TDA - fall 2025

Stats + ML
techniques
in persistence

Last time:

A whole mess of statistics.

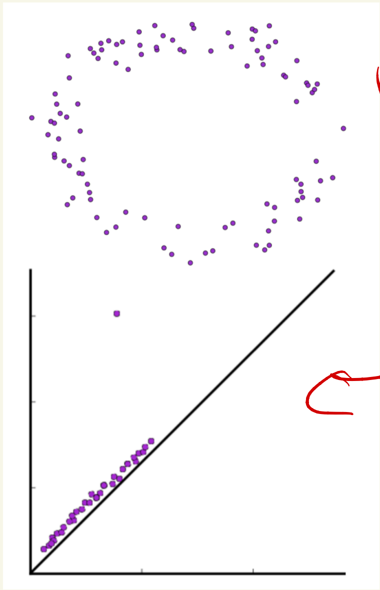
Prime theme:

Given a collection of PDs, how can we:

- Compute averges
- guarantee some amount of statistical significance
- eventually, maybe consider some methods that play well with ML

Changing the question: Fasy et al 2014

What is an estimate for the averge (true?) diagram, & how far off am I?



unknown space

how good?

- Want to estimate PD for a set $M \subseteq \mathbb{R}^d$
- Don't know M
  - ↳ but, have a sample $S_n = \{X_1, \dots, X_n\}$ drawn uniformly from M.
- Persistence diagram for S_n is used as an estimator for X → denoted $\hat{X}$

Confidence intervals

Given a collection of points $X = \{x_1, \dots, x_n\}$ from $\mathbb{R}$, the $100 \cdot (1-\alpha)\%$ confidence interval for the mean μ is the interval $[u(X), v(X)]$ such that

$$P(\mu \in [u(X), v(X)]) = 1-\alpha$$

$\geq ?$

Equivalently: find c & an estimate for μ called $\hat{\mu}$ s.t.

$$P(\|\mu - \hat{\mu}\| \geq c) = \alpha$$

How to use in persistence?

Fix $\alpha \in (0,1)$

Want $c_n := c_n(X_1, \dots, X_n)$ s.t.

$$\limsup_{n \to \infty} P\left(d_B(\hat{X}, X_n) > c_n\right) \leq \alpha$$

Then, $[0, c_n]$ is an asymptotic $(1-\alpha)$ confidence set for the bottleneck distance $d_B(\hat{X}, X)$.

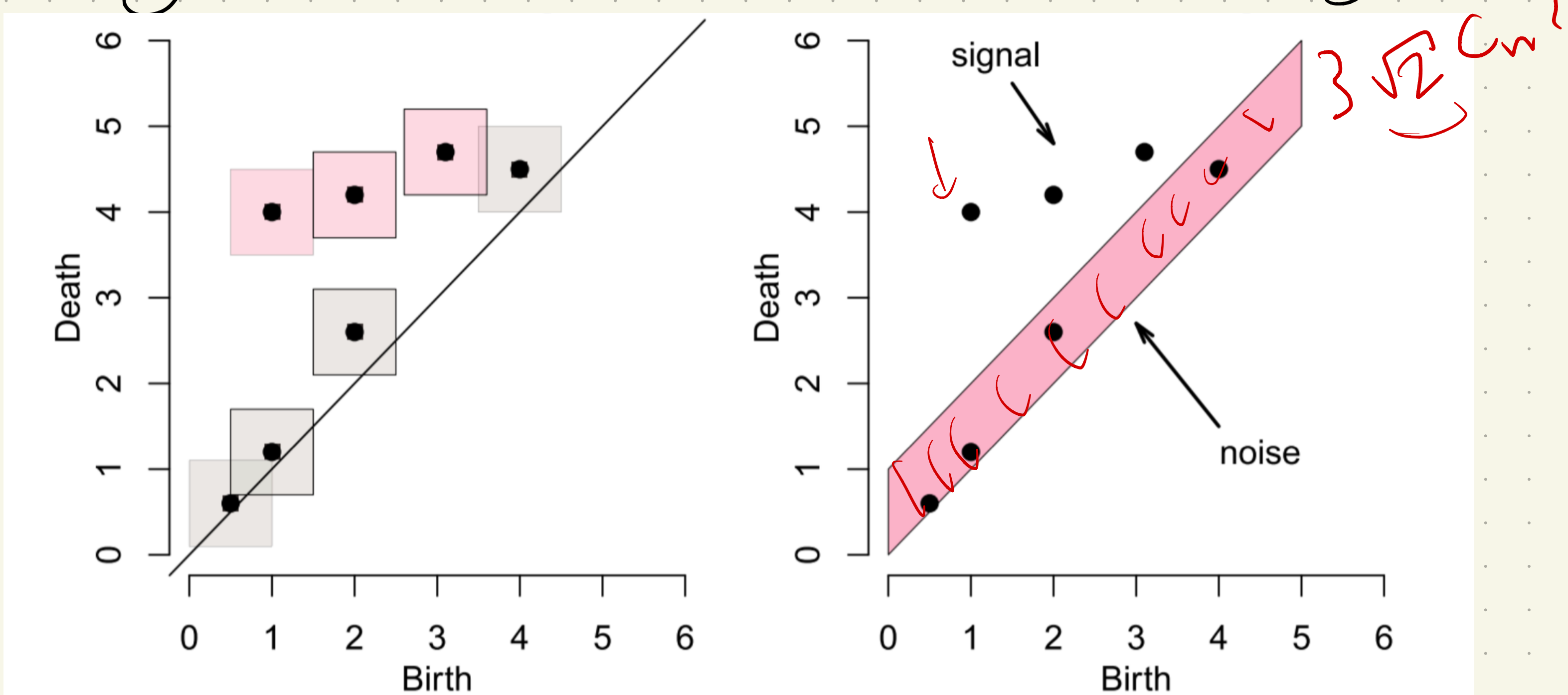
The confidence set C_n is the set of diagrams whose distance to $\hat{X}$ is $\leq c_n$

