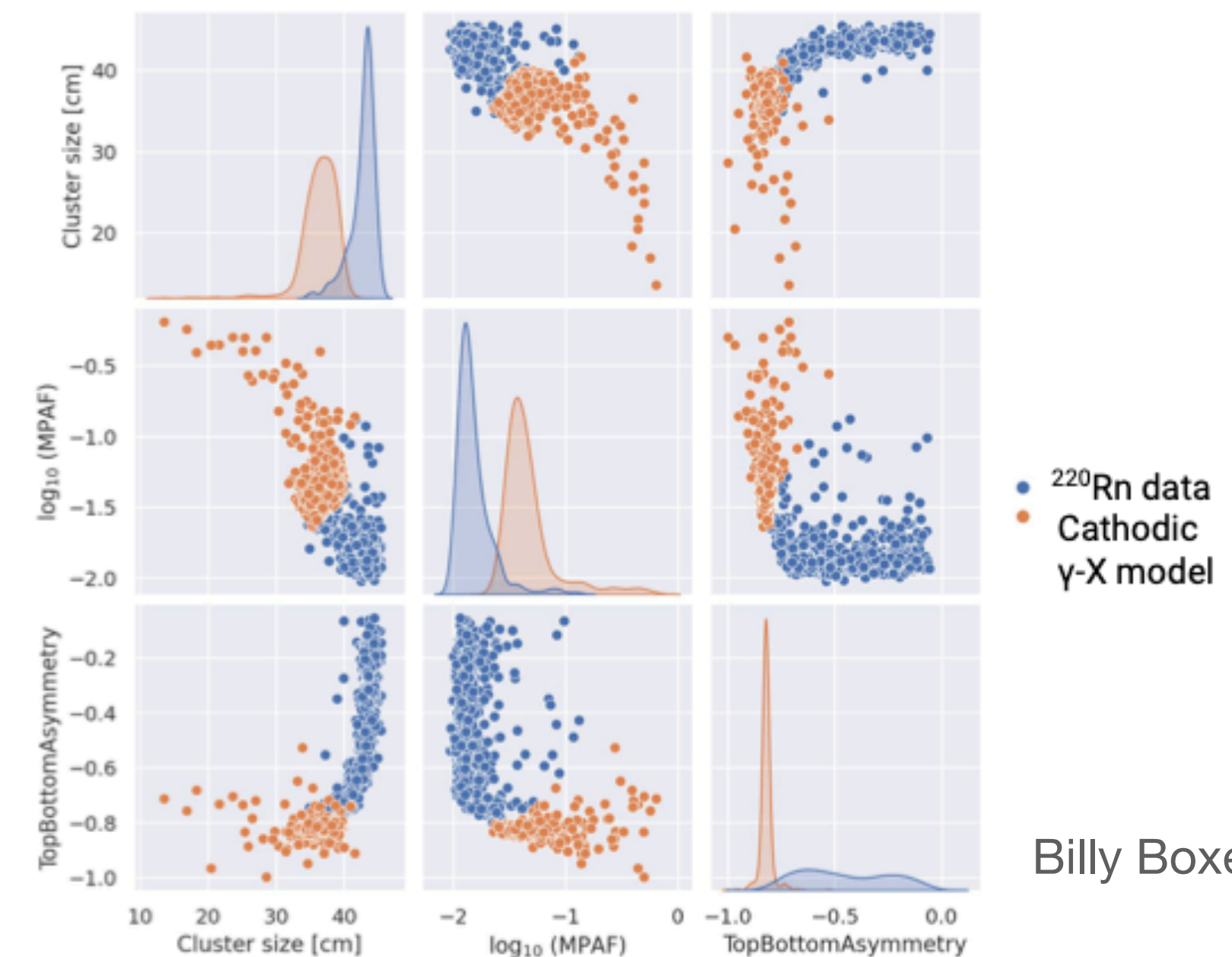

Monzani, CHEP 2023

2207.03764

Classification in DM experiments

Need to classify something? You can never go wrong with

Boosted Decision Trees!!!
- **Easy to understand how the algorithm uses input data**
- **Select high-level features**

Inference (Analysis)

signal-like MC

high-level (event) data

simulations

mid-level data

low-level data

Raw data

Information

^{220}Rn data
Cathodic γ-X model

Billy Boxer, DM 2023, Spain

A. Higuera, MODE workshop, Princeton University, July 2023

Classification in DM experiments

Inference
(Analysis)

signal-like MC

high-level
(event) data

simulations

mid-level data

low-level data
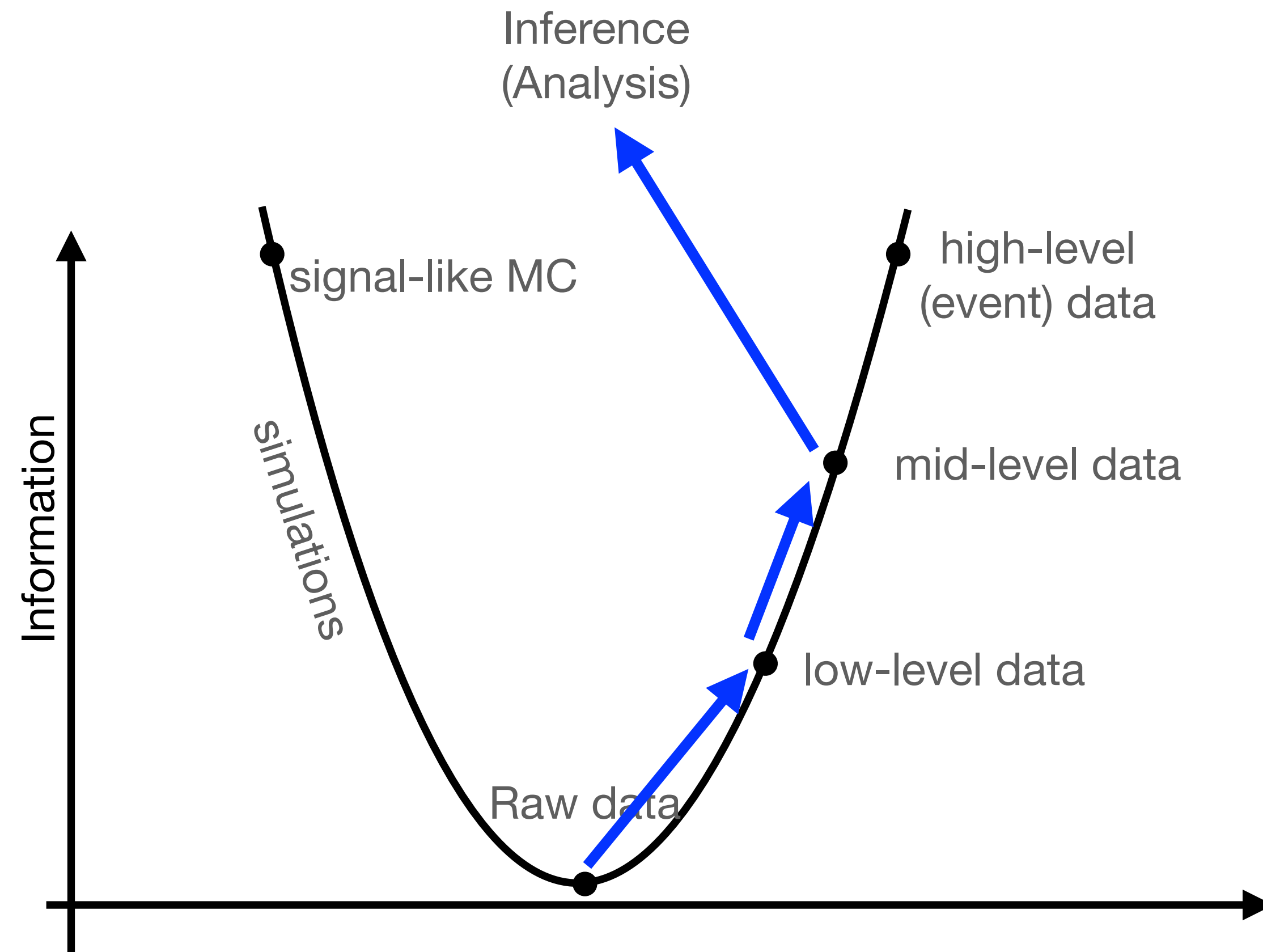
Information

Raw data
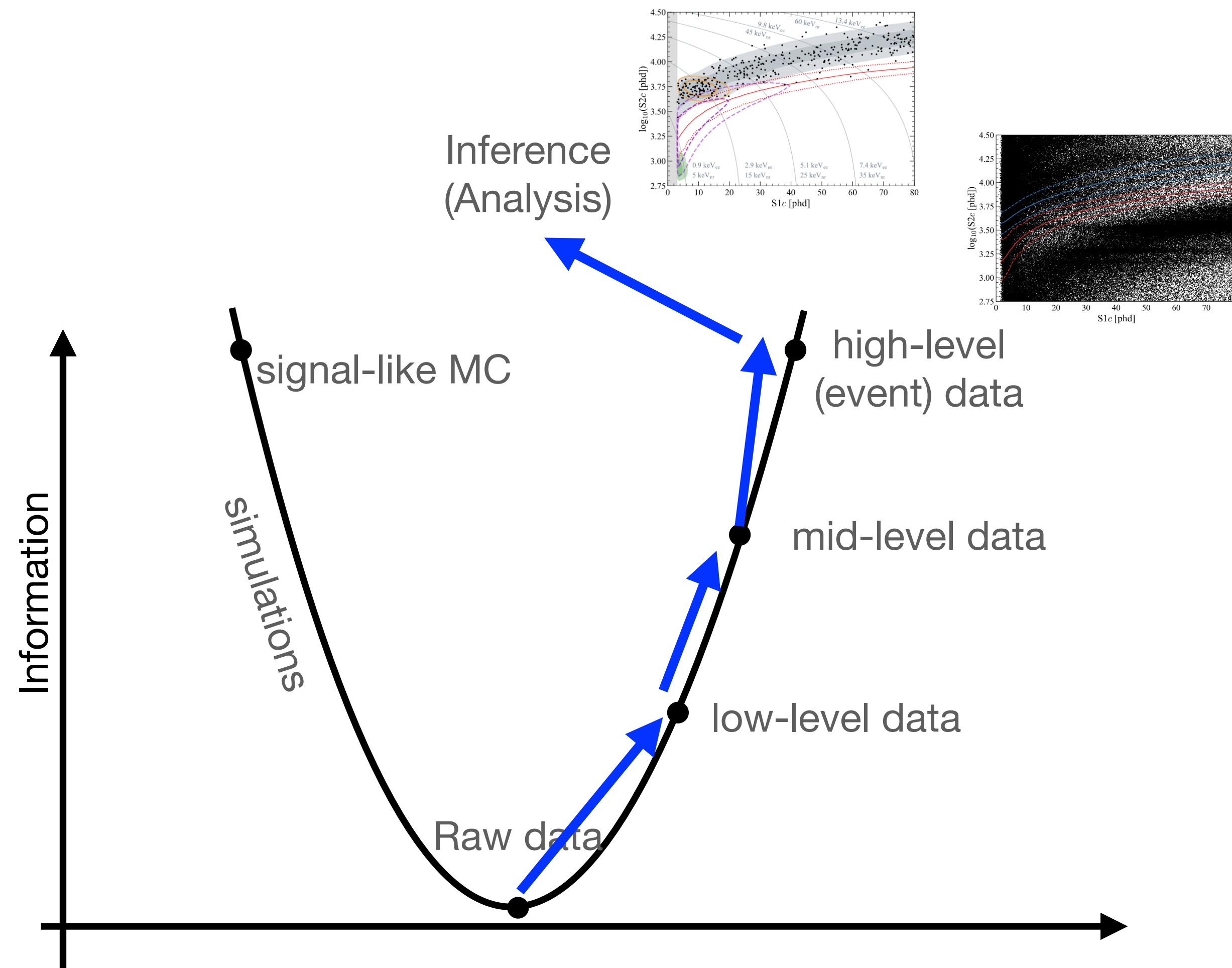
Need to classify something? You can never go wrong with

Boosted Decision Trees!!!

