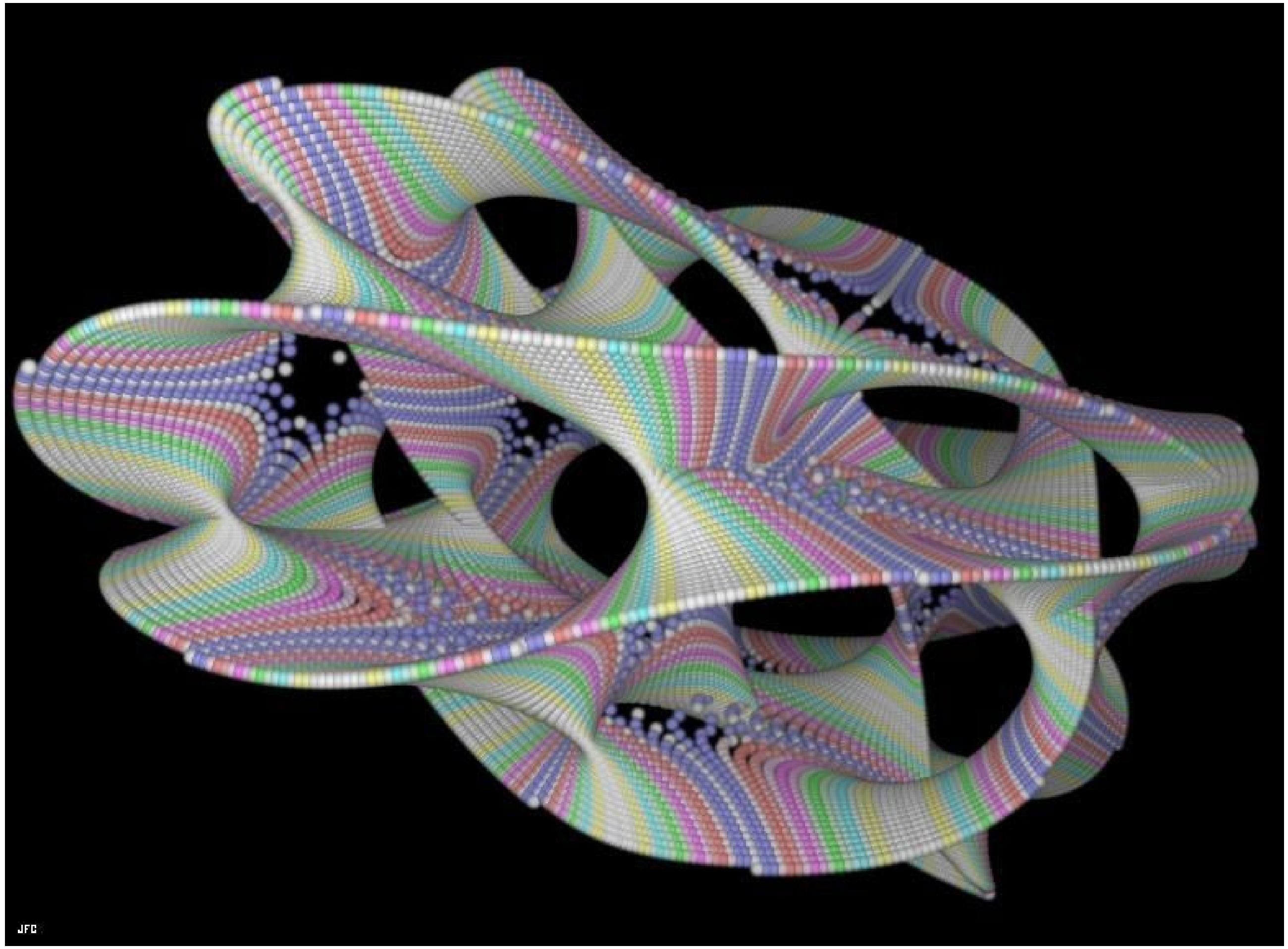
Euclideanized Signals
Facilitating pheno-focused model exploration