$$C_n = \{ Y \mid d_B(\hat{X}, Y) \leq c_n \}$$

Assume you have $\hat{X}$ + c_n!

Put a box of width $2c_n$ at every point in $\hat{X}$.

A point is noise if its box intersects the diagonal → or, put strip along diagonal!



How to get c_n though? (don't have M)

- Start with data $S = \{X_1, \ldots, X_n\}$
- Choose $b = b_n$ such that $b = o\left(\frac{n}{\log n}\right)$
- Pretend we have all $N = \binom{n}{b}$ subsamples $S^1, \ldots, S^N$

  ↳ "bootstrapping"

  (In reality: just do a few + pray)
- Calculate $d_H(S^j, S)$, $j = 1 \ldots N$
- Set $L_b(t) = \frac{1}{N} \sum_{j=1}^{N} I(T_j > t)$ &

  set $c_b = 2 L_b^{-1}(\alpha)$

"subsampling method"

(In paper, present one other)

What now??
Using a theorem here:

Theorem

For mild assumptions on the space M, and for all large n,

$$\mathbb{P}(d_B(\hat{X}, X) > c_b) \leq \mathbb{P}(d_H(\mathcal{S}_n, \mathbb{M}) > c_b) \leq \alpha + O\left(\frac{b}{n}\right)^{\frac{1}{4}}$$

[Note: there is every chance you may be better at probability than me...]

Some "toy examples" to demo:

Fasy et al 2014, Annals of Statistics

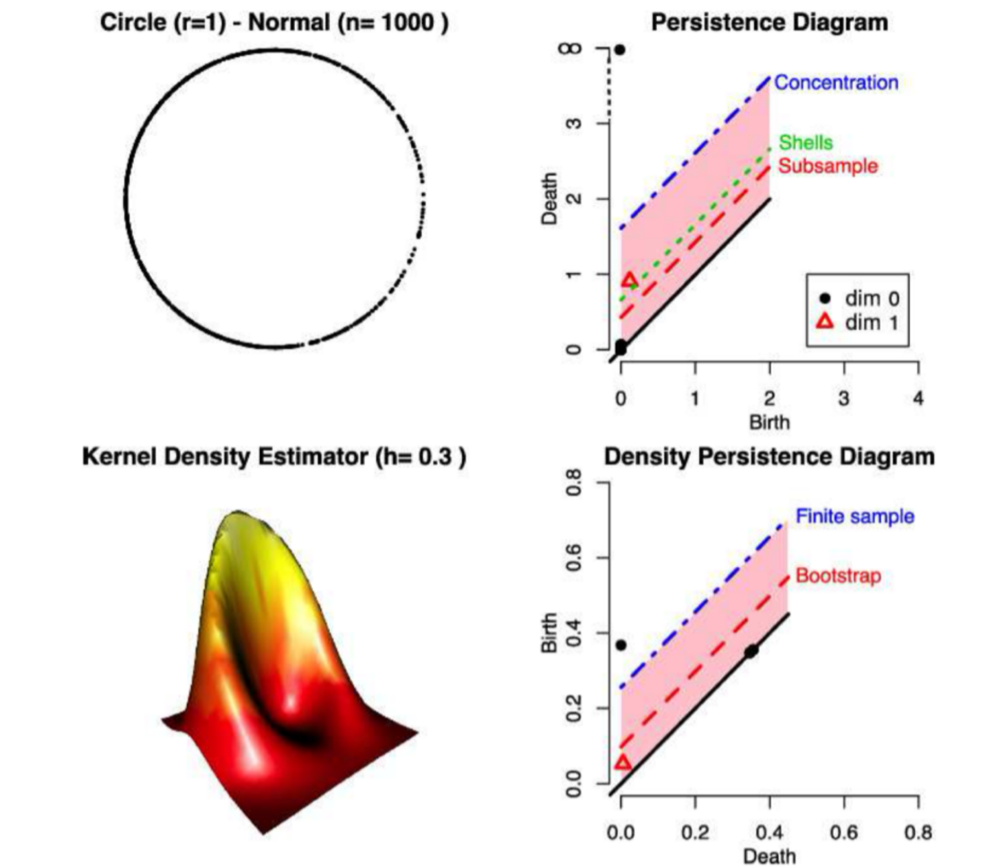


FIG. 7. *Truncated Normal distribution over the unit Circle. (Top left) sample $\mathcal{S}_n$. (Top right) corresponding persistence diagram. The black circles indicate the life span of connected components, and the red triangles indicate the life span of 1-dimensional holes. (Bottom left) kernel density estimator. (Bottom right) density persistence diagram. For more details see Example 14.*

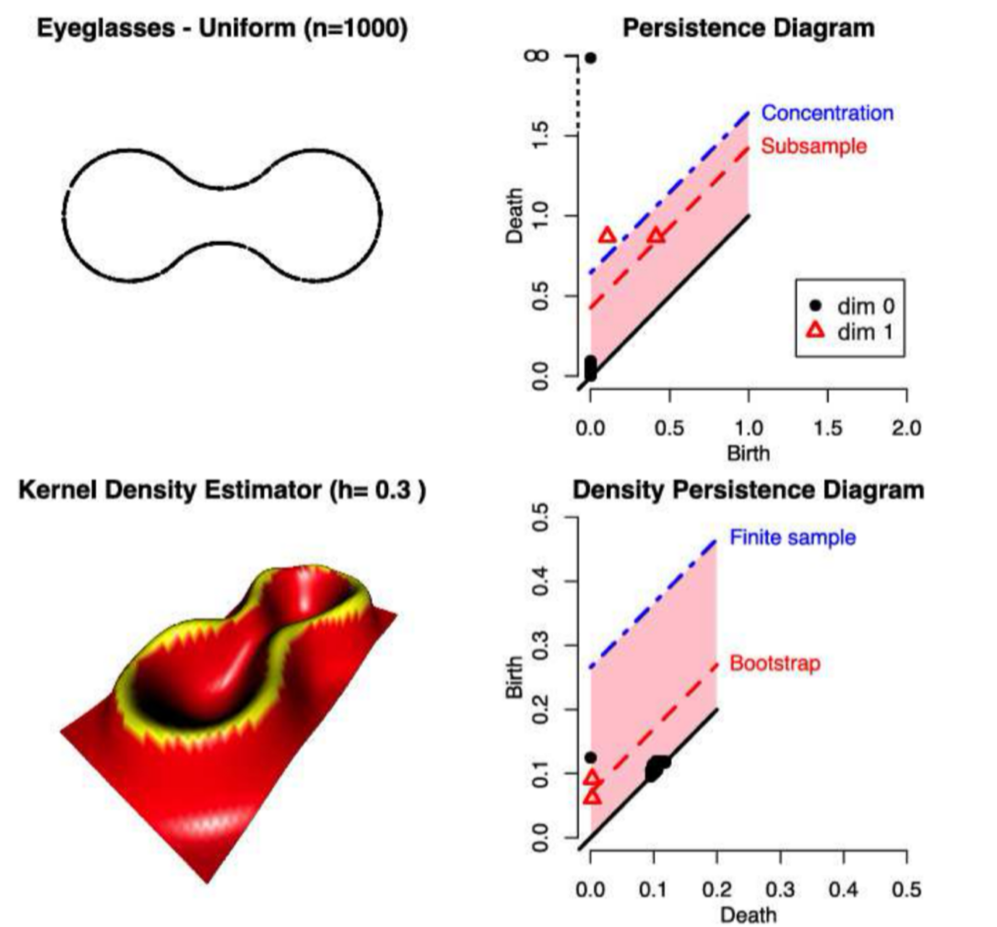
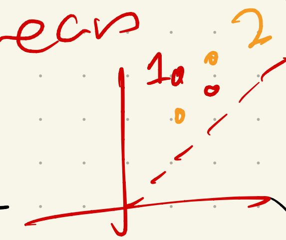


