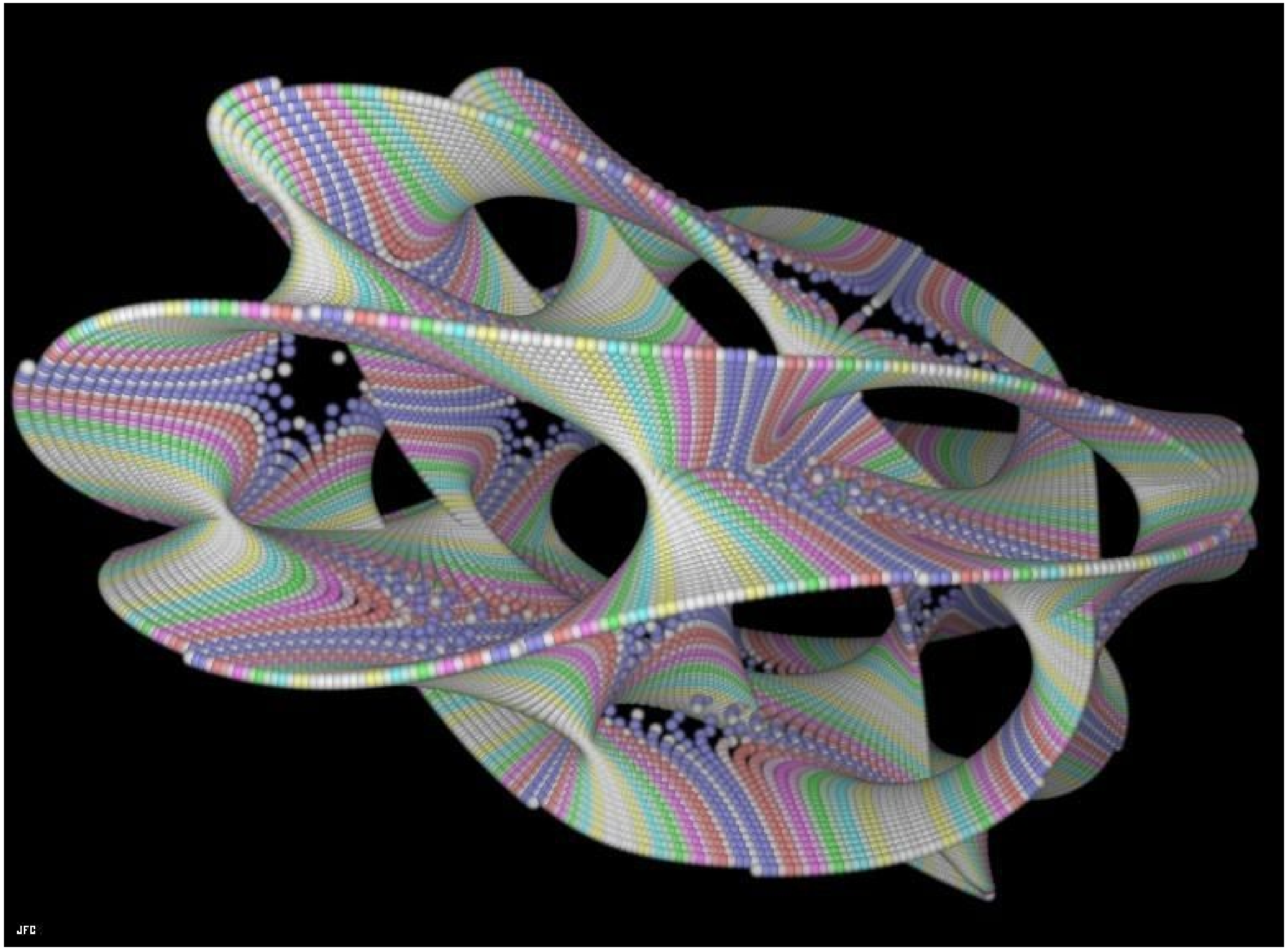
Euclideanized Signals
Facilitating pheno-focused model exploration



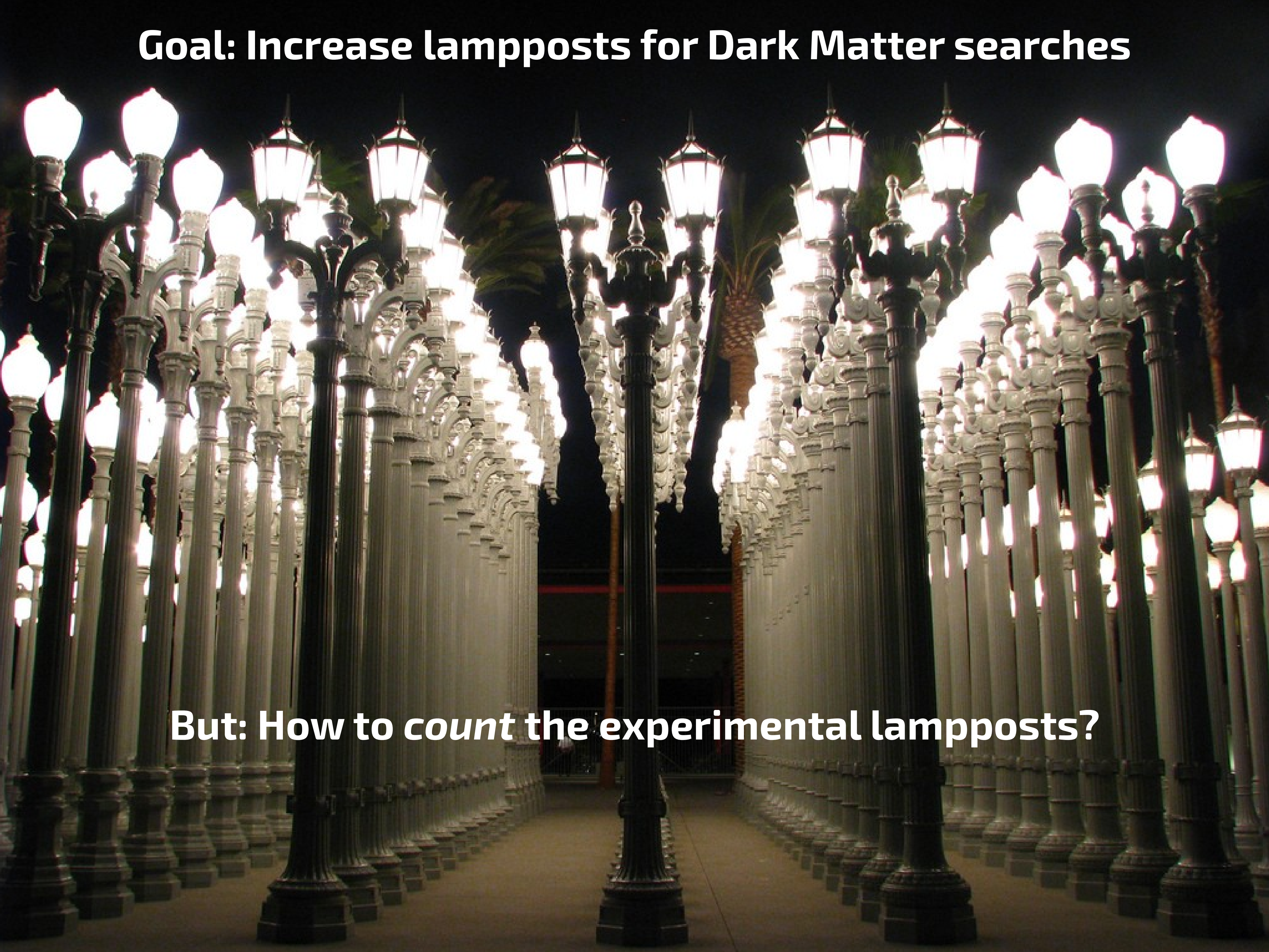
Edwards & Weniger, 1704.05458, 1712.05401