Together with Thomas Edwards (GRAPPA) & GAMBIT collaboration

Lorentz Center
ML & DM Workshop
15^{th} Jan 2018, Leiden

GRAPPA
GRavitation AstroParticle Physics Amsterdam

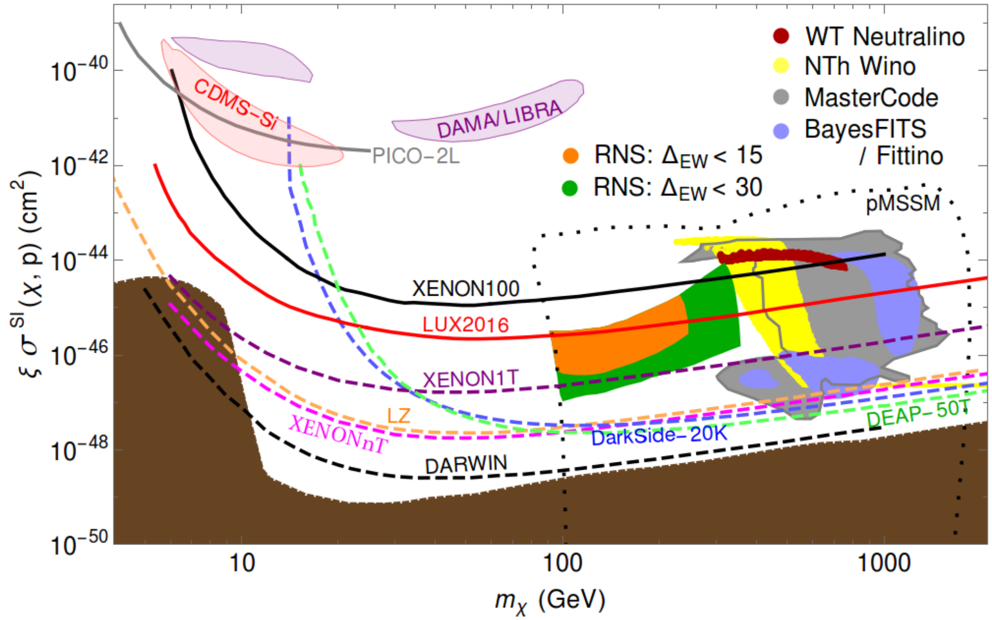
Christoph Weniger
University of Amsterdam

Motivation

Q: How to study and optimize the sensitivity of future instruments for your favourite dark matter model?

Standard approach

- **Upper limits**: Estimate what part of the parameter space can be killed.



Baer+ 2016

- **Confidence contours**: How well can benchmark points be reconstructed?

More interesting but hard to address

- Can one discriminate model A, B, C, …, Z? Where do models overlap?
- Where do additional experiments break model parameter degeneracies?
- What are the distinct phenomenological features of a model?

Fisher information

Log-likelihood ratio quantifies difference between parameter points

$$\mathrm{TS} = -2 \ln \frac{\mathcal{L}(\vec{\theta}_2 | \mathcal{D}(\vec{\theta}_1))}{\mathcal{L}(\vec{\theta}_1 | \mathcal{D}(\vec{\theta}_1))}$$

(Note: Not always easy to translate into significance level)

Fisher information matrix is the Talyor expansion of this

$$\mathrm{TS} \approx (\vec{\theta}_2 - \vec{\theta}_1)^T \mathcal{I} (\vec{\theta}_2 - \vec{\theta}_1)$$

$$\mathcal{I}_{kl}(\boldsymbol{\eta}) = -\left\langle \frac{\partial^2 \ln \mathcal{L}(\mathcal{D}|\boldsymbol{\eta})}{\partial \eta_k \partial \eta_l} \right\rangle_{\mathcal{D}(\boldsymbol{\eta})}$$

1704.05458 Edwards & CW

Appealing aspects

- describes parameter degeneracies and covariance
- provides a metric on the model parameter space → Information geometry!

Reasons why it is hard to use in practice

- Fisher matrix is often singular, changes rank
- Parameter boundaries not taken into account
- *Only local information* (no comparison between models)

Possible solution: ~Isometric embedding

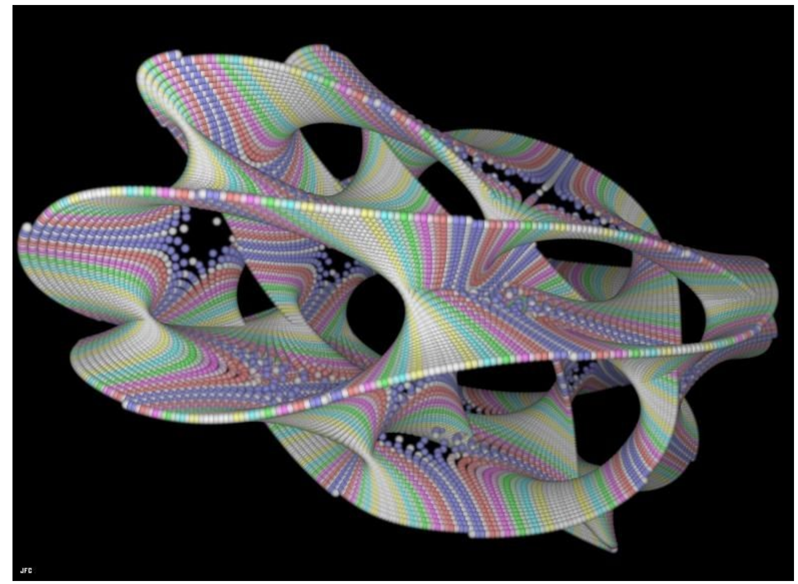
Model parameters

$\vec{\theta} \in \Omega_{\mathcal{P}} \subset \mathbb{R}^d$

Embedding in higher-dimensional space
with unit Fisher information matrix.