FIG. 8. *Uniform distribution over the eyeglasses curve. (Top left) sample $\mathcal{S}_n$. (Top right) corresponding persistence diagram. The black circles indicate the life span of connected components and the red triangles indicate the life span of 1-dimensional holes. Bottom left: kernel density estimator. (Bottom right) density persistence diagram. For more details see Example 15.*

Some issues + takeaways

Pros:
- Can in some sense prove what is noise versus a feature
  - ↳ confidence intervals
- Can provide some notion of average
  - ↳ Frechet mean



Cons:
- Averages are not unique (even for simple diagrams)
- Confidence intervals require diagrams which we get from subsamples of point clouds
  - ↳ not always the data!
- Unclear how to use with ML
  - ↳ what is an SVM for PDs?

Next: Changing the diagrams

We'll see 2 methods to alter diagrams so we can have better results:

- Persistence landscapes: PLs
- Persistence images: PIs

2018

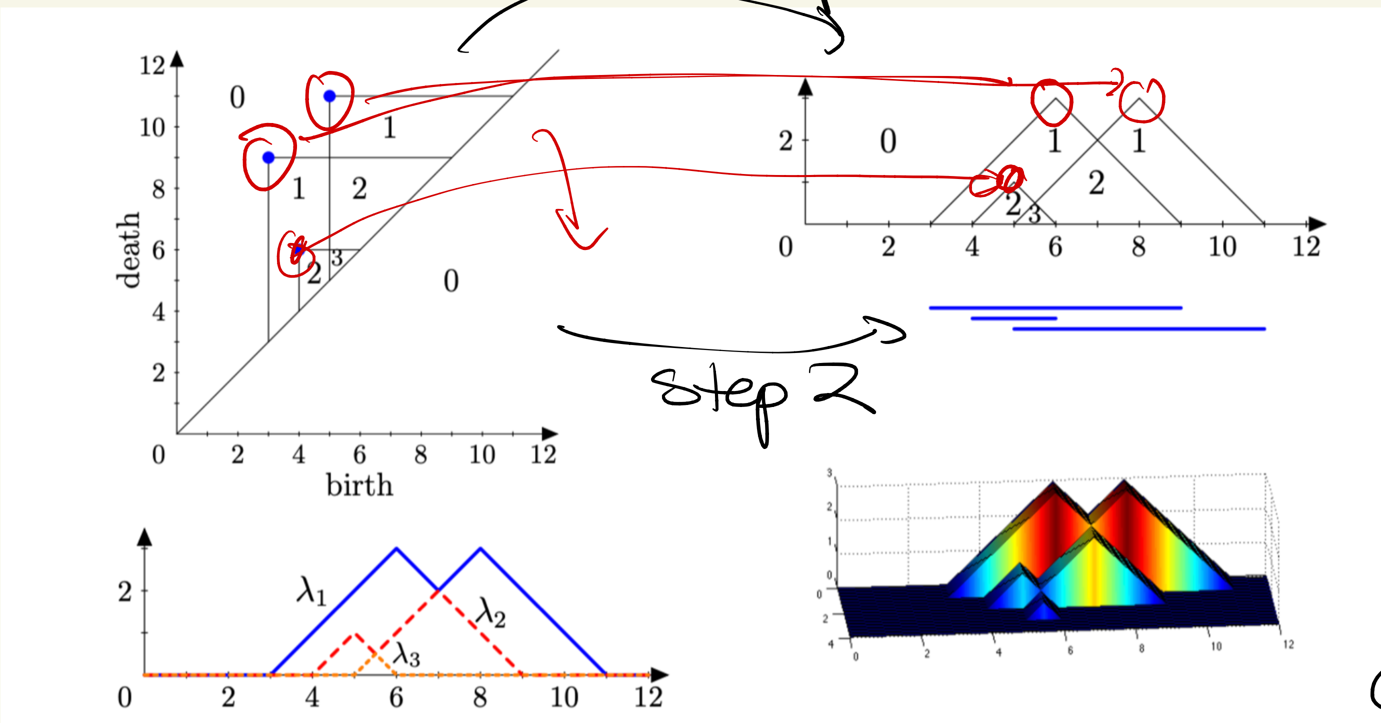
There are others! See Ch. 13 of book, or any NeurIPS or ICML paper mentioning "persistence" in the last 5 years. :)

Pesistence Landscapes

Bubenik, 2014 - JMLR

Definition (by picture): Fix dimension K.