GRAPPA
GRavitation AstroParticle Physics Amsterdam

DM stats
1st Mar 2018, Banff

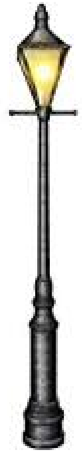
Christoph Weniger
University of Amsterdam

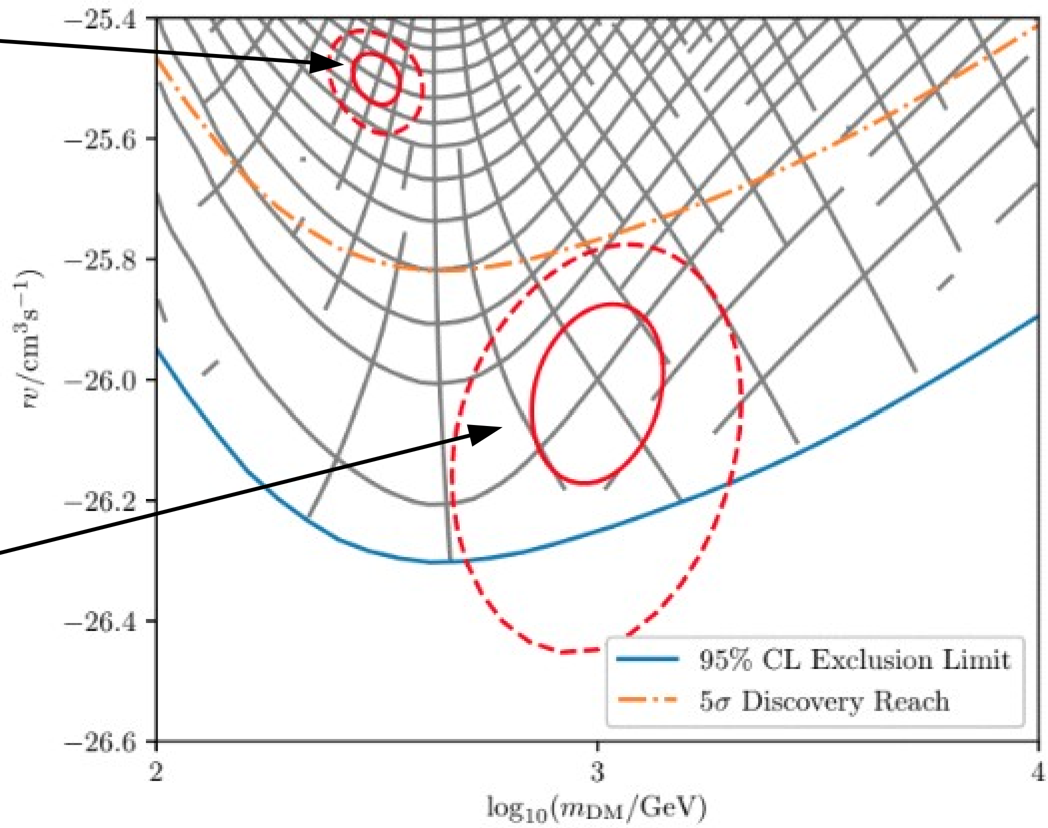
Goal: Increase lampposts for Dark Matter searches



But: How to *count* the experimental lampposts?

Definition of “Number of lampposts”