$\vec{\theta} \mapsto \vec{x}(\vec{\theta})$

$\vec{x} \in \mathbb{R}^n \qquad \mathcal{I} = \mathbb{1}$



Likelihood ratios --> Euclidean distances

$$\mathrm{TS} = -2 \ln \frac{\mathcal{L}(\vec{\theta}_2 | \mathcal{D}(\vec{\theta}_1))}{\mathcal{L}(\vec{\theta}_1 | \mathcal{D}(\vec{\theta}_1))} \approx \|\vec{x}_1 - \vec{x}_2\|^2$$

Makes problem accessible to many Machine Learning tools:
Dimensionality reduction, Clustering algorithms, Manifold learning, …

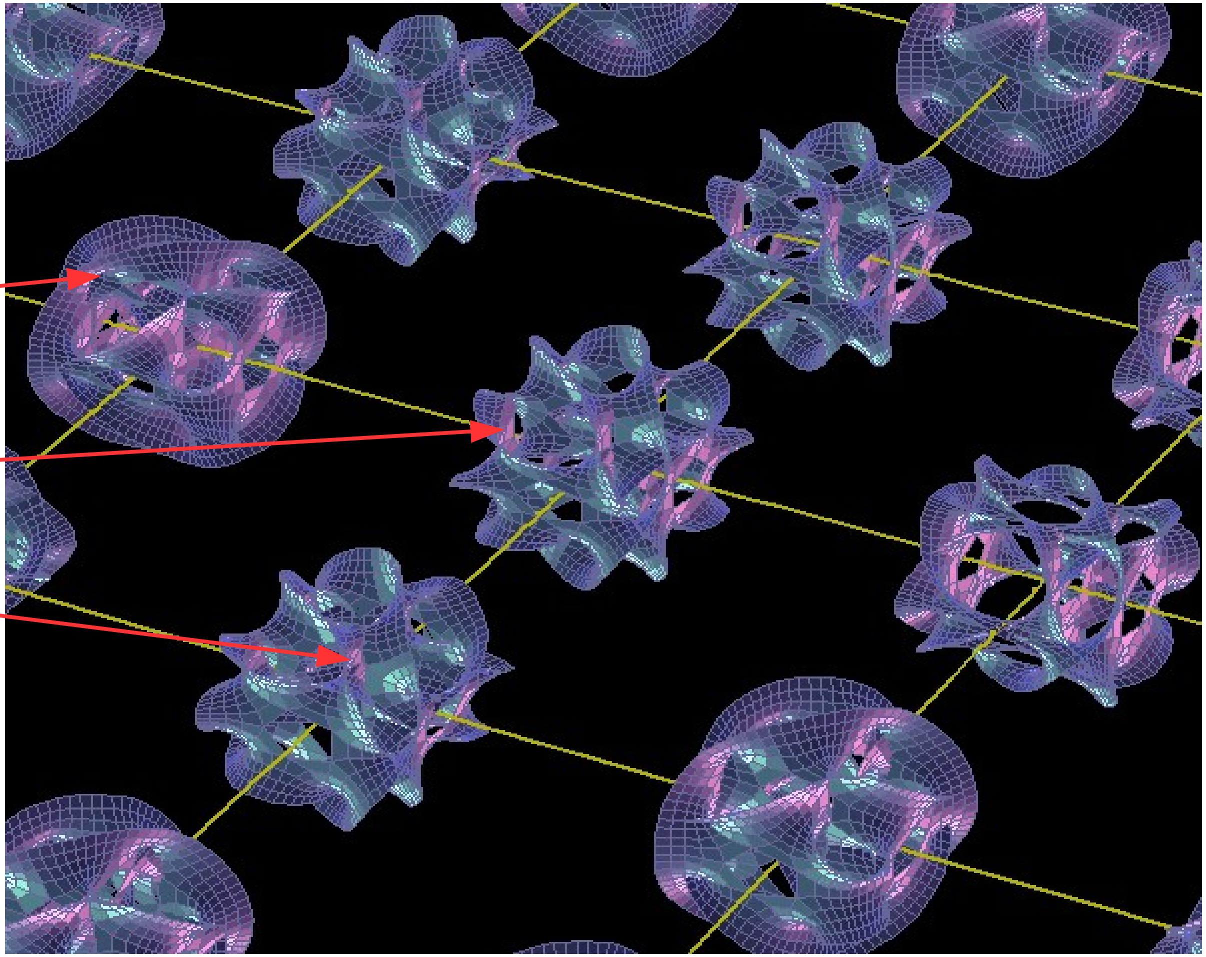
Embedding depends on instrument

Each instrument has its own embedding

CTA

ATLAS

XENON-nT



Combination of instruments is straightfoward

$$\mathrm{TS} \approx \|\vec{x}_1 - \vec{x}_2\|^2 + \|\vec{y}_1 - \vec{y}_2\|^2 + \ldots$$