Why do tree-based models still outperform deep learning on tabular data? (arxiv.2207.08815)

- Improving sensitivity to low-mass dark matter in LUX using an novel electrode background mitigation technique (arxiv:2011.0960)
- Study of background from accidental coincidence signals in the PandaX-II experiment (arxiv:2204.11175)
- Boosted decision trees approach to neck alpha events discrimination in DEAP-3600 experiment (arxiv.2009.00895)
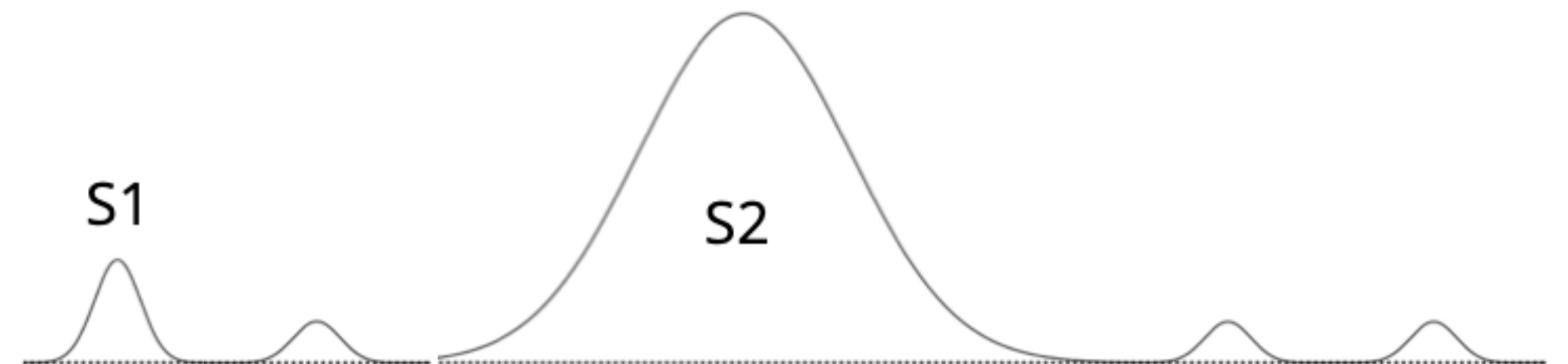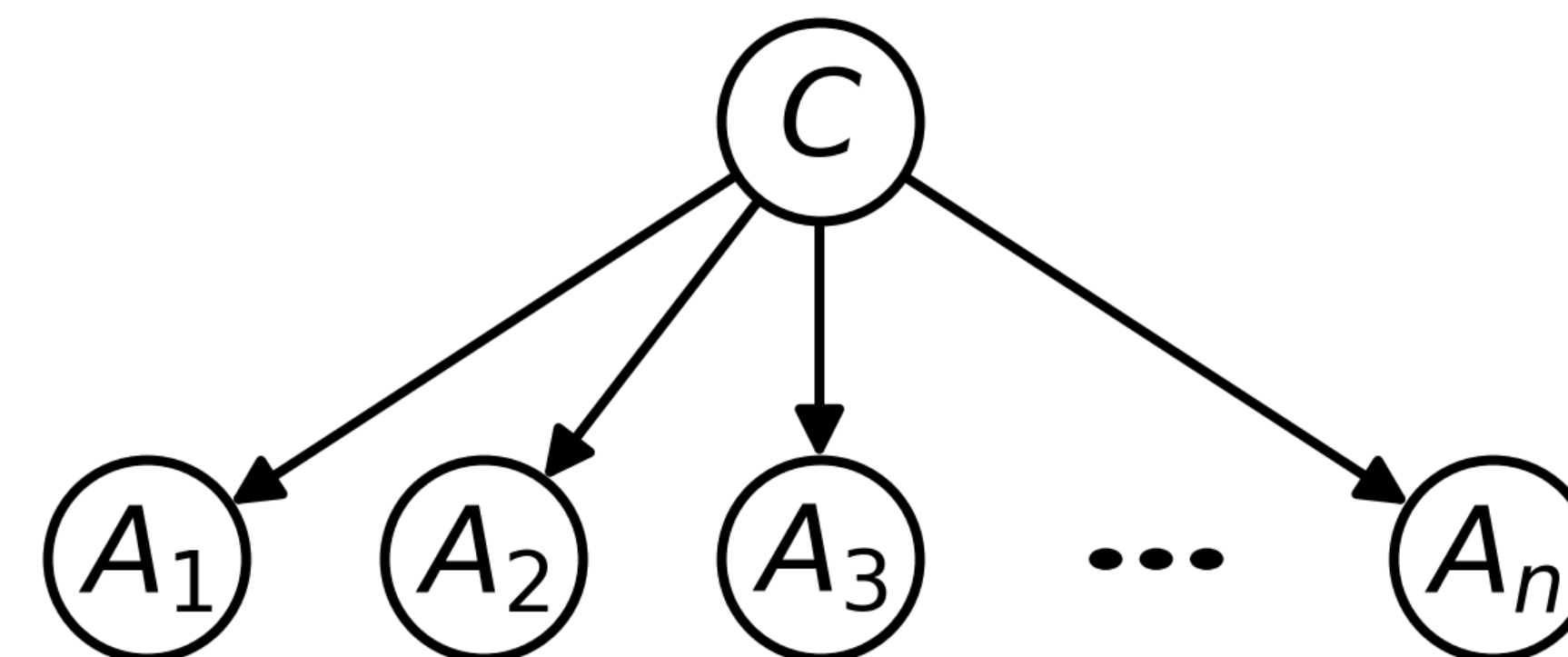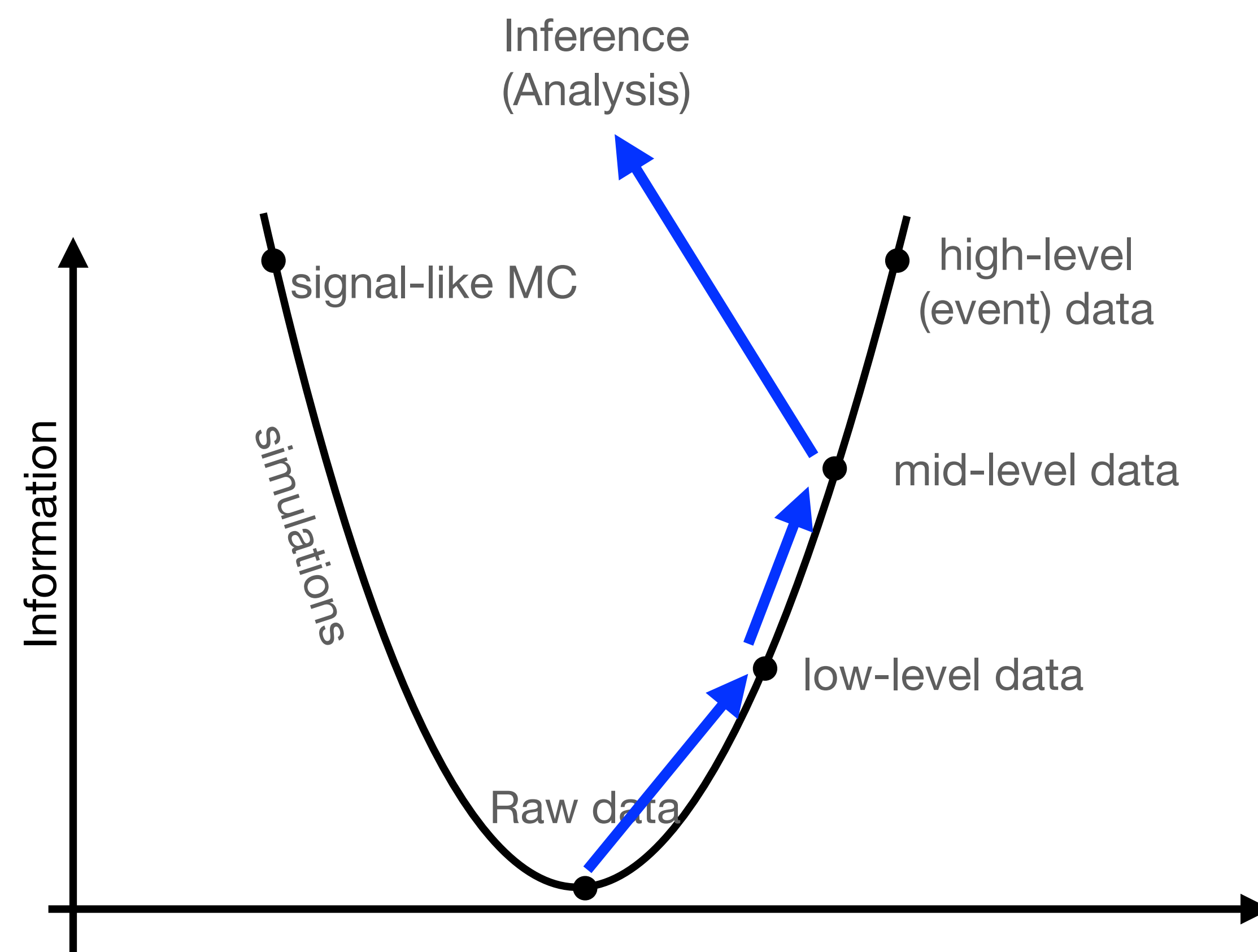
A. Higuera, MODE workshop, Princeton University, July 2023

Classification in DM experiments

Inference
(Analysis)

signal-like MC

high-level
(event) data

simulations

mid-level data

Information

low-level data

Raw data

Why do tree-based models still outperform deep learning on tabular data? (arxiv.2207.08815)

What if you are not dealing with tabular data (event-level data)?

Pulse data (waveform)

S1

S2

Classification in DM experiments

Inference
(Analysis)