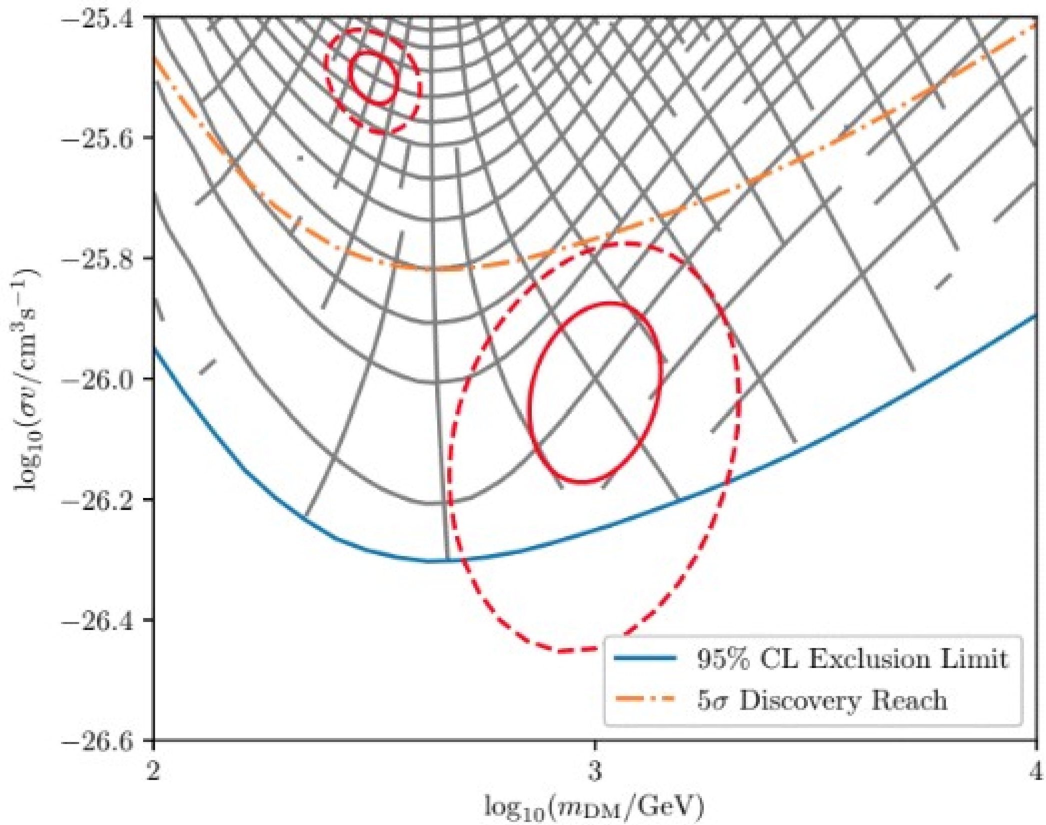
Heuristic definition

- Frequentist: Number of models that can be discriminated at 1 sigma level.
- Bayesian: Jefferys prior $p\left(\vec{\theta}\right) \propto \sqrt{\det \mathcal{I}\left(\vec{\theta}\right)}$

Quantifying Sensitivity of DM Searches

Standard approaches

- Expected upper limits
- Expected discovery reach
- Benchmark point reconstruction



Questions that remain usually unanswered

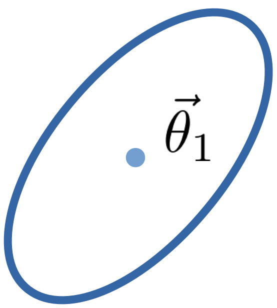
- Can one discriminate model A, B, C, …? Where do models "look the same", where do they differ?
- Would additional experiments break model degeneracies globally?
- What are the *distinct* phenomenological features of a model?

Fisher information

Log-likelihood ratio quantifies difference between parameter points

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

→ Expected confidence regions



Fisher information matrix is the Talyor expansion of this

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

$$\mathcal{I}_{kl} = -\left\langle \frac{\partial^2 \ln \mathcal{L}(\mathcal{D}|\vec{\theta})}{\partial \theta_k \partial \theta_l} \right\rangle_{\mathcal{D}(\vec{\theta})}$$

- describes parameter uncertainties
- provides a *metric* on the model parameter space → *Information geometry!*

Technical challenges

- Fisher matrix is
  - …often singular, changes rank
  - …unaware of parameter boundaries
  - …unaware of non-local model degeneracies
- Need to pair-wise compare compare (often millions of) parameter points