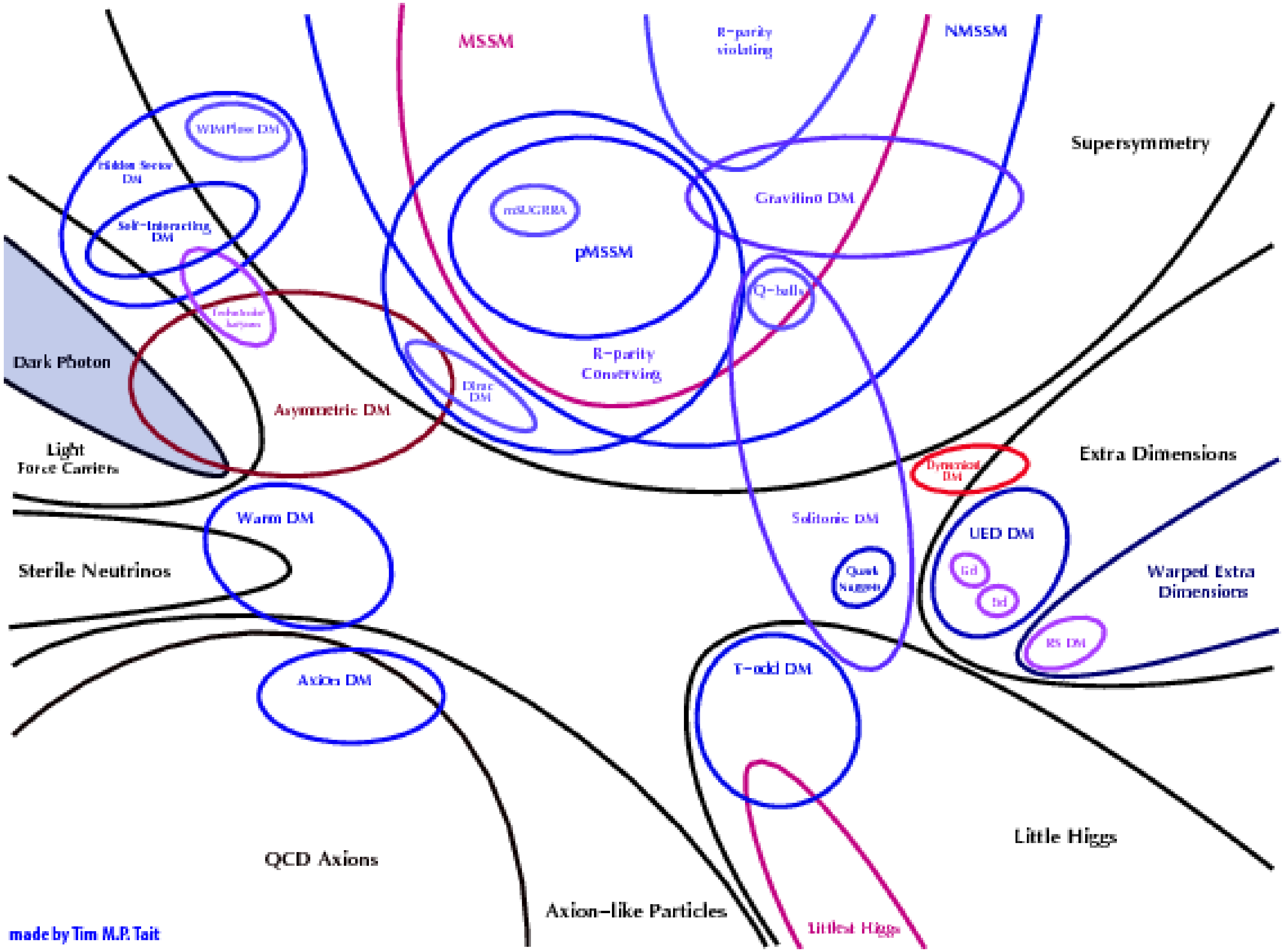
“Volume”: of embedded region corresponds to the number of models that can be discriminated by the experiment

Experimental design: Maximize volume of embedding

Feature extraction: Identify branches of embedded manifold

In the future?