signal-like MC

high-level
(event) data

Information

simulations

mid-level data

low-level data

Raw data

C
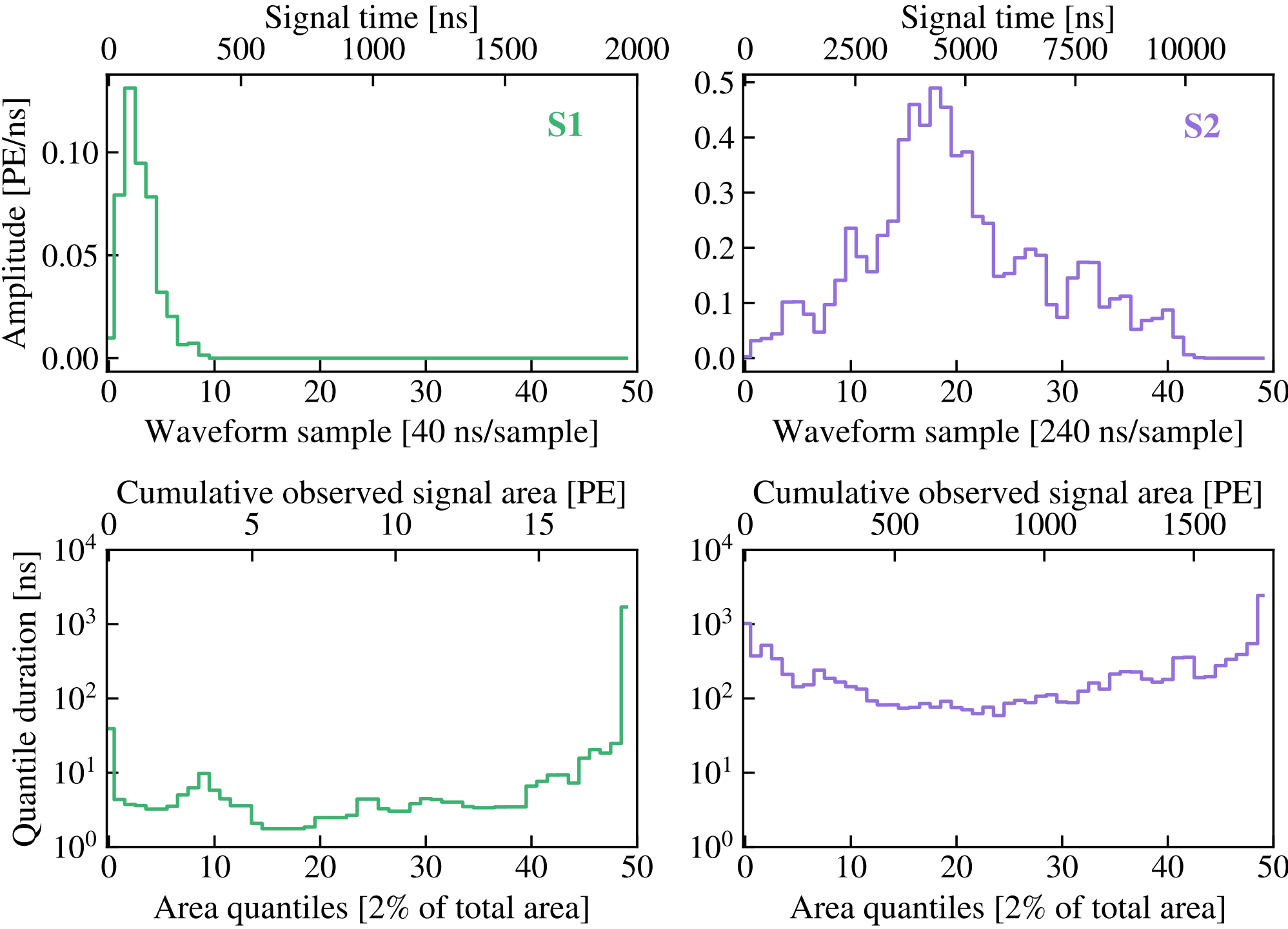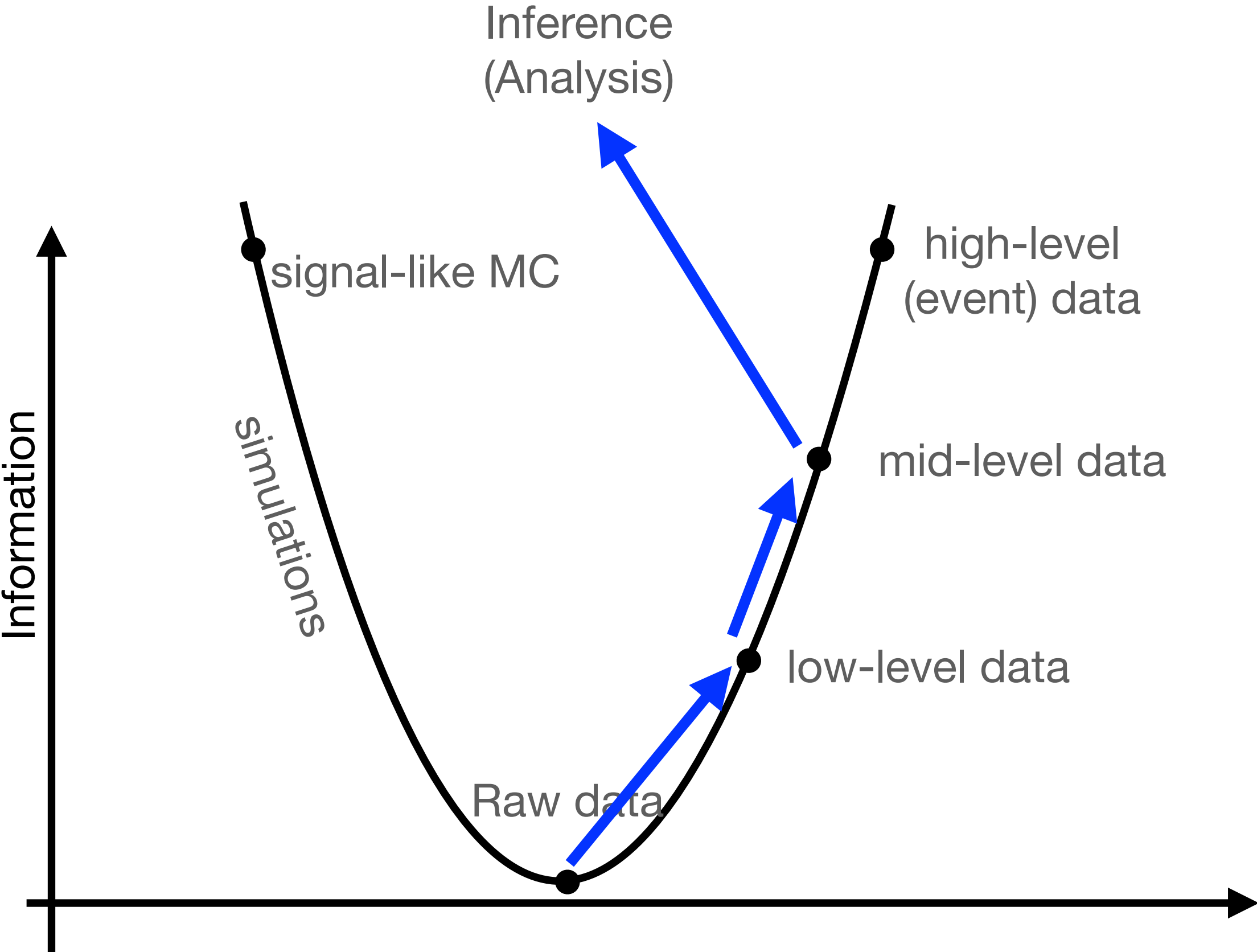
A_1 A_2 A_3 $\cdots$ A_n

**Bayesian Network
Input: waveform attributes
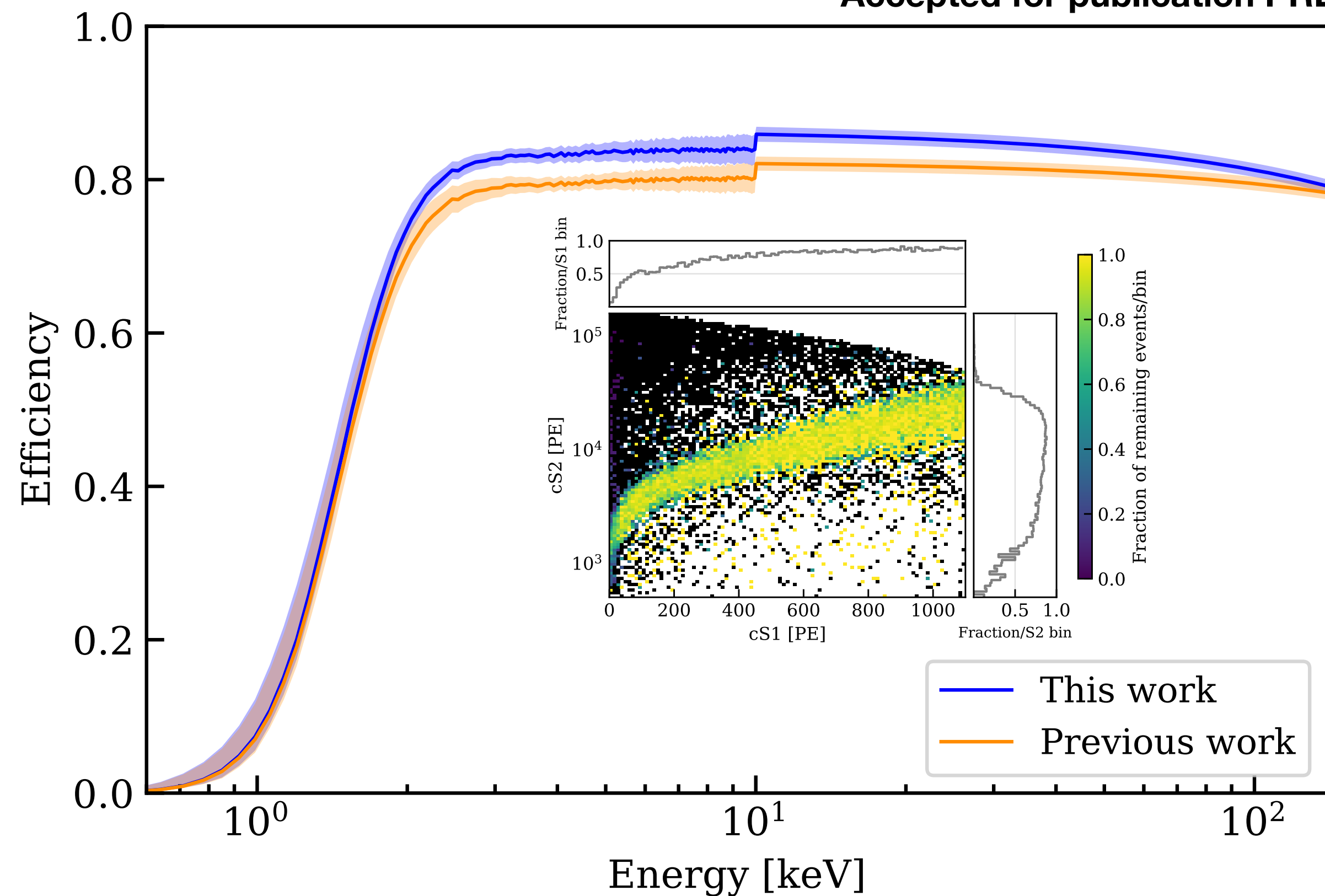Output: Pseudo probability of being signal-like**