Isometric embedding of model parameter space

Model parameters

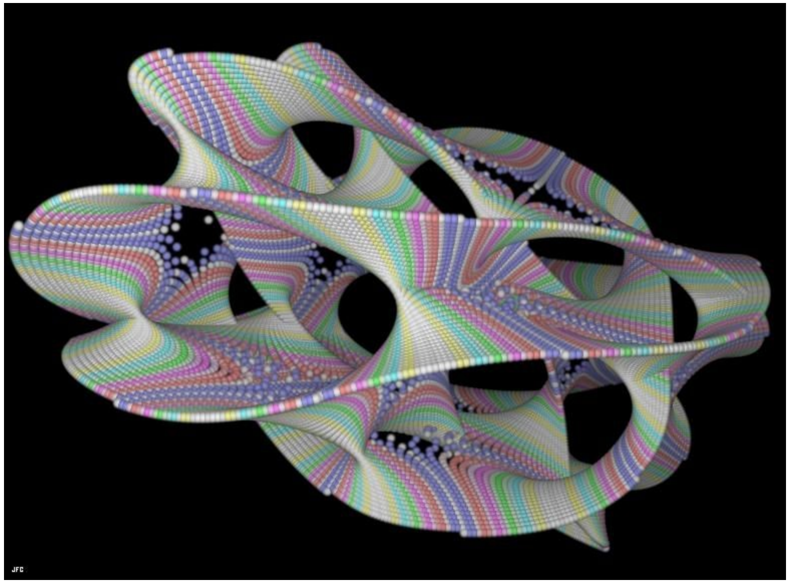
$\vec{\theta} \in \mathbb{R}^d$

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

$\mathcal{I} = \mathbb{1}$

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

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



d: Number of model parameters

n: Number of experimental data bins

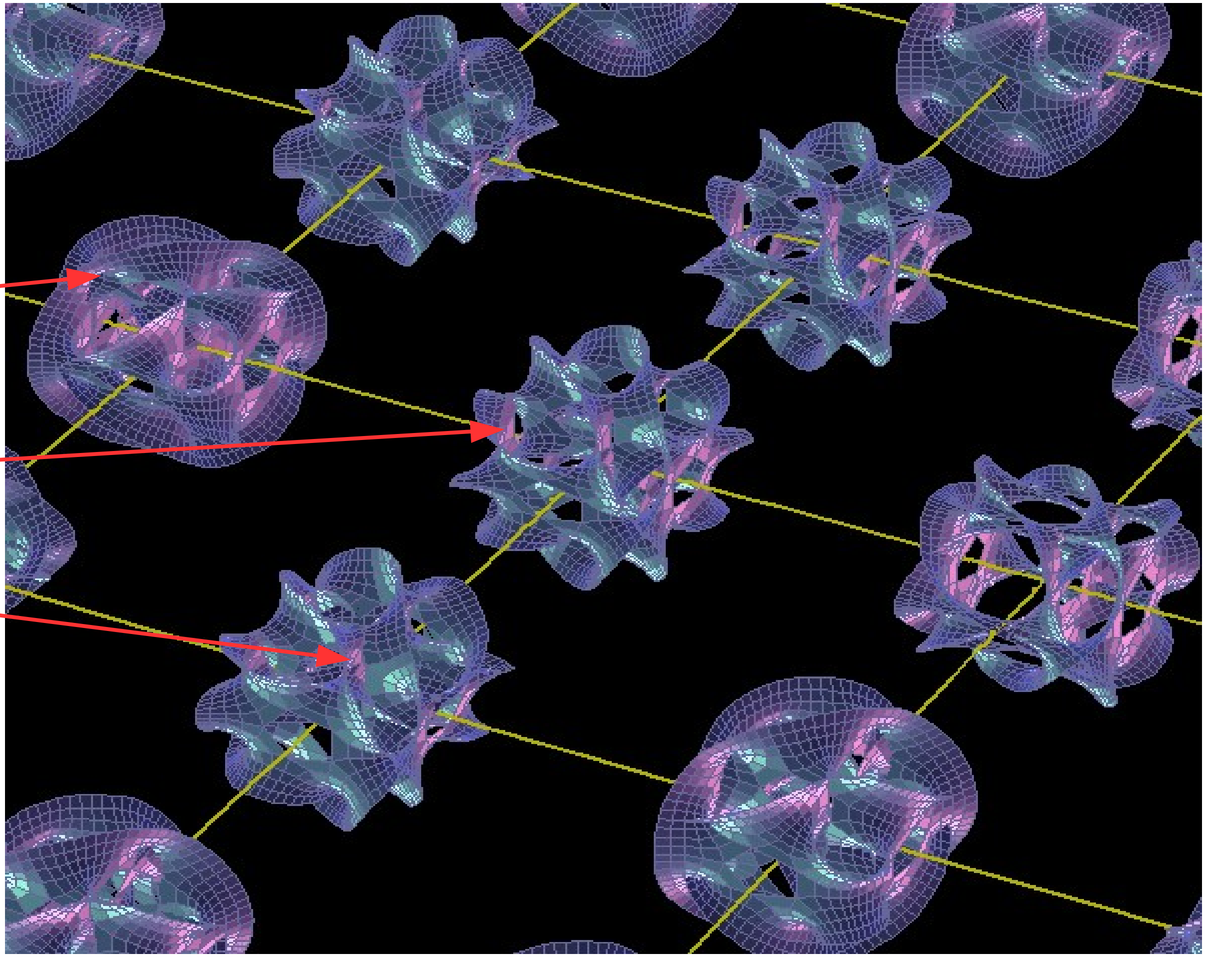
Likelihood ratios → Euclidean distance

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

Number of lampposts = “Volume” of projected hypersurface

Combining instruments

Each instrument has its own embedding



CTA

ATLAS

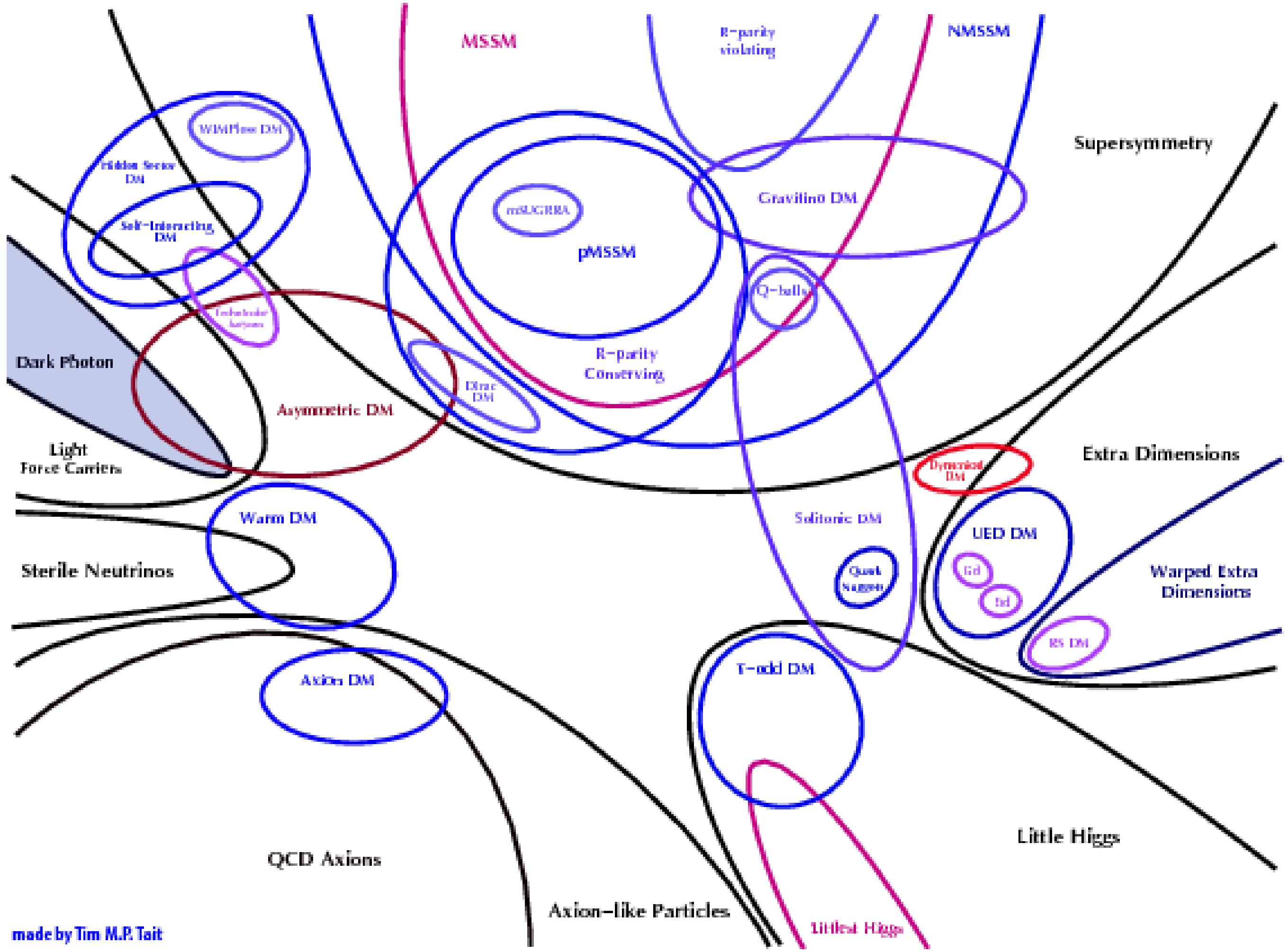
XENON-nT

Combination of instruments is straightfoward

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

Experimental design: Maximize volume of embedding

***Quantify* Venn diagrams of dark matter models**



Volumina: Number of detectable models