step 3



1) Compute the persistence diagram

2) Rank function:

$$\beta^{a,b} = \dim\left(\mathrm{Im}\left(H_K(X_a) \xrightarrow{f_{a,b}} H_K(X_b)\right)\right)$$

$$H_K(X_1)_a \to \cdots \, _b H_K(X_n)$$

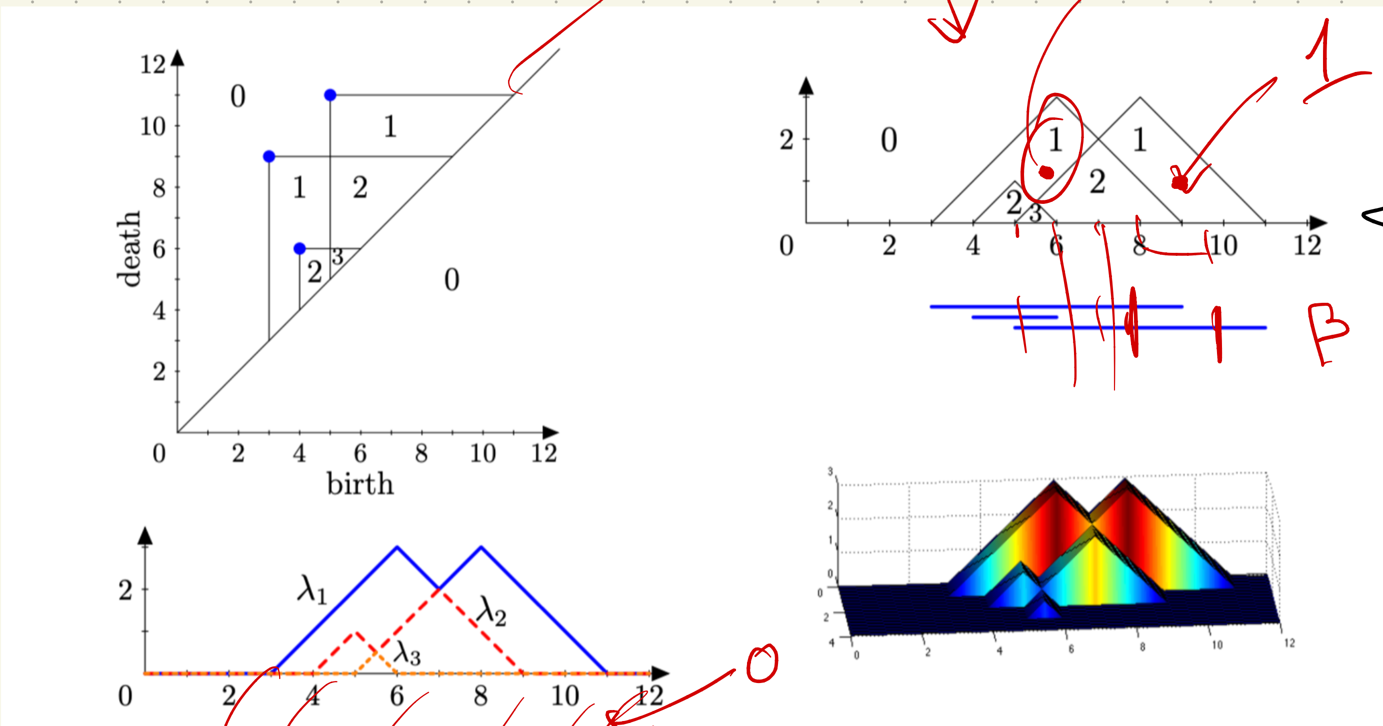
3) Rotate $(x,y) \mapsto \left(\frac{x+y}{2}, \frac{y-x}{2}\right)$,

eg $(3,9) \mapsto \left(\frac{3+9}{2}, \frac{9-3}{2}\right) \to (6,3)$

(cont)

PLs (cont):