S1

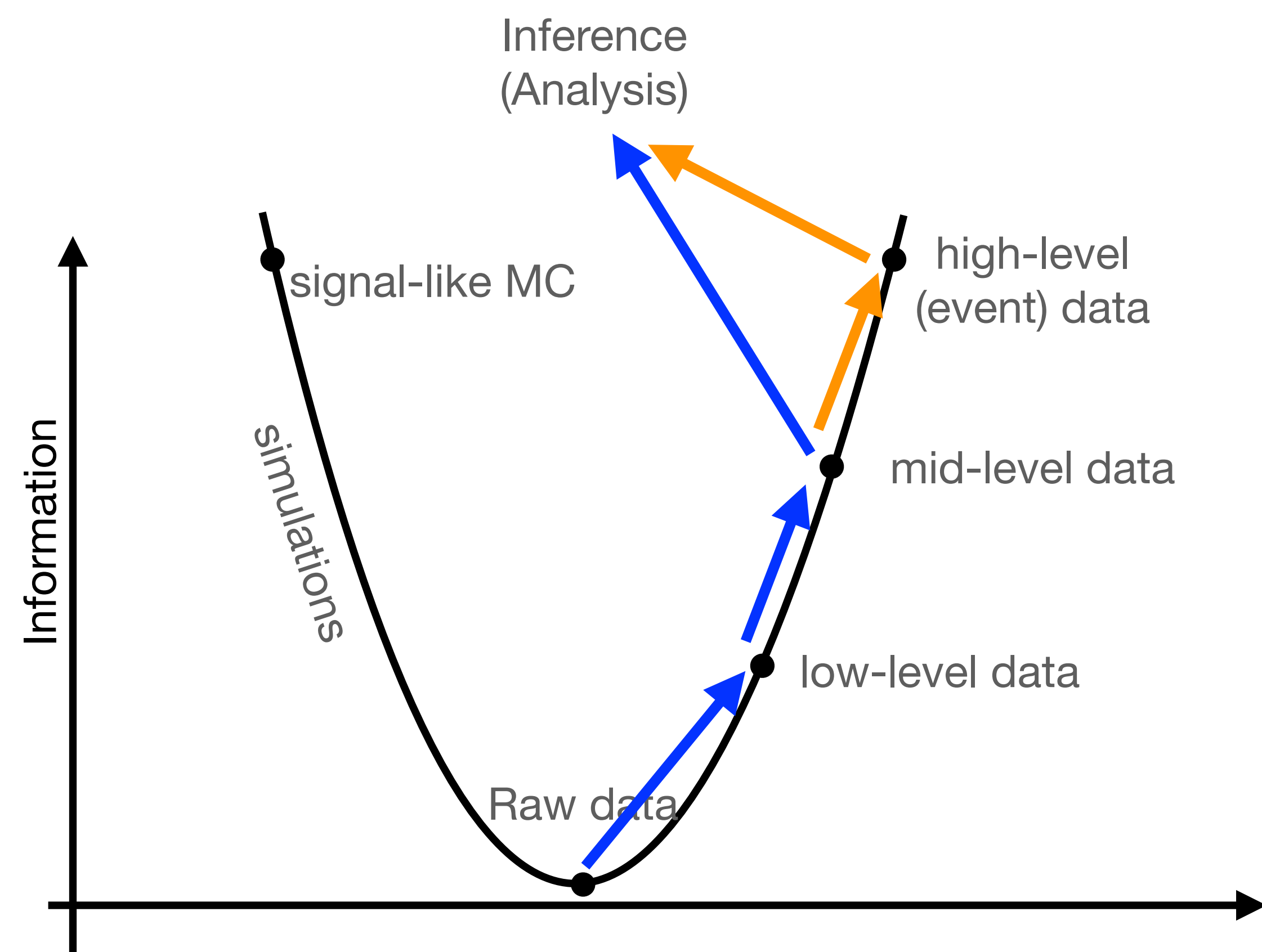S2

A. Higuera, MODE workshop, Princeton University, July 2023

Classification in DM experiments

26
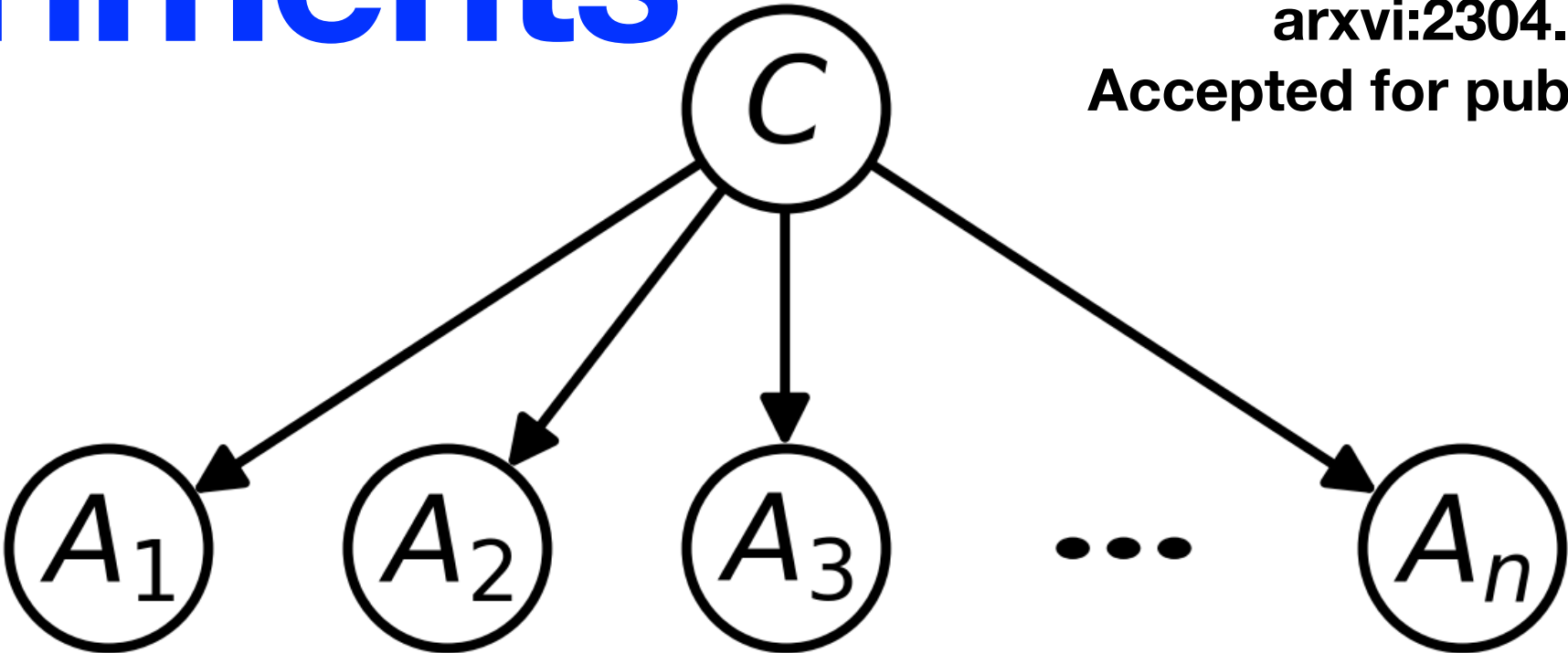
Classification in DM experiments

Using NBC based metric in event selection reduces significantly detector signals outside of the ROI and solves the need for additional selection criteria based on detector signal quality

Increase the selection efficiency from previous method

A. Higuera, MODE workshop, Princeton University, July 2023

27

Anomaly detection in DM experiments

Inference
(Analysis)

signal-like MC

high-level
(event) data

simulations

mid-level data

Information
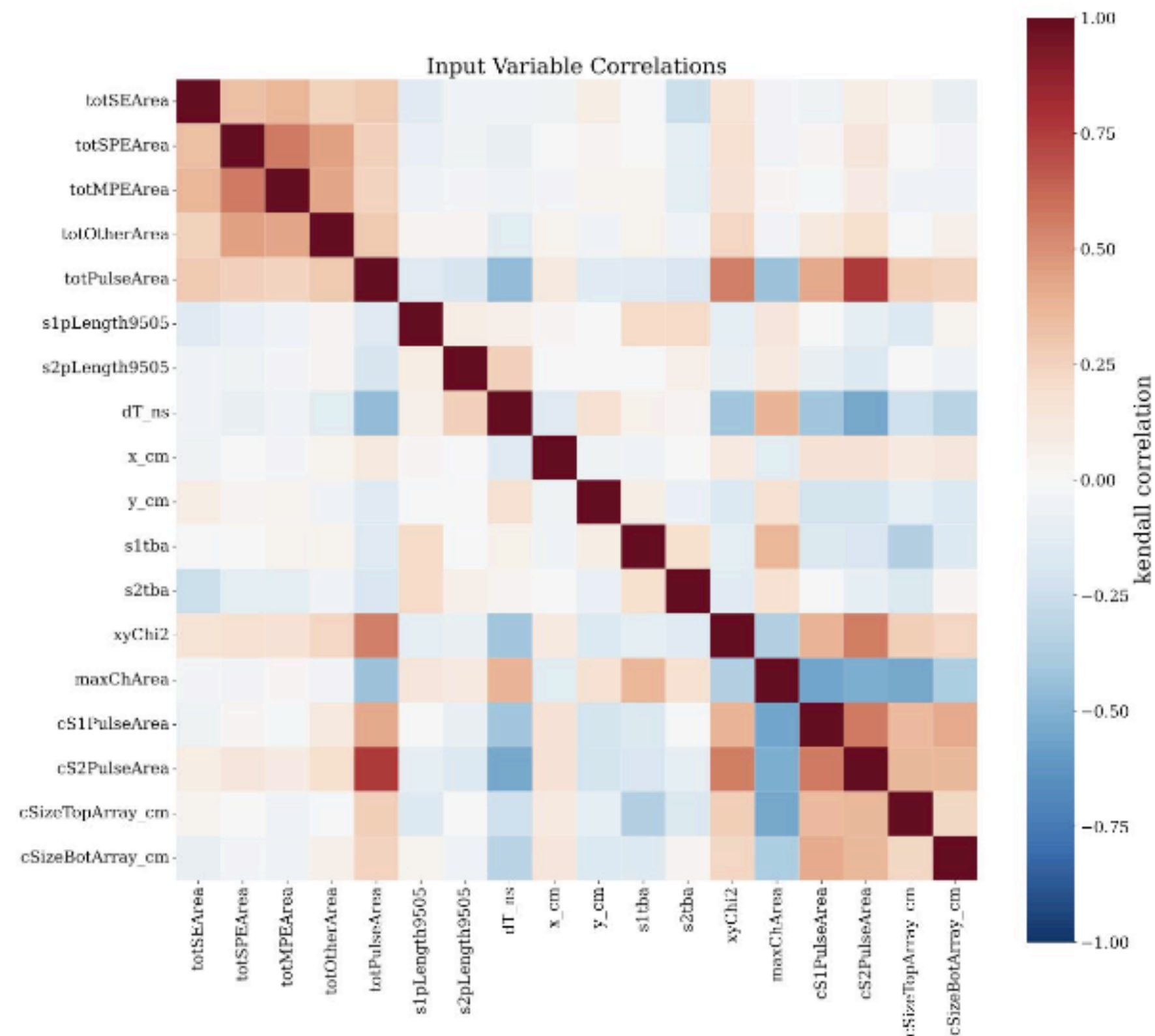
low-level data

Raw data
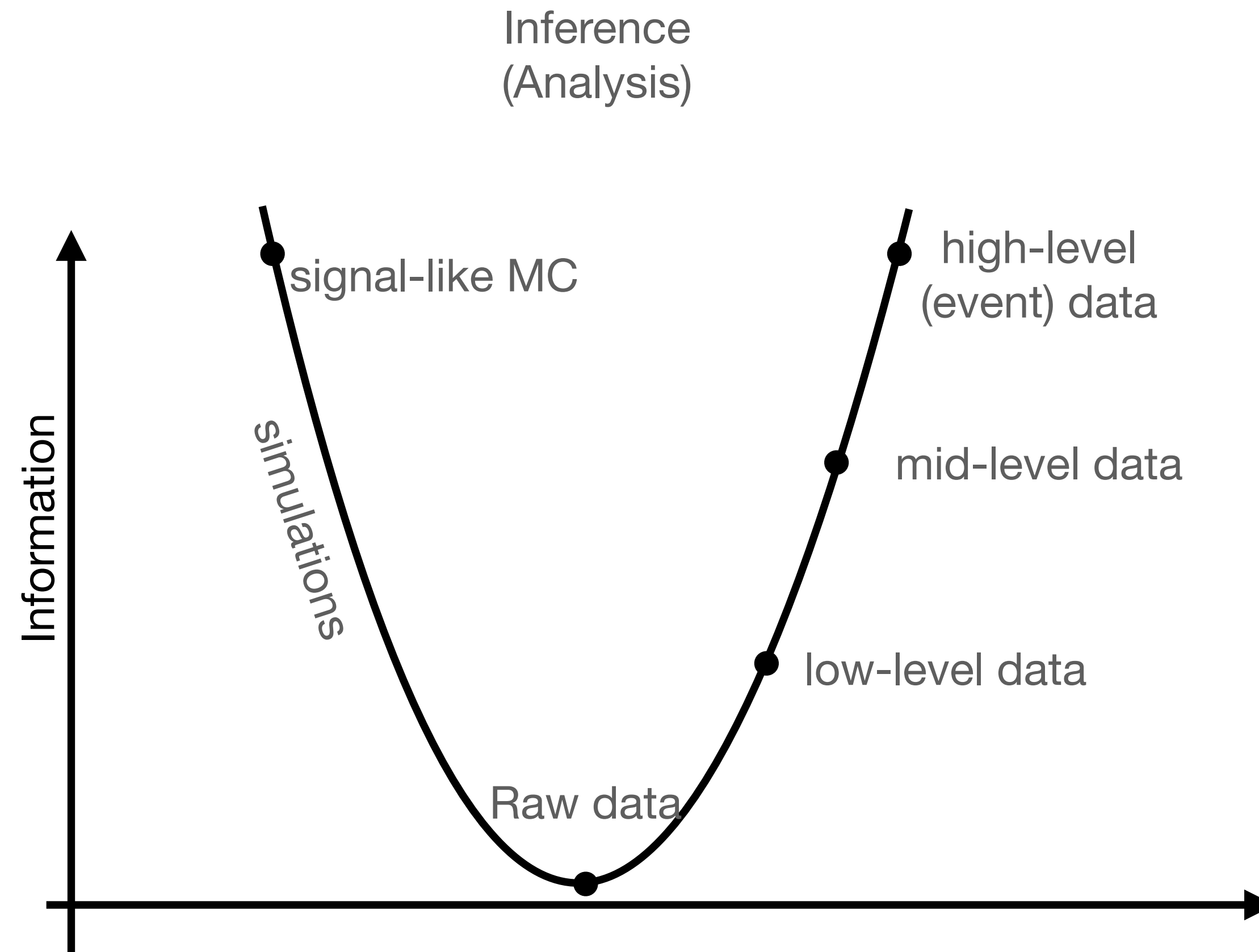
Identify and interpret anomalous data