Overlaps: Numer of detections that would be compatible with both models

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).

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

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

n_{b} : Dimensionality of measurement

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

Motivation of embedding equations

Starting point: **Fisher information matrix**

$$\mathcal{I}_{lk}(\boldsymbol{\theta}) = \sum_{ij} \frac{\partial S_i}{\partial \theta_k} D^{-1}_{ij} \frac{\partial S_j}{\partial \theta_l}$$

with

$$D_{ij} = K_{ij} + \delta_{ij} \frac{S_i(\boldsymbol{\theta}) + B_i}{E_i}$$

Noise + bkg covariance

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

This motivates the **embedding equation**

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

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

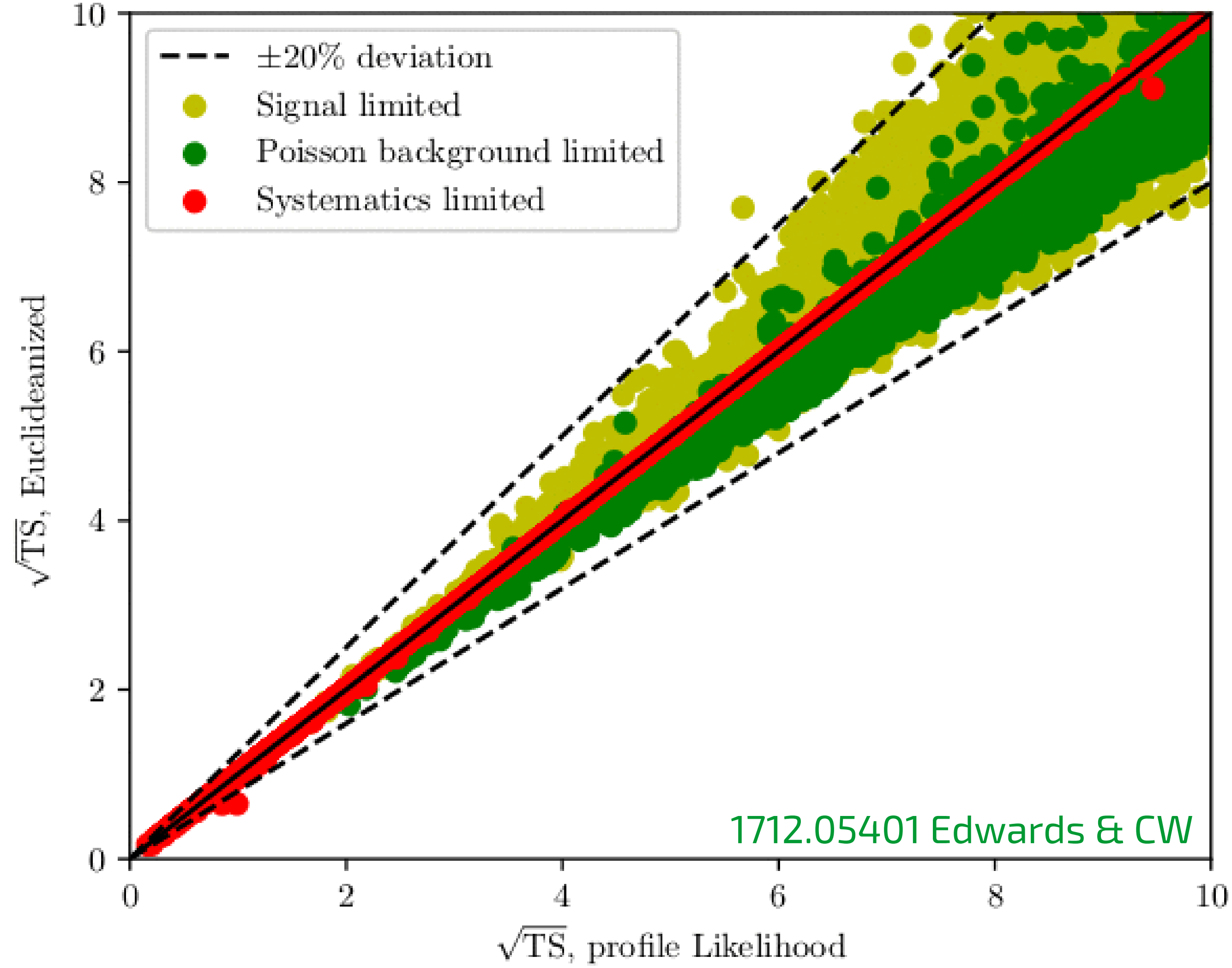
Fudge factor for signal limited regime, $R = 0.1$