Quantify **Venn diagrams of dark matter models**



How different/similar are models from the perspective of actual observations?

An example implementation

The forecasting pipeline is build around the statistical model implemented in *swordfish*. This is a Poisson process with Gaussian uncertainties (aka Cox-process).

$$\ln \mathcal{L}_\mathrm{p}(\mathcal{D}|\mathbf{S}) = \max_{\delta\mathbf{B}} \left(\underbrace{\sum_{i=1}^{n_\mathrm{b}} \left(d_i \cdot \ln \mu_i(\mathbf{S}, \boldsymbol{\delta}\mathbf{B}) - \mu_i(\mathbf{S}, \boldsymbol{\delta}\mathbf{B})\right)}_{\text{Poisson likelihood}} - \underbrace{\frac{1}{2} \sum_{i,j=1}^{n_\mathrm{b}} \delta B_i \left(K^{-1}\right)_{ij} \delta B_j}_{\text{Bkg covariance}} \right)$$

$$\mu_i(\mathbf{S}, \boldsymbol{\delta}\mathbf{B}) = \Big(\underbrace{S_i + B_i}_{\text{Signal + background}} + \underbrace{\delta B_i}_{\text{Bkg perturbations}} \Big) \cdot \underbrace{E_i}_{\text{Exposure}}$$

n_b : Dimensionality of measurement

Covers: Indirect, direct & collider searches, various cosmology observables, …

Motivation of embedding equations

Starting point: **Fisher information matrix**

$$\mathcal{I}_{lk}(\boldsymbol{\theta}) = \sum_{ij} \frac{\partial S_i}{\partial \theta_k} D^{-1}_{ij} \frac{\partial S_j}{\partial \theta_l} \quad \text{with} \quad D_{ij} = K_{ij} + \delta_{ij} \frac{S_i(\boldsymbol{\theta}) + B_i}{E_i}$$

Noise + bkg covariance

$\vec{\theta} \in \mathbb{R}^d$ $\quad \mathcal{I}: \ (d \times d)$ matrix $\quad D: \ (n_b \times n_b)$ matrix

This motivates the **embedding equation**

$$x_i \equiv \left(\sum_j (D^{-1/2})_{ij} S_j E_j \right) \left(1 + \frac{R \cdot S_i}{R \cdot S_i + B_i + K_{ii} E_i} \right)$$