Using high-dimension spaces with a combination of high-level data and mid-level data

With dimensional reduction and/or clustering

A. Higuera, MODE workshop, Princeton University, July 2023

28

Anomaly detection in DM experiments
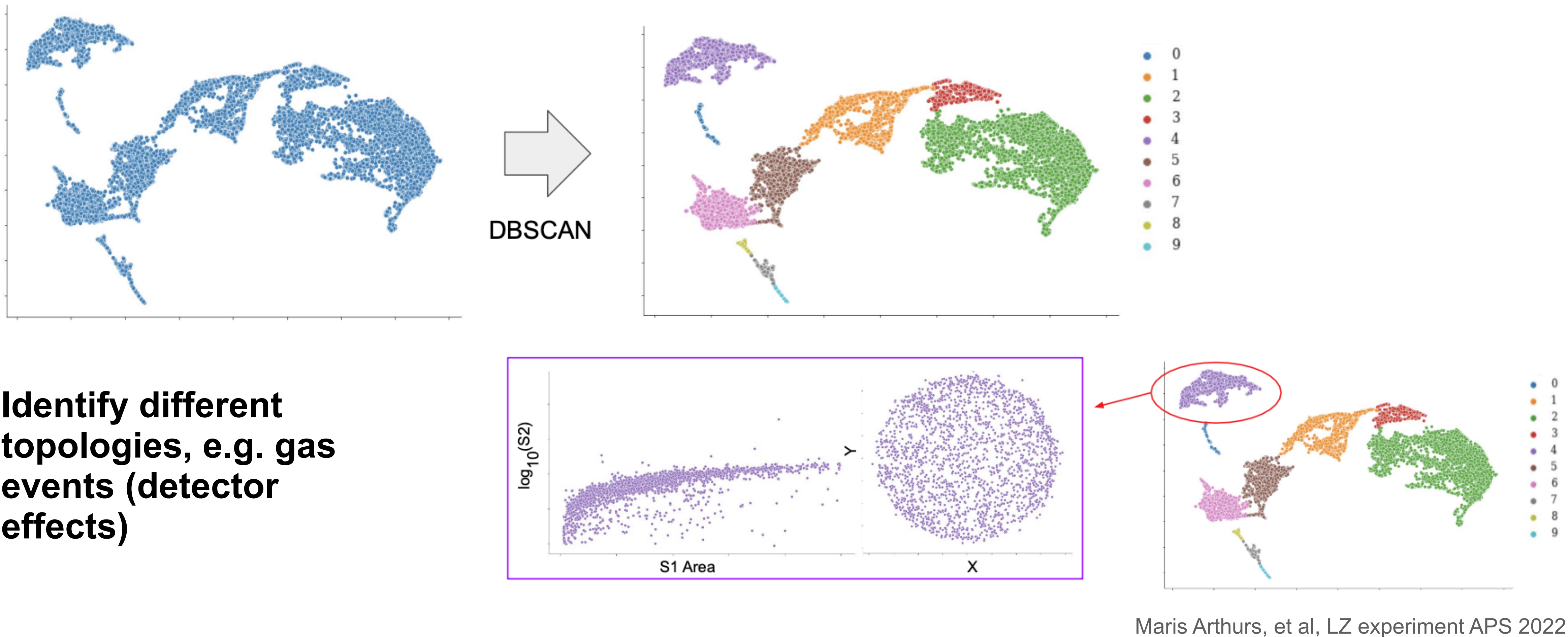
Maris Arthurs, et al, LZ experiment APS 2022

UMAP

Uniform Mainfold Approximation and Projection
A novel manifold learning technique for dimension reduction.
UMAP is an algorithm for dimension reduction based on manifold
learning techniques and ideas from topological data analysis, it
preserves more of the global structure with superior run time
performance (arxiv:1802.03426)

A. Higuera, MODE workshop, Princeton University, July 2023

Anomaly detection in DM experiments

DBSCAN

Identify different topologies, e.g. gas events (detector effects)

Maris Arthurs, et al, LZ experiment APS 2022

A. Higuera, MODE workshop, Princeton University, July 2023
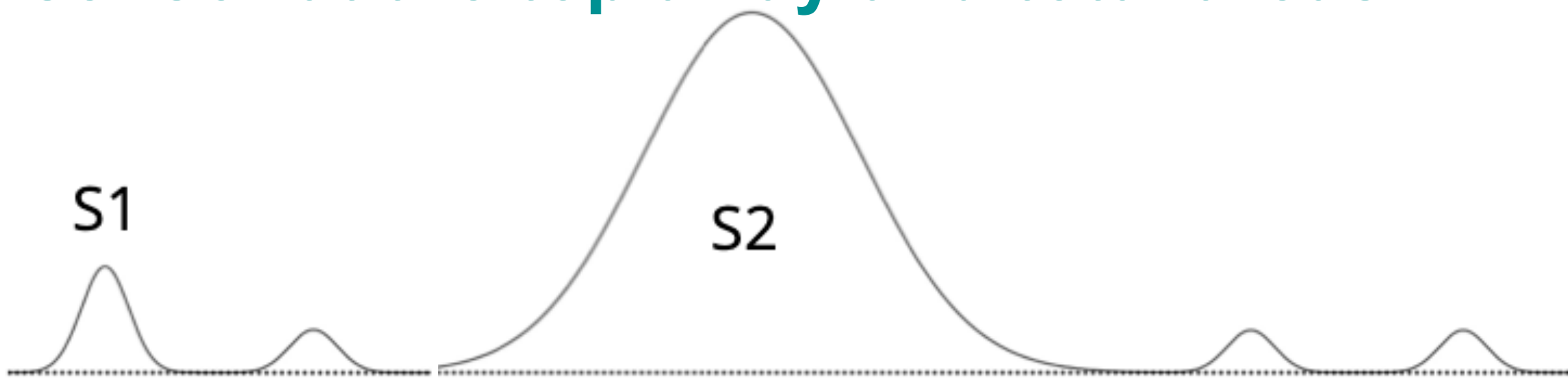
Anomaly detection in DM experiments

Luis Sanchez & Sanya Arora, Rice

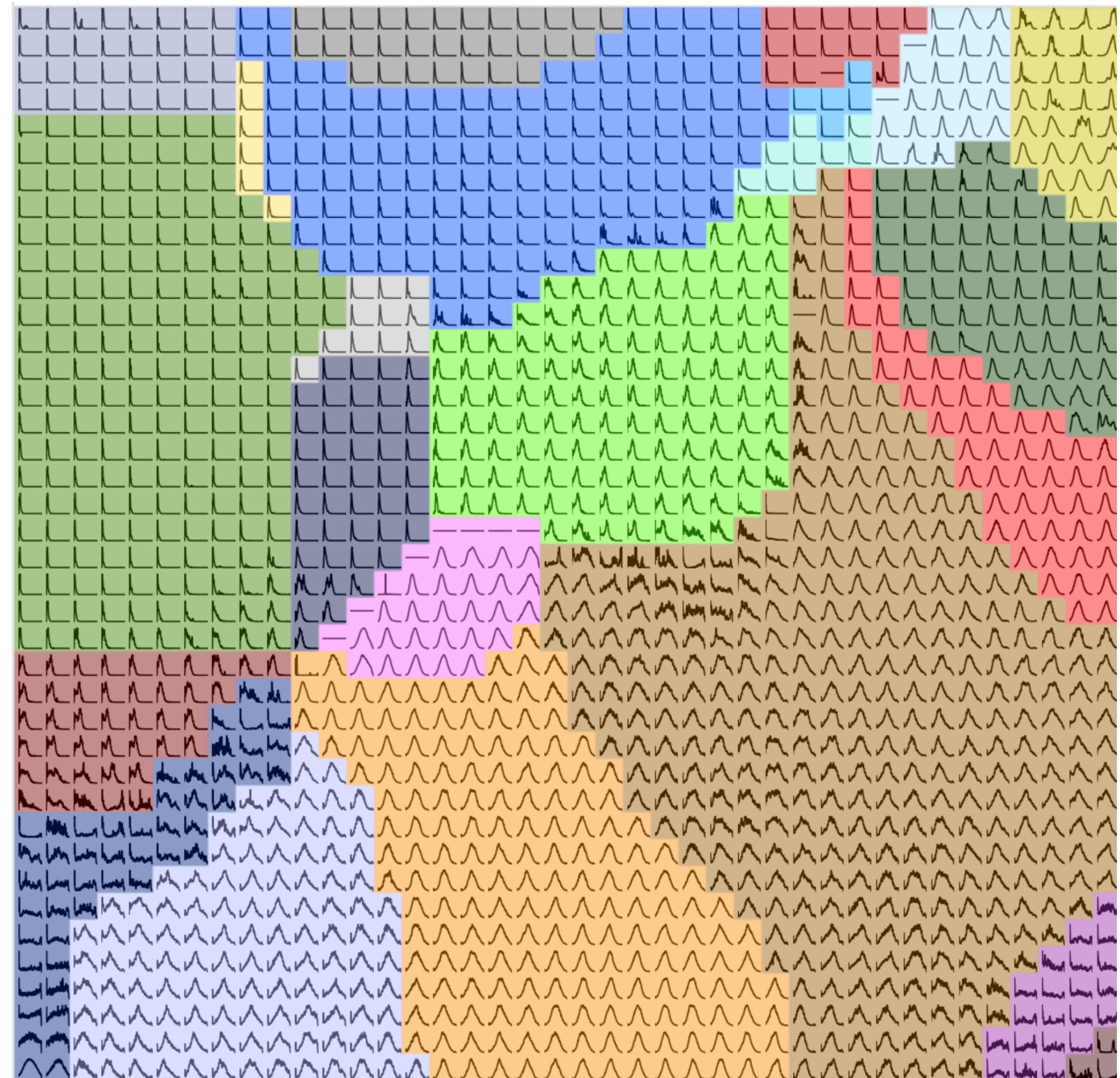**Identify and interpret anomalous data**

Self Organizing Maps

An unsupervised technique using competitive training to produce a low-dimensional representation of a higher dimensional data set while preserving the topological structure of the data