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

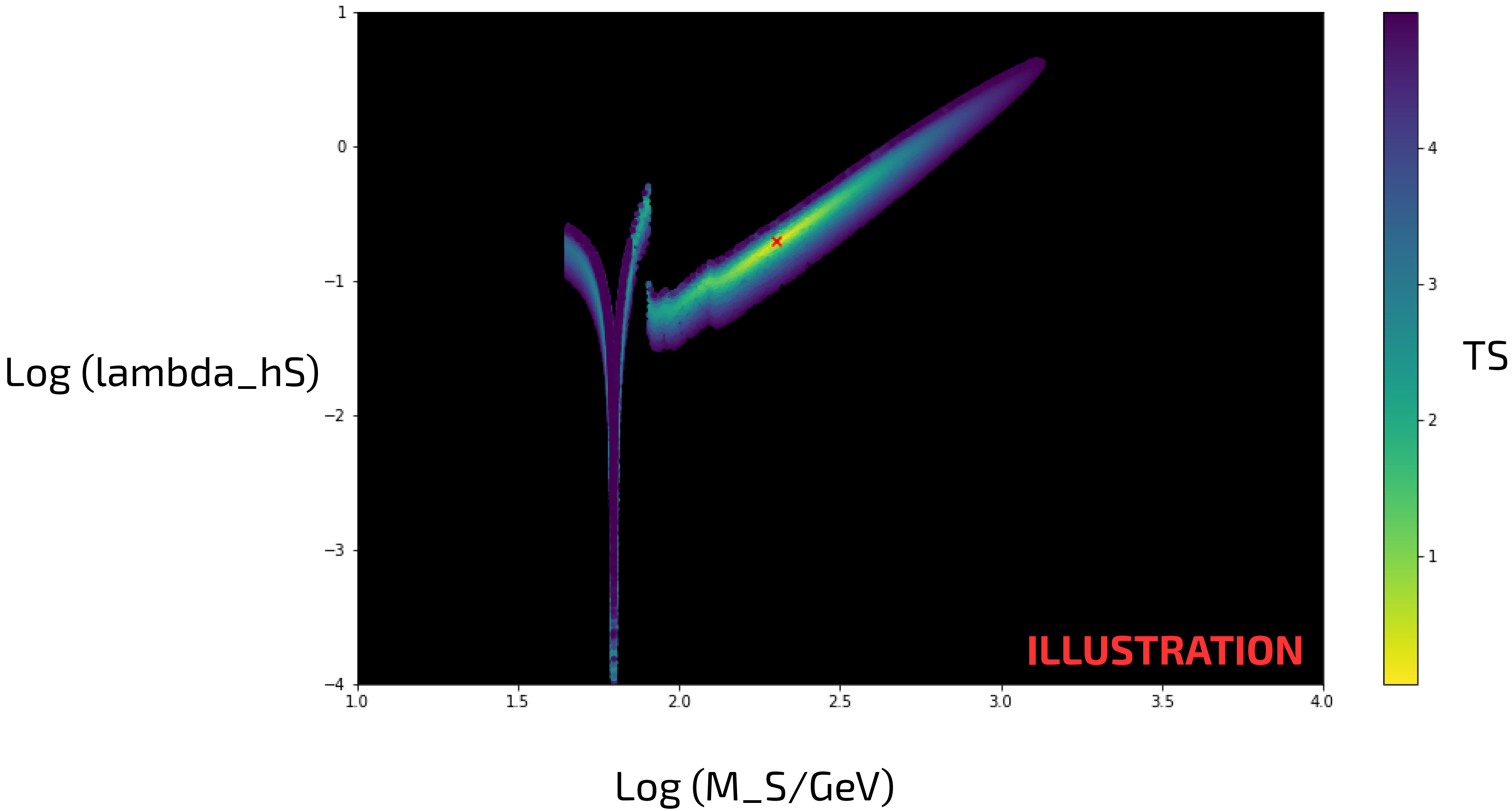
Comparison of exact (profile likelihood) and approximate (euclideanized signal) TS values, for randomly generated models.



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

A simple example: Singlet DM

Expected confidence region around benchmark point (red cross)
(assuming some toy indirect search likelihood)



Based on the chains from GAMBIT, Singlet DM, 2017

CTA example (illustration)