$\vec{x} \in \mathbb{R}^{n_b}$

Fudge factor for signal limited regime, $R = 0.1$

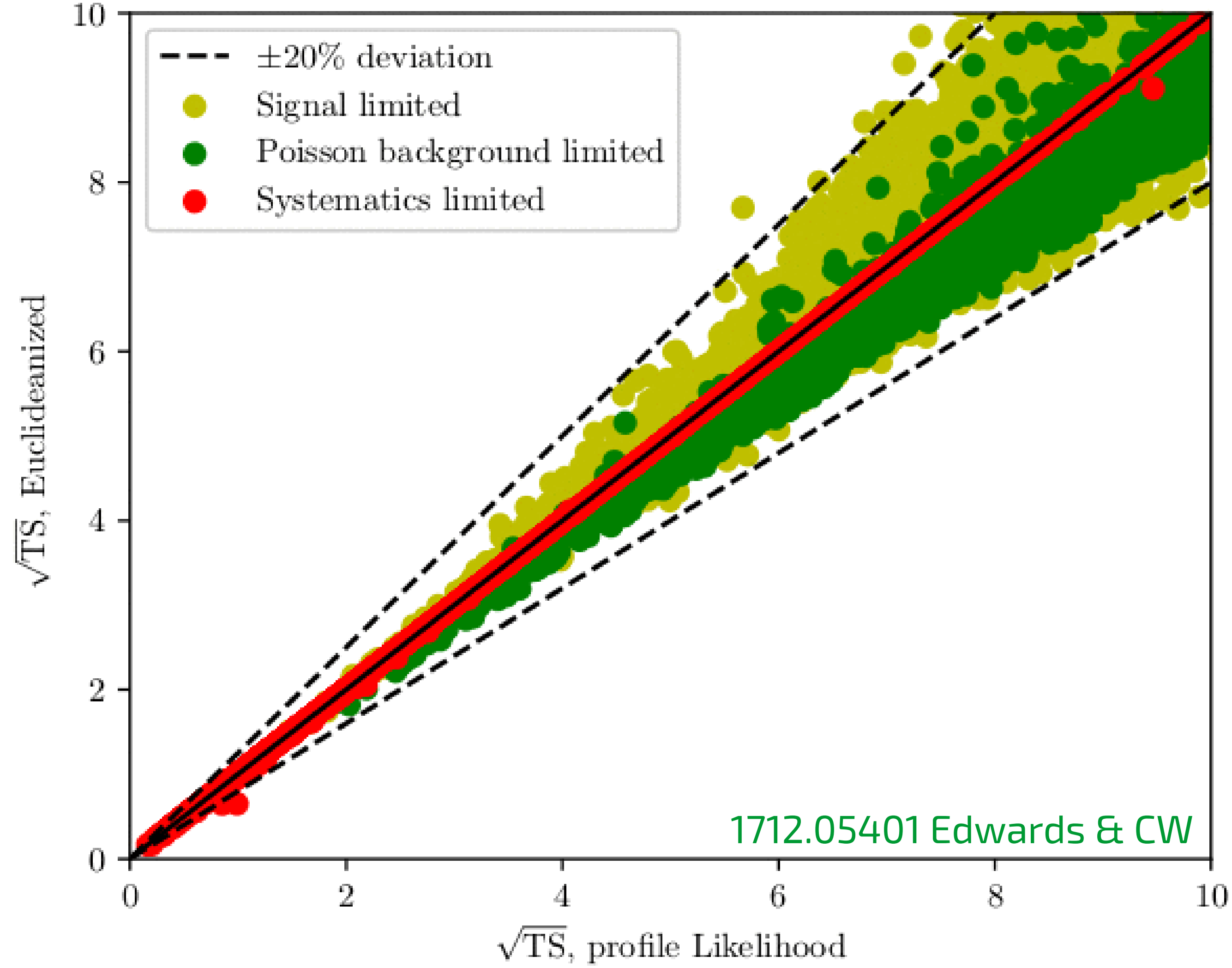
1712.05401 Edwards & CW

Then:

$$\mathrm{TS} = -2 \ln \frac{\mathcal{L}(\vec{\theta}_2 | \mathcal{D}(\vec{\theta}_1))}{\mathcal{L}(\vec{\theta}_1 | \mathcal{D}(\vec{\theta}_1))} \approx \|\vec{x}_1 - \vec{x}_2\|^2$$

Comparison of exact and approx TS

Comparison of exact (profile likelihood) and approximate (euclideanized signal) TS values, for randomly generated models.

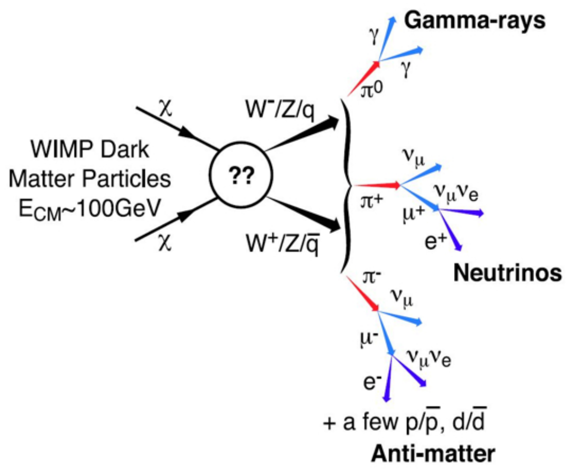


Agreement within 20%, for signal-limited, Poisson background limited and systematics limited regions.

A first simple application

Problem

- Millions of samples from singlet DM parameter space
- Task: Identify expected 68.3% and 95.4%CL reconstruction region for future CTA (here only toy instr.)
- Naively, this requires the calculation of a million likelihood ratios
- Has to be repeated for each benchmark point of interest → Impossible to explore all aspects of model