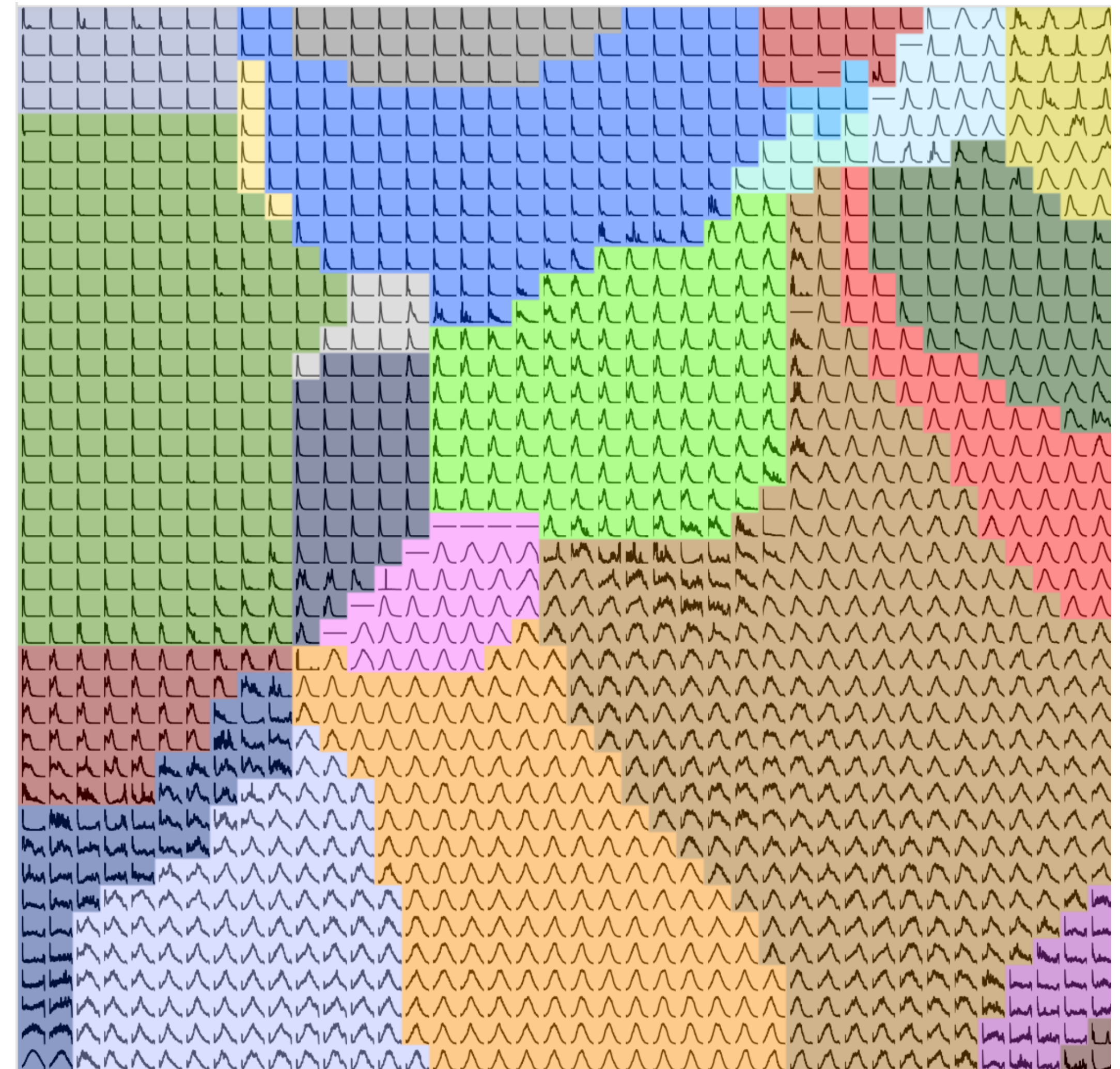**Input data: deciles + fraction area per sensor at the top array and total areas**

S1   S2

A. Higuera, MODE workshop, Princeton University, July 2023

Anomaly detection in DM experiments

Luis Sanchez & Sanya Arora, Rice

Identify and interpret anomalous data
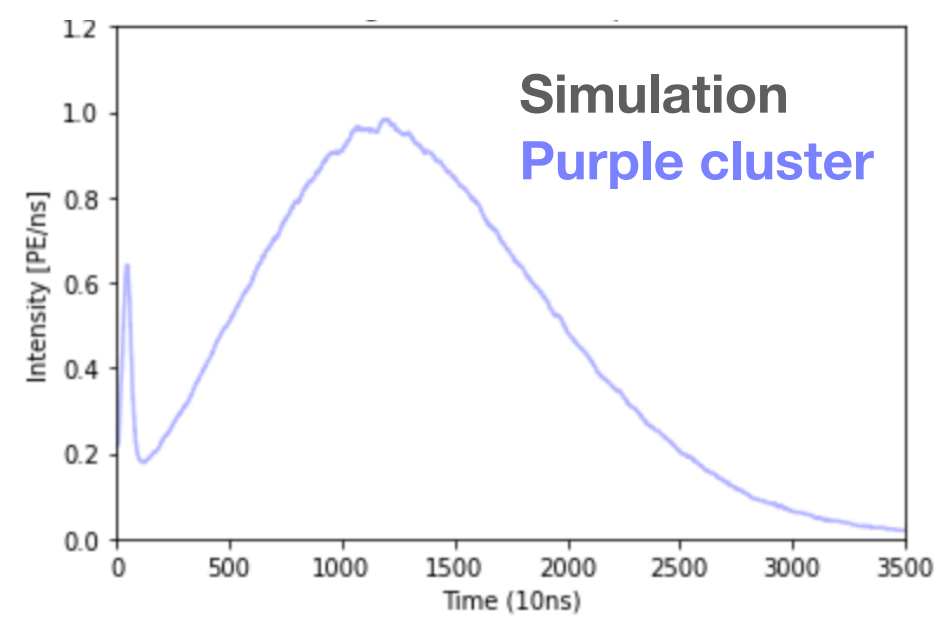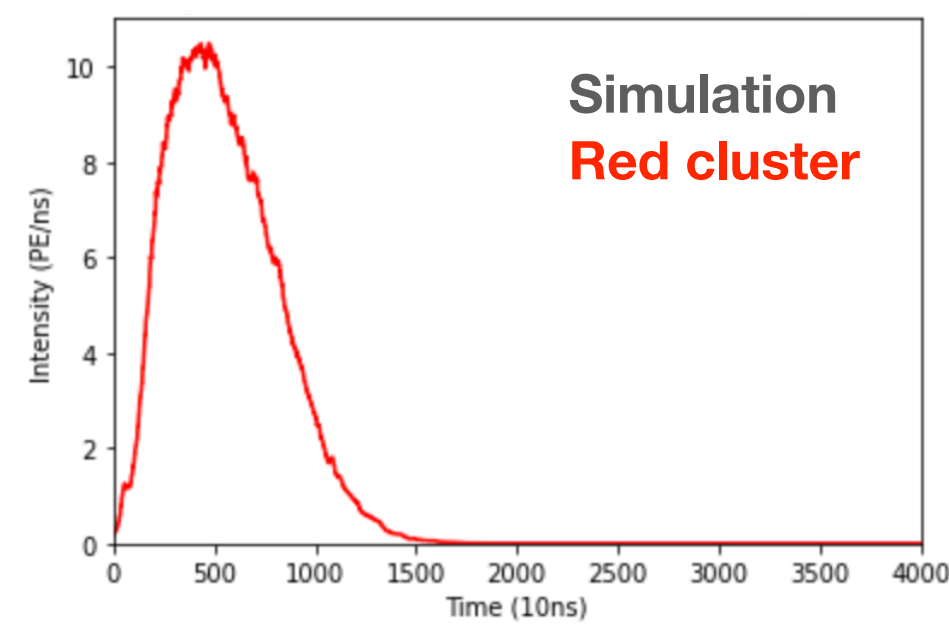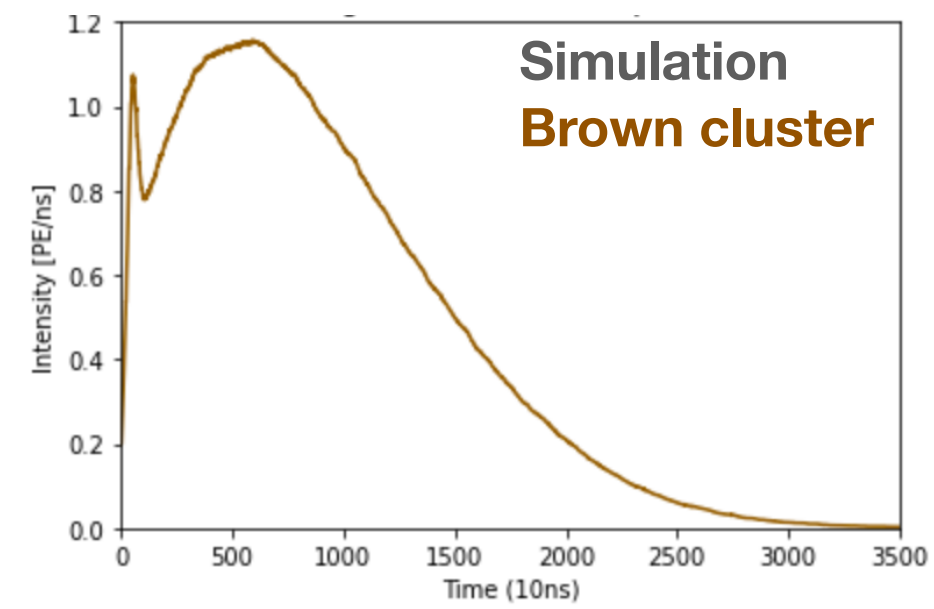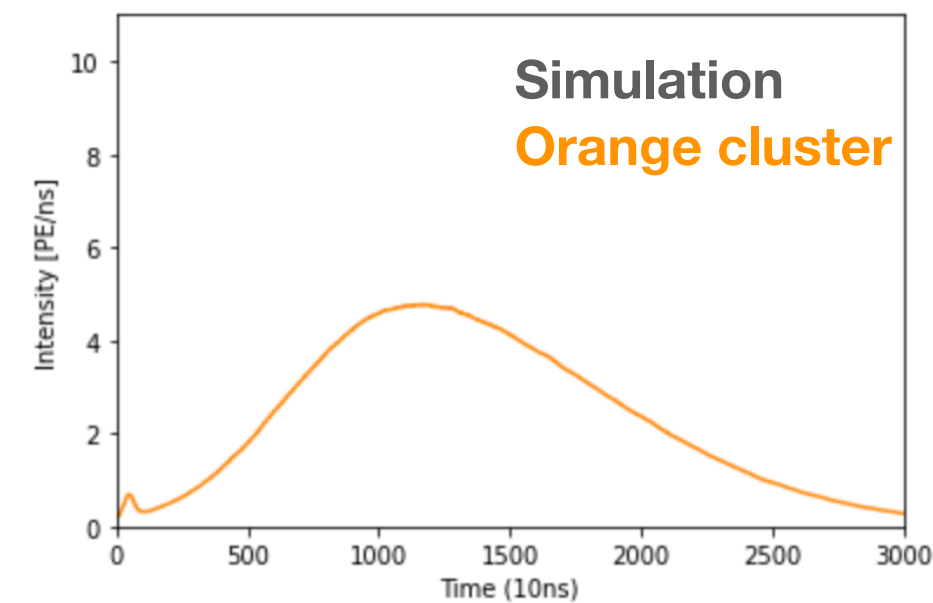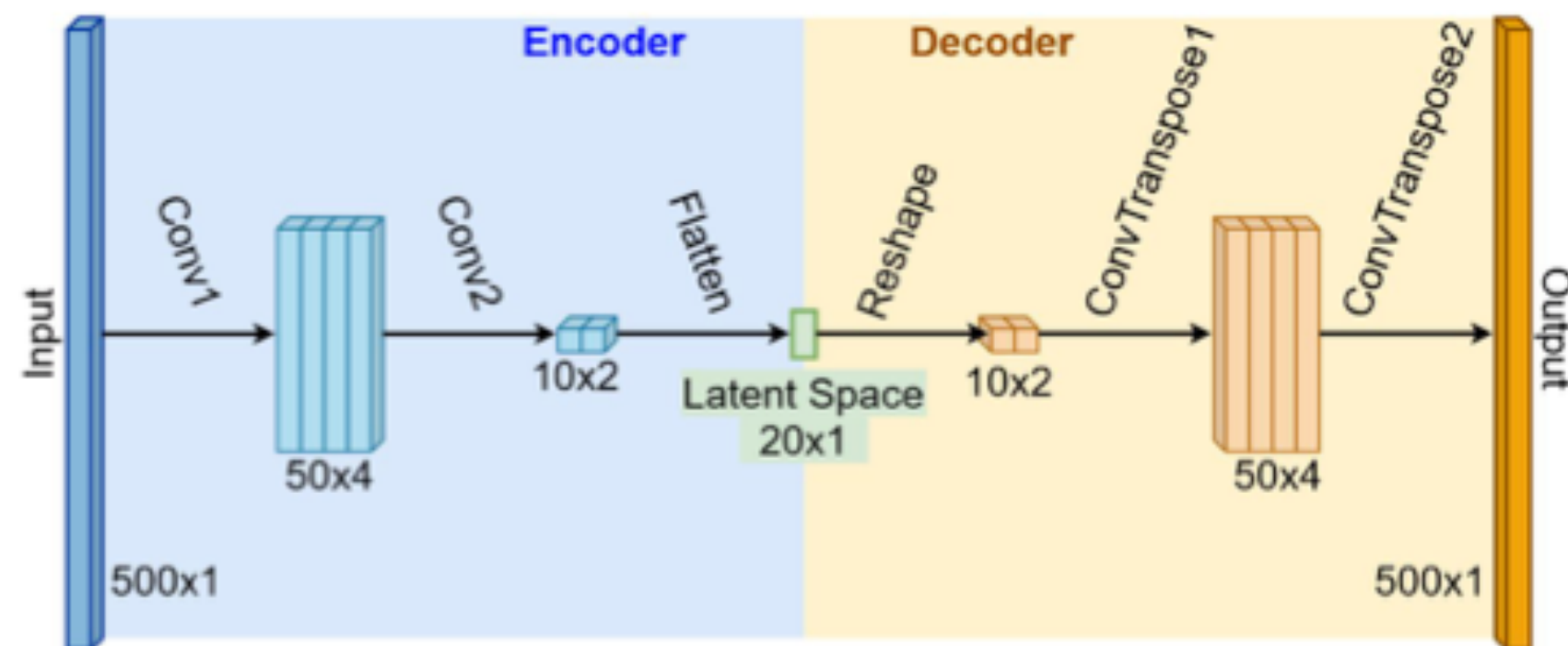### Self Organizing Maps