Scenario

- DM annihilation

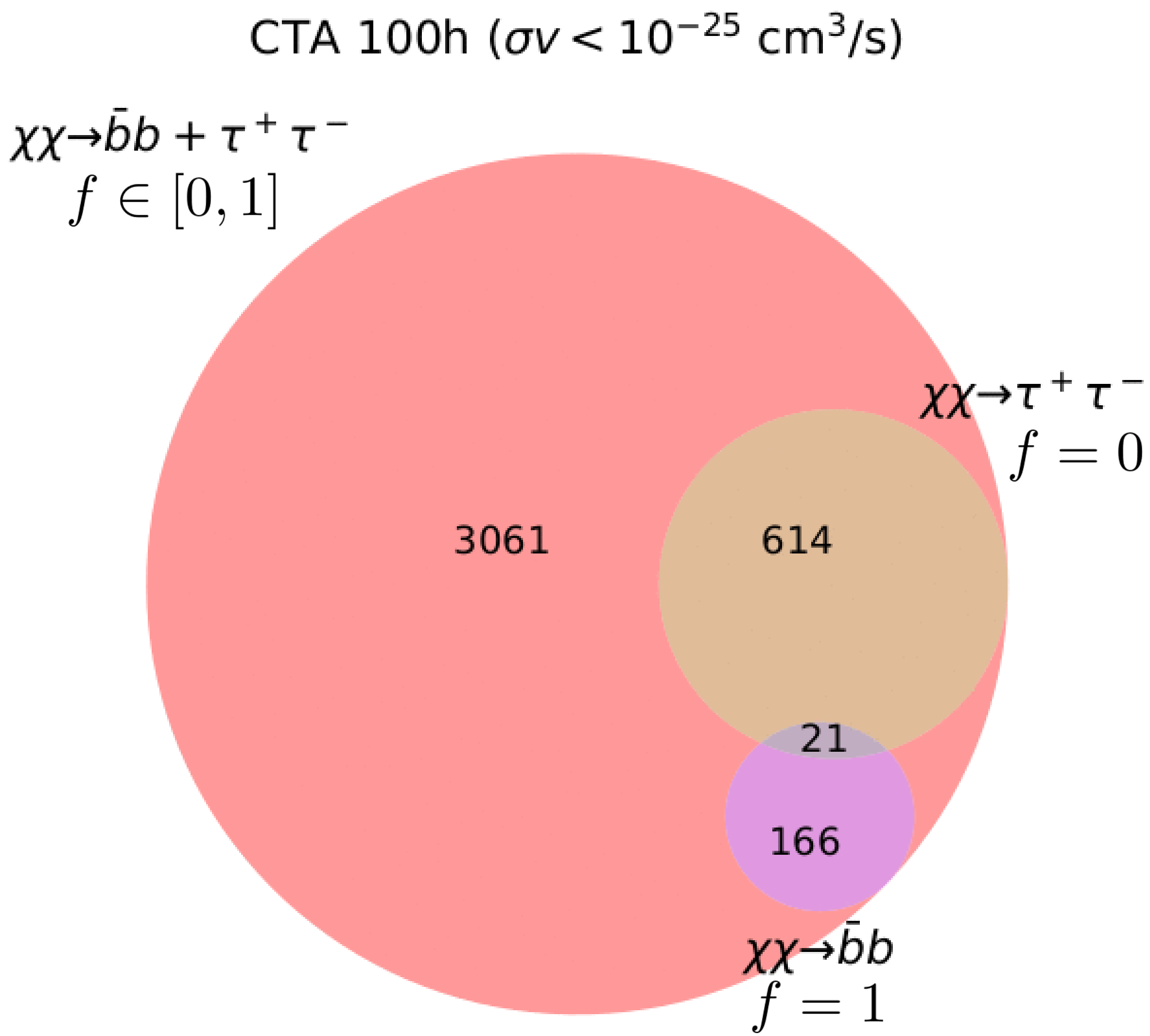
  $$\mathrm{BR}(\chi\chi \to \bar{b}b) = f$$
  $$\mathrm{BR}(\chi\chi \to \tau^+\tau^-) = 1 - f$$

- Model parameters

  $$\langle \sigma v \rangle, m_\chi, f$$
  $$\langle \sigma v \rangle \leq 10^{-25} \mathrm{cm}^3 \mathrm{s}^{-1}$$

- CTA likelihood
  100h GC observations

Venn diagram for possible CTA DM detections



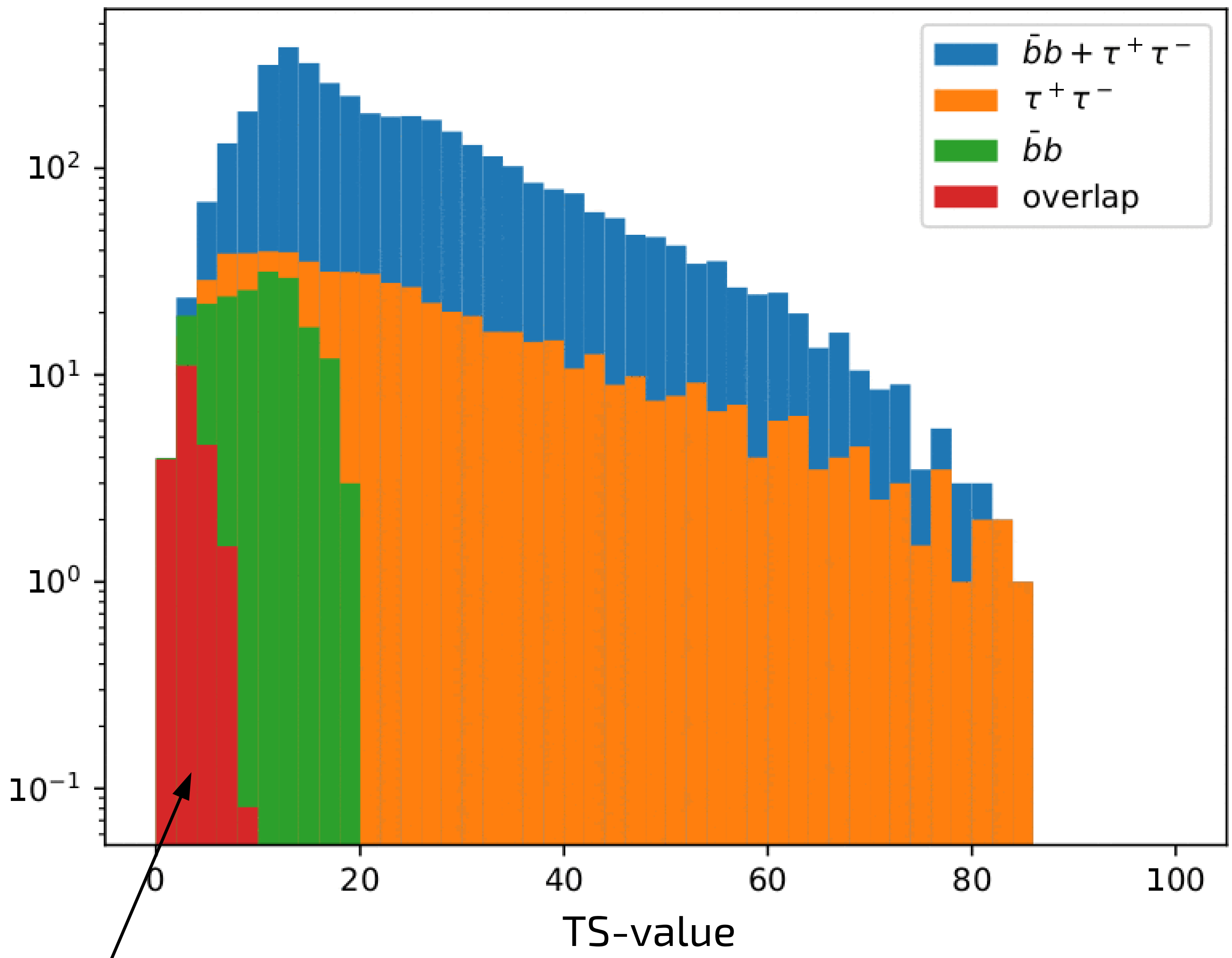
→ 3862 lampposts

Estimate number of 1-sigma regions

$$N_{\text{regions}} = \sum_{i=1}^{n_{\text{points}}} \frac{1}{N(\vec{\theta}_i)}$$