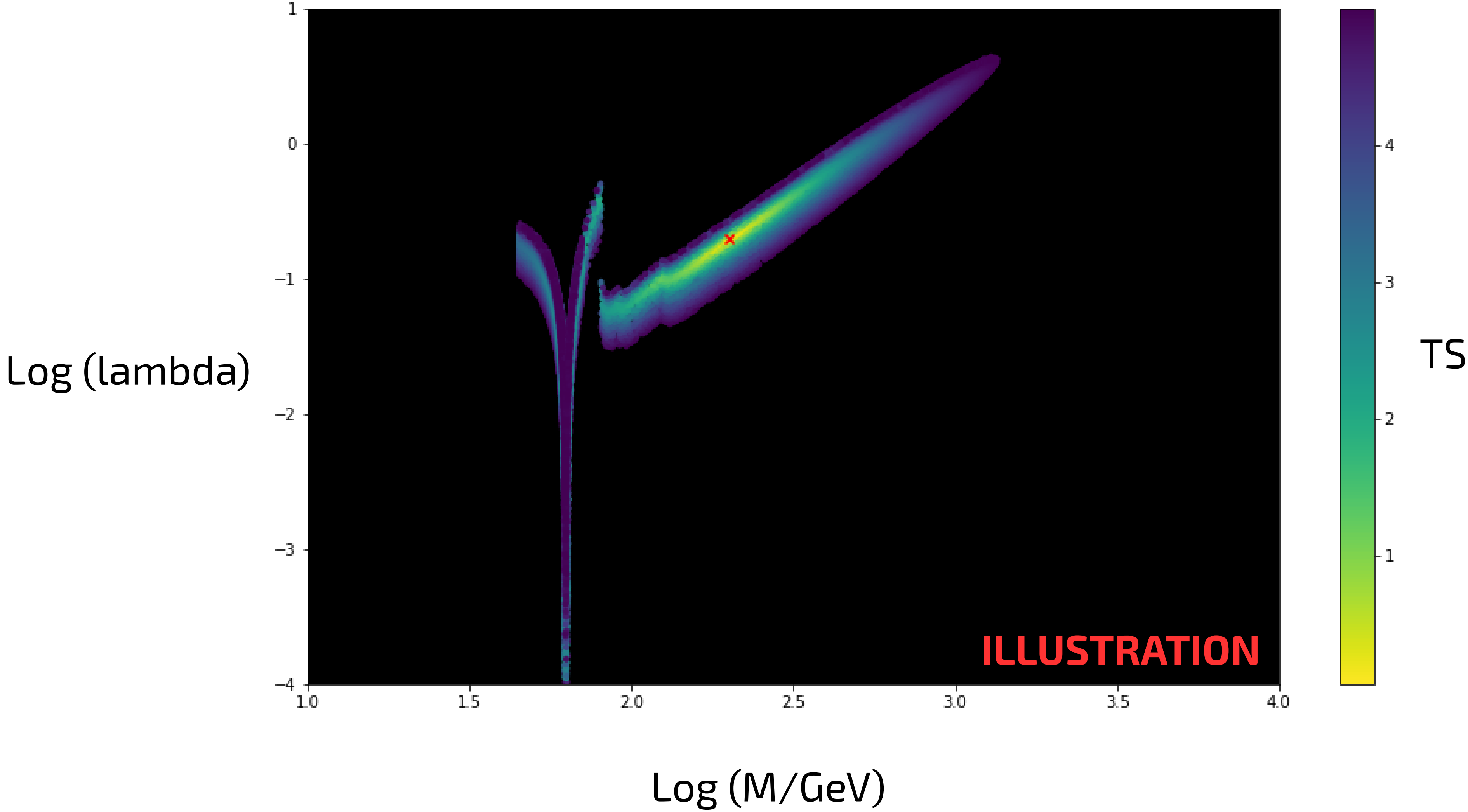


Approach

- Map singlet DM parameter space onto Euclidenaized signals (swordfish*)
- Use nearest neighbour search (sklearn BallTree) to identify closest projected points
- Project back onto model parameter space
- Make plot