A. Higuera, MODE workshop, Princeton University, July 2023

Anomaly detection in DM experiments

Identify and interpret anomalous data

Autoencoders

Waveform preparation:
- Downsampled to 200 MHz
- Normalized by max amplitude
- Smoothed
- Zero-padded to 500 samples

Anomalous Quality S2s

Likely Multiple Scatter!

One class of anomalous S2 in the training dataset looks inconsistent with single scatters. This method is able to flag these as likely backgrounds.

Low Energy

The second class comprises jagged low energy S2s. These are not necessarily backgrounds and show the limitations of this method with low energy S2s.

A. Higuera, MODE workshop, Princeton University, July 2023

Inferences in DM experiments

Inference (Analysis)

signal-like MC

high-level (event) data

Information

simulations

mid-level data

low-level data

Raw data

"The standard HEP workflow"

- We are searching for processes that have been predicted by a model but not yet seen
- If no observation we report the expected significance (sensitivity)

PandaX-4T (2021)

XENON1T (2018)

XENONnT (this work)

LZ (2022)

WIMP-nucleon cross-section $\sigma^{\mathrm{SI}}[\mathrm{cm}^2]$

WIMP Mass M_{DM} [GeV/c^2]

A. Higuera, MODE workshop, Princeton University, July 2023

Inferences in DM experiments

Inference (Analysis)

Information

simulations
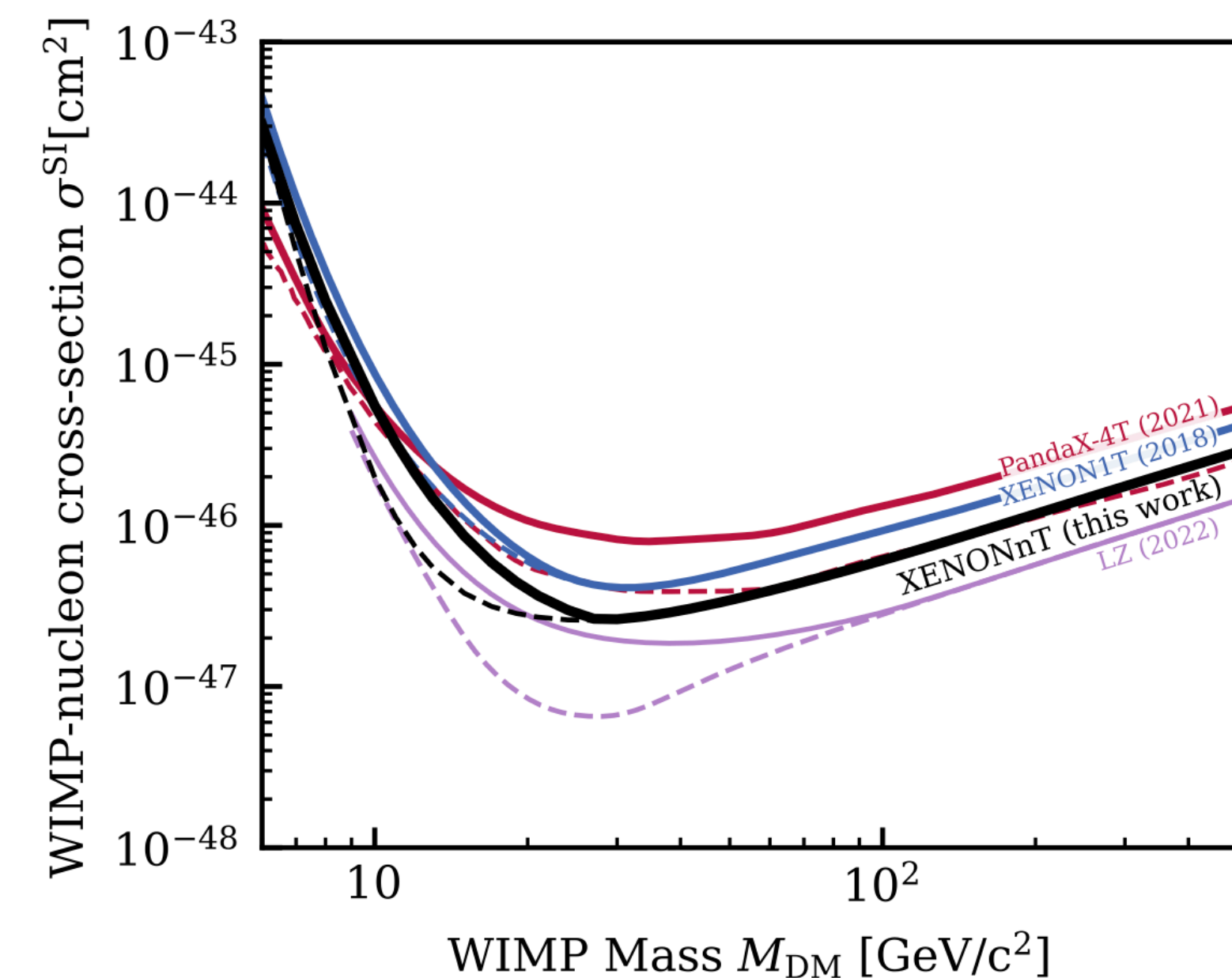
signal-like MC

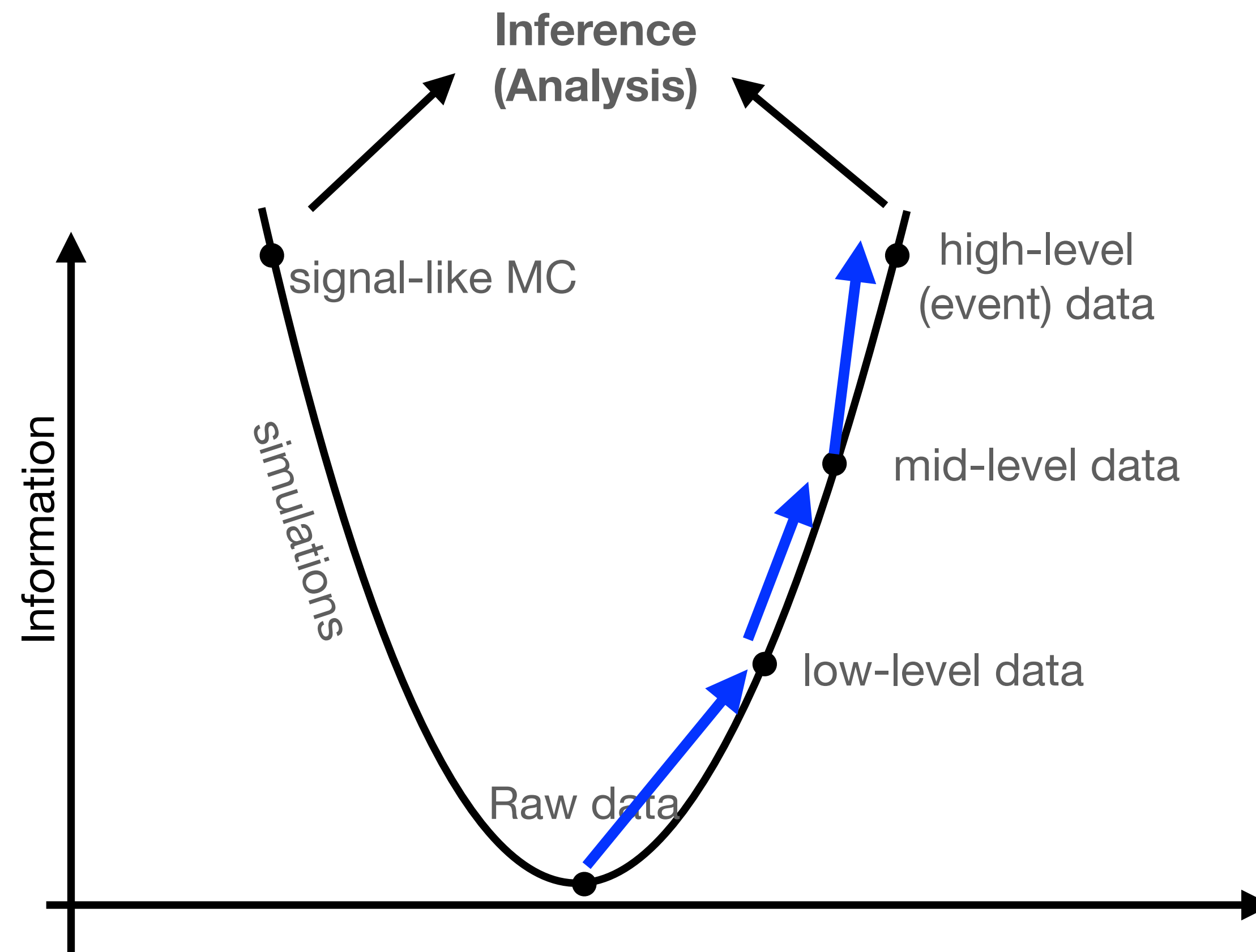high-level (event) data

mid-level data

low-level data

Raw data

"The standard HEP workflow"

- We are searching for processes that have been predicted by a model but not yet seen
- If no observation we report the expected significance (sensitivity)
- The most common approach is calculating a profile likelihood ratio … significant computational burden