$(6,1): \beta^{5,7}$



1

So far: here ✓

β

0

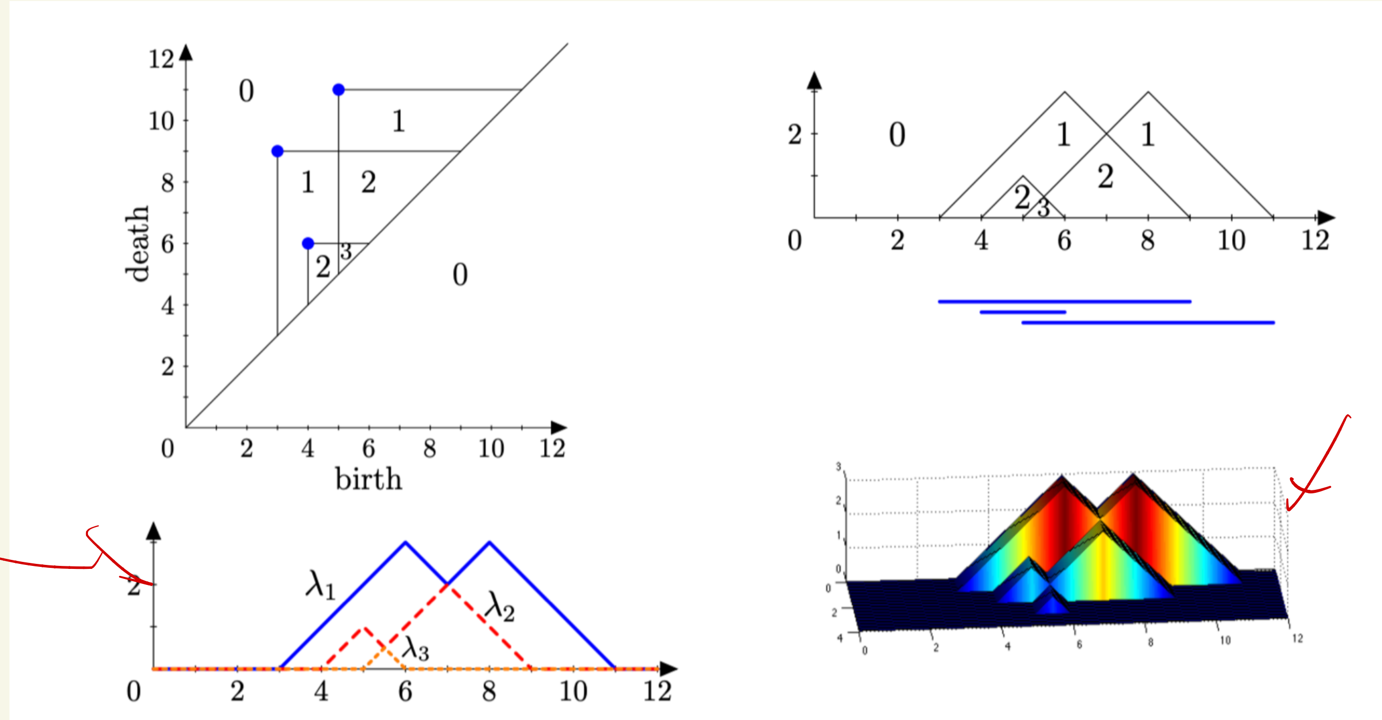
Step 4: Compute $\lambda : \mathbb{R}^2 \to \mathbb{R}$

$$\lambda(m,h) = \begin{cases} \beta^{m-h,\, m+h} & \text{if } h \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

ie: $\lambda(9,1) = \beta^{9-1,\, 9+1} = \beta^{8,10}$

[Calculate via barcodes if easier!]

Step 5:



Next,

let $\lambda_k(t) = \sup\{m \geq 0 \mid \beta^{t-m,t+m} \geq k\}$

So λ_1: max s.t. Betti number is ≥ 1

λ_2: max s.t. " " ≥ 2

etc.

$\Rightarrow$ layes of where Betti # changes

Why? Can get unique averages!
Let's recall Frechet means, + compare:

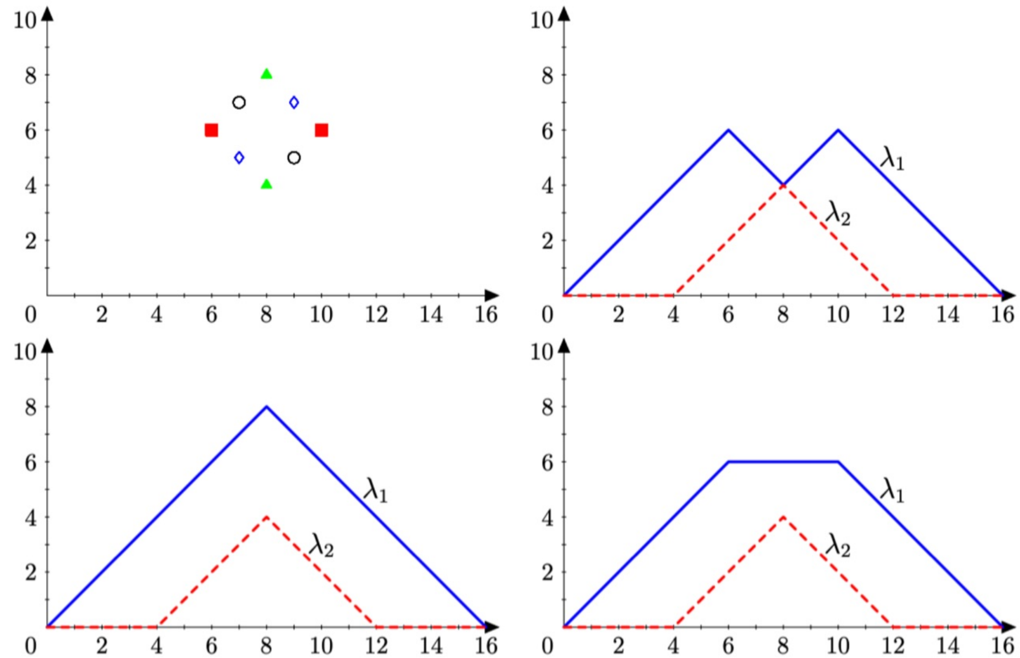


Figure 3: Means of persistence diagrams and persistence landscapes. Top left: the rescaled persistence diagrams $\{(6,6),(10,6)\}$ and $\{(8,4),(8,8)\}$ have two (Fréchet) means: $\{(7,5),(9,7)\}$ and $\{(7,7),(9,5)\}$. In contrast their corresponding persistence landscapes (top right and bottom left) have a unique mean (bottom right).

Can also derive (better?) confidence intervals.

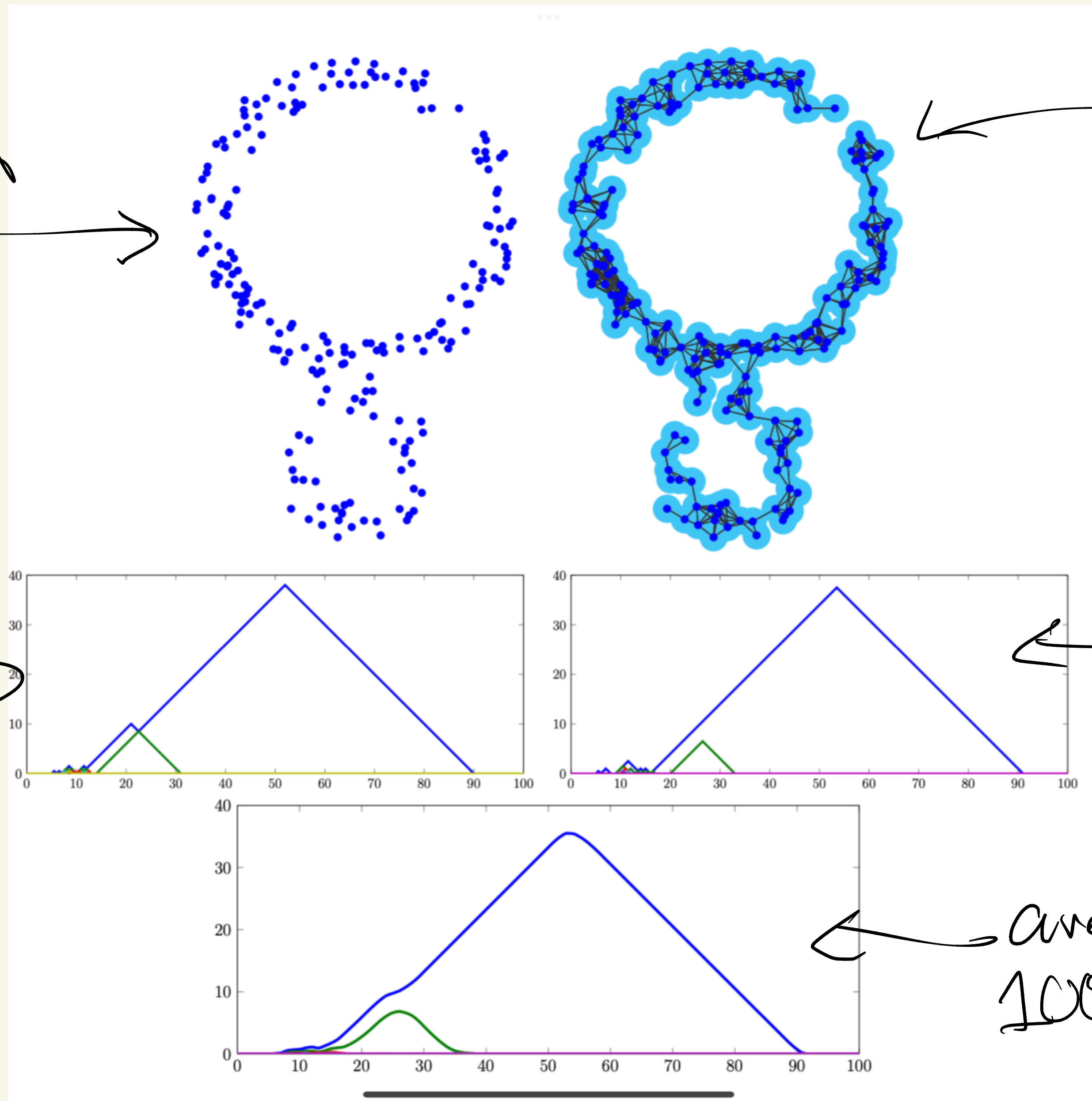
Toy examples : averages

Subsampled points from linked annuli

1 skeleton of Čech complex

Sample 1 for H_1 P.L.o

Sample 2

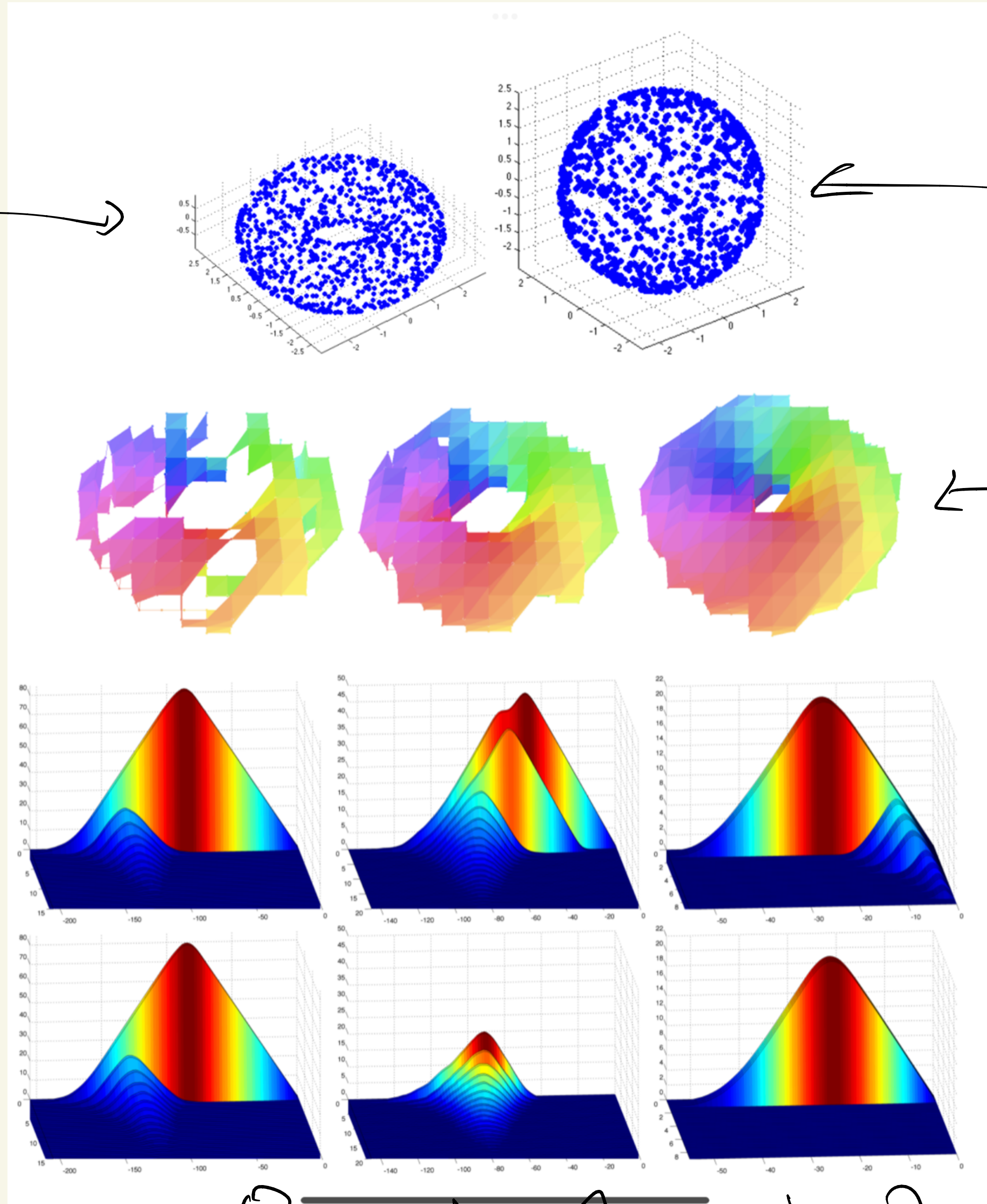


average of 100 samples

Another

Torus sampled

Sphere sampled



Build filtration (torus here)

Torus

Sphere