$N(\vec{\theta})$: Number of nearest neigbours within unit ball.
(needs to be corrected for effective dimensionality, filling factor)

CTA example (illustration)



Only low-significant signals contribute to overlap

Outlook

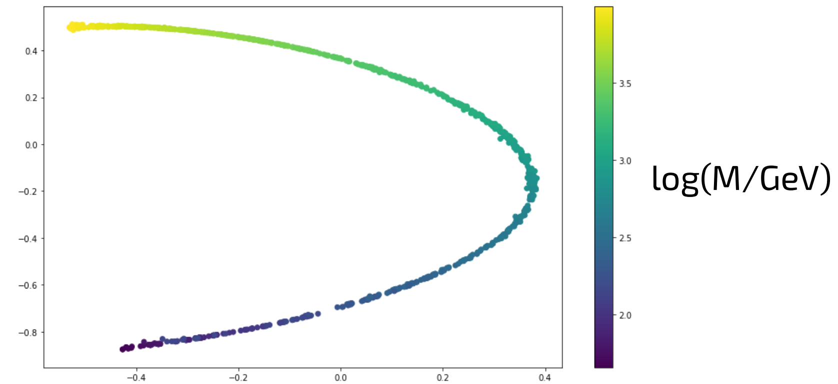
What else?

- "Lampposts" define **minimal search grid for new physics searches** (information geometry already used in, e.g., pulsar searches)

- Dimensionality of euclideanized signal manifold is proxy for effective degrees of freedom of model
  → Frequentist **p-values**

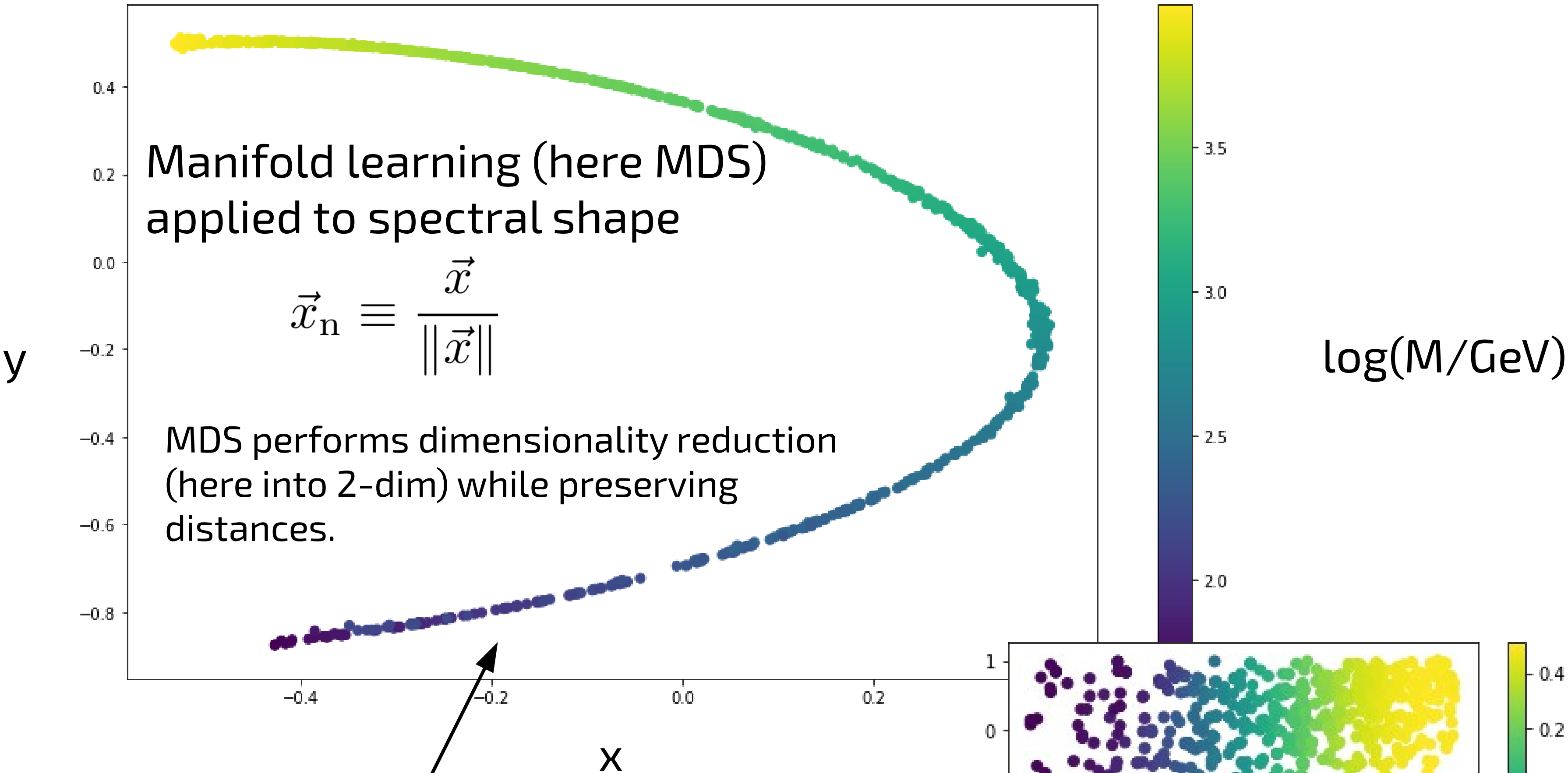
- Euclideanized signals can be used to estimate for Fisher information matrix in original model parameter space
  → Can be used to **optimize parameter scans**? (e.g. optimal kernal for distributed Gaussian processes)

- Euclideanized signals can be used as starting point for dimensionality reduction (manifold learning)
  → Automatic feature extraction




Thank you!

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.