Inferences in DM experiments
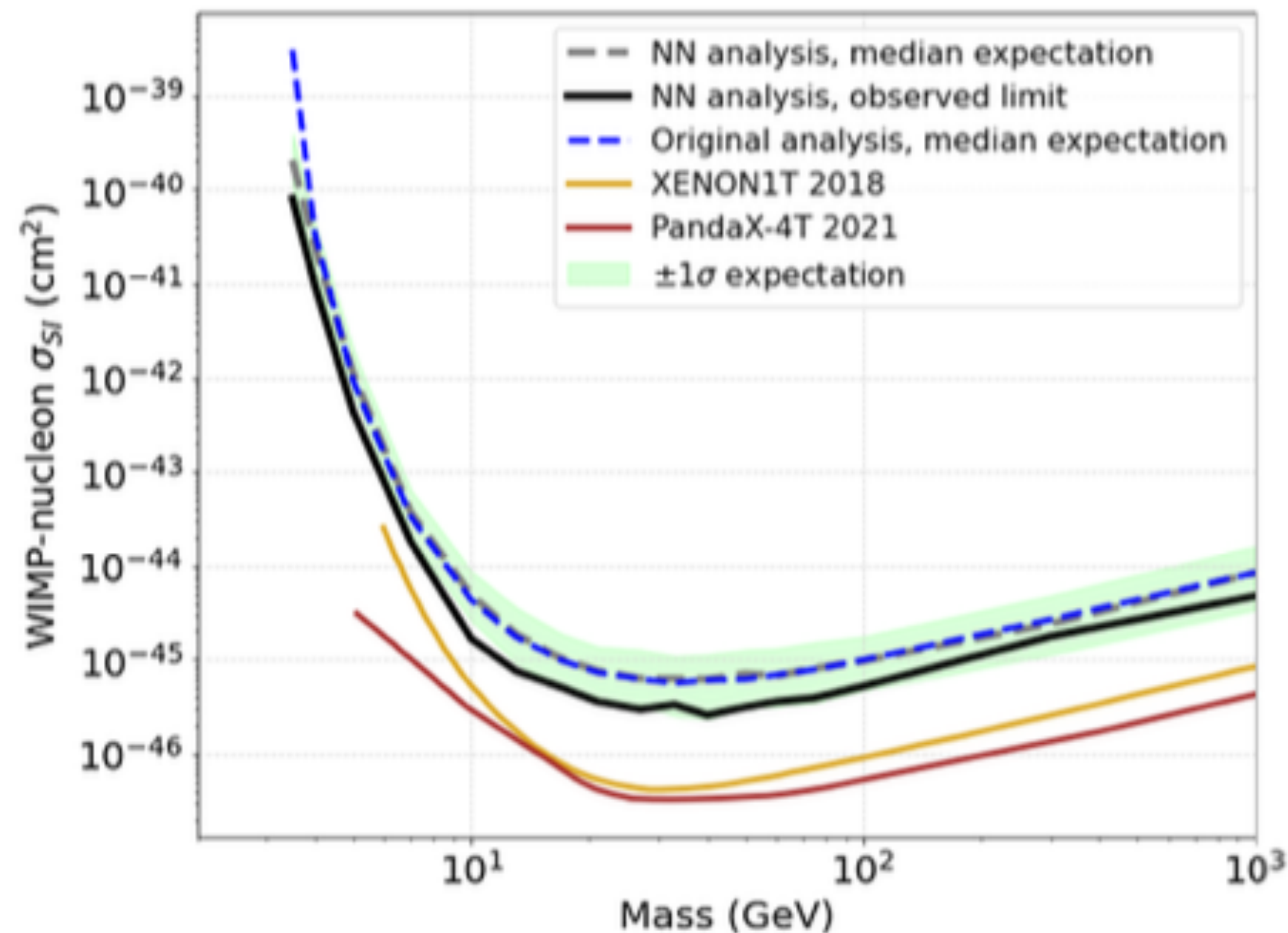
Fast and flexible analysis of direct dark matter search data with machine learning

- Combining ML with profile likelihood fit procedure

- ML is more efficient at capturing information to create models in the form of PDFs

- Faster computation time: ML collapses multiple variables for signal/background discrimination into only one 1D discriminant, taking much of the work out of generating PDFs

- Evaluation of systematic uncertainties through the variation of nuisance parameters in the fit is made more feasible due to the faster calculation and lower statistics requirements

**A. Higuera, MODE workshop, Princeton University, July 2023**

Inferences in DM experiments
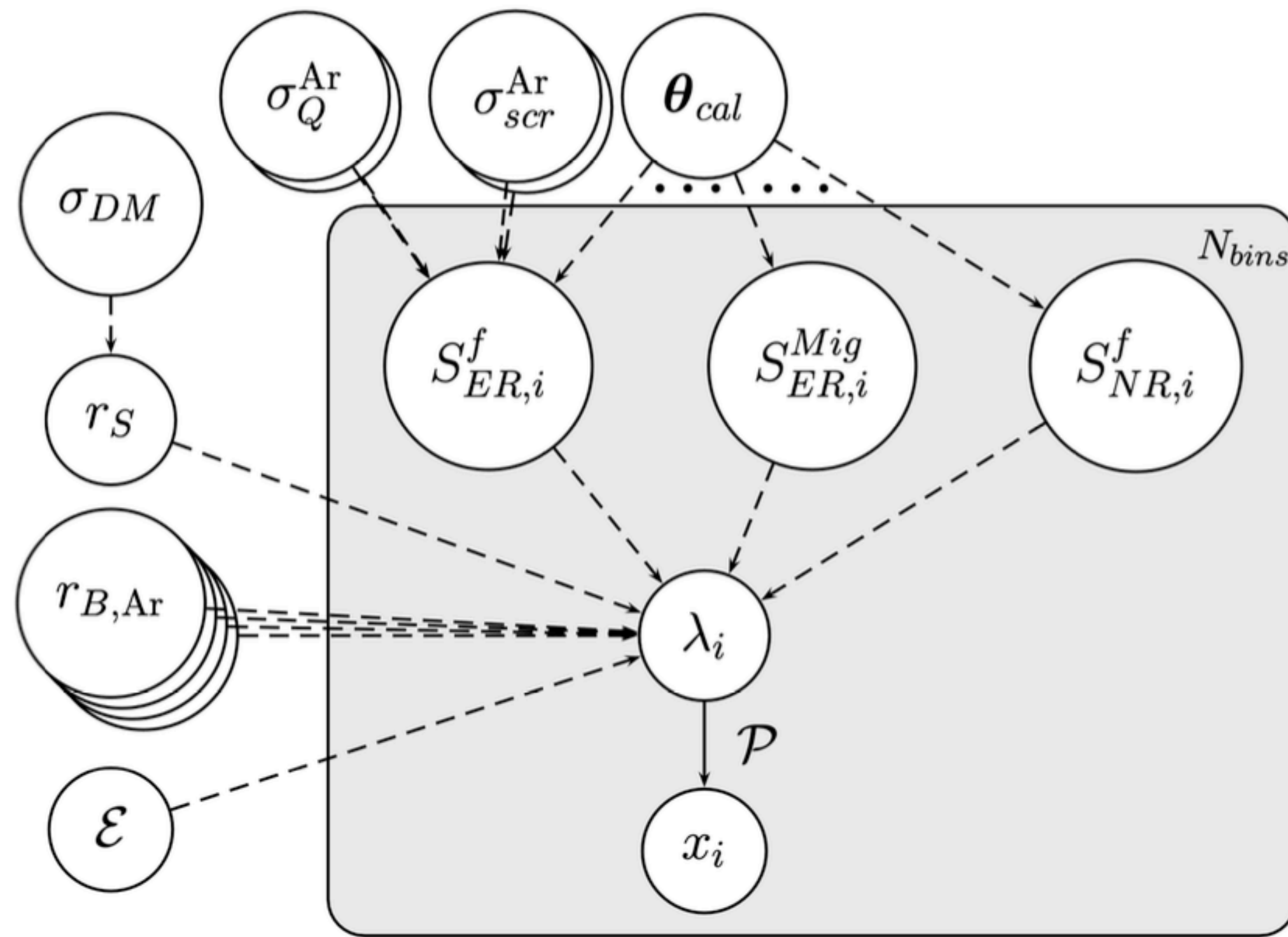
Search for low mass dark matter in DarkSide-50: the bayesian network approach

- The method incorporates the detector response model into the likelihood function

- No assumptions about the linearity of the problem or the shape of the probability distribution functions are required

- This method keeps the dependence of the parameter of interest on the response model parameters, allowing also a possible constraint on the response model

A. Higuera, MODE workshop, Princeton University, July 2023

Discussion
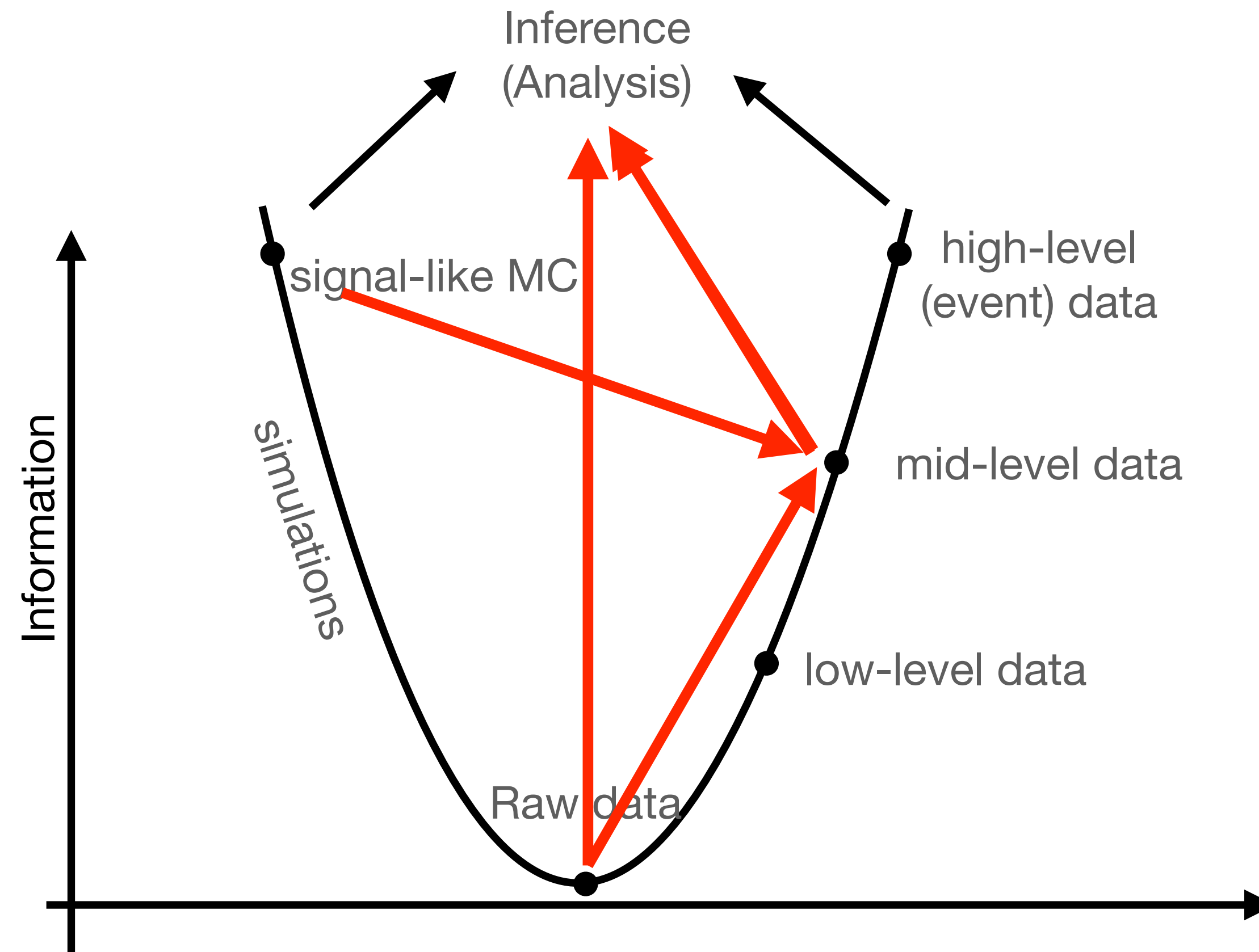
- There is a strong science case to search for dark matter

- The DM community is pushing the boundaries(low-background & low-threshold experiments)

- Data analysis are become more complex and ML is the tool to tackle complex problems

- We physicists have domain knowledge and should leverage that when using ML

A. Higuera, MODE workshop, Princeton University, July 2023

Extas

Neutrinos

- Neutrinos will become a dominant "background"