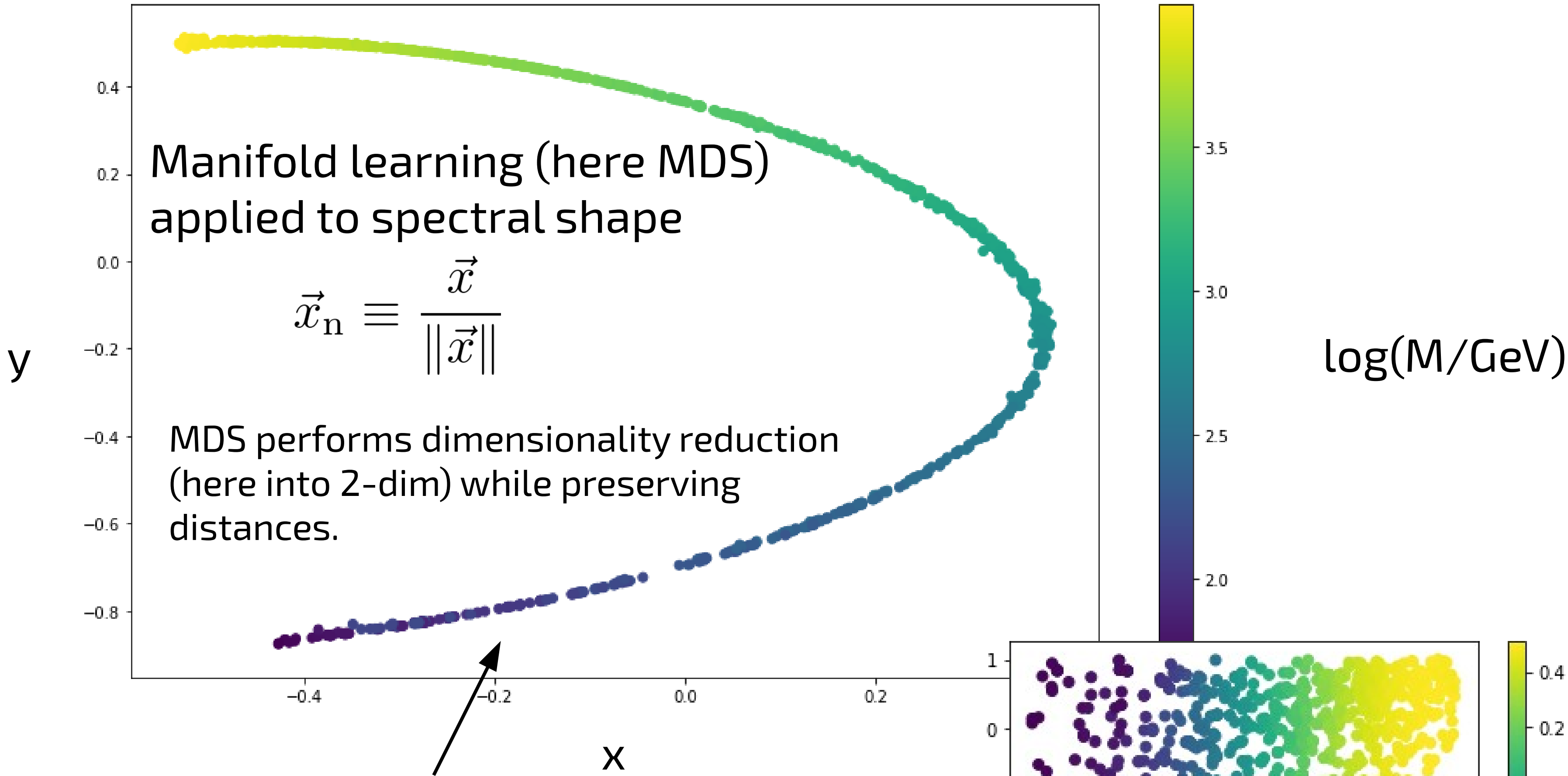
A first simple application

Resulting confidence region around benchmark point (red cross)



Based on the chains from GAMBIT, Singlet DM, 2017

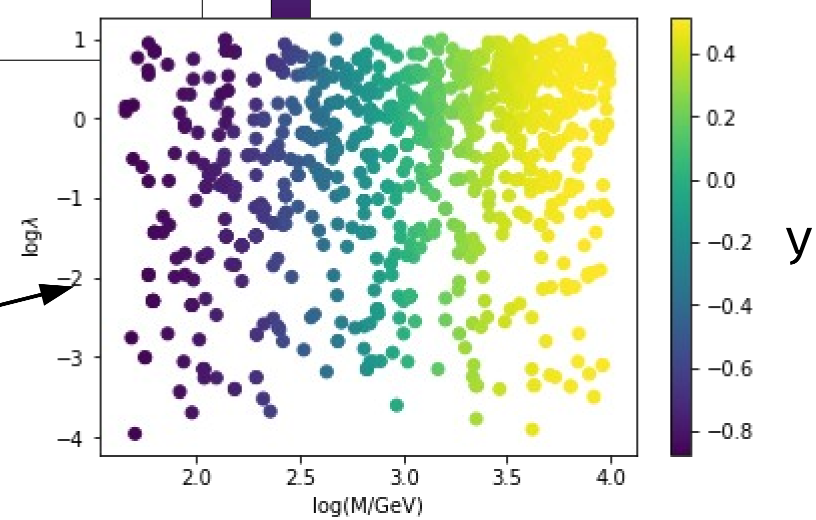
Automatic feature classification?



Manifold learning (here MDS) applied to spectral shape

$$\vec{x}_{\rm n} \equiv \frac{\vec{x}}{\|\vec{x}\|}$$

MDS performs dimensionality reduction (here into 2-dim) while preserving distances.



MDS figures out that there is only one relevant parameter (mass) that determines the spectral shape.

Conclusions

- Fisher forecasting is useful, but has serious limitations
- I propose an (approximate) isometric embedding
  Model parameters → high-dim. space with trivial Fisher metric
  which removes many of the limitations

- There is a **analytic version of such an embedding** for the Cox-process
- Example: Fast calculation of reconstruction contours for CTA

- Embedded manifold (aka Euclideanized signals) provides **starting point for applying Machine Learning tools** to study model phenomenology

- **Possible applications**
  - Derivation of detection thresholds / limits / confidence regions
  - **Automatized model comparison** on the phenomenological level, for *many* models
  - Identify **unique signatures** of DM models
  - Starting point for **dimensionalty reduction** (manifold learning) to identify characteristic observable features of method
  - Starting point for **experimental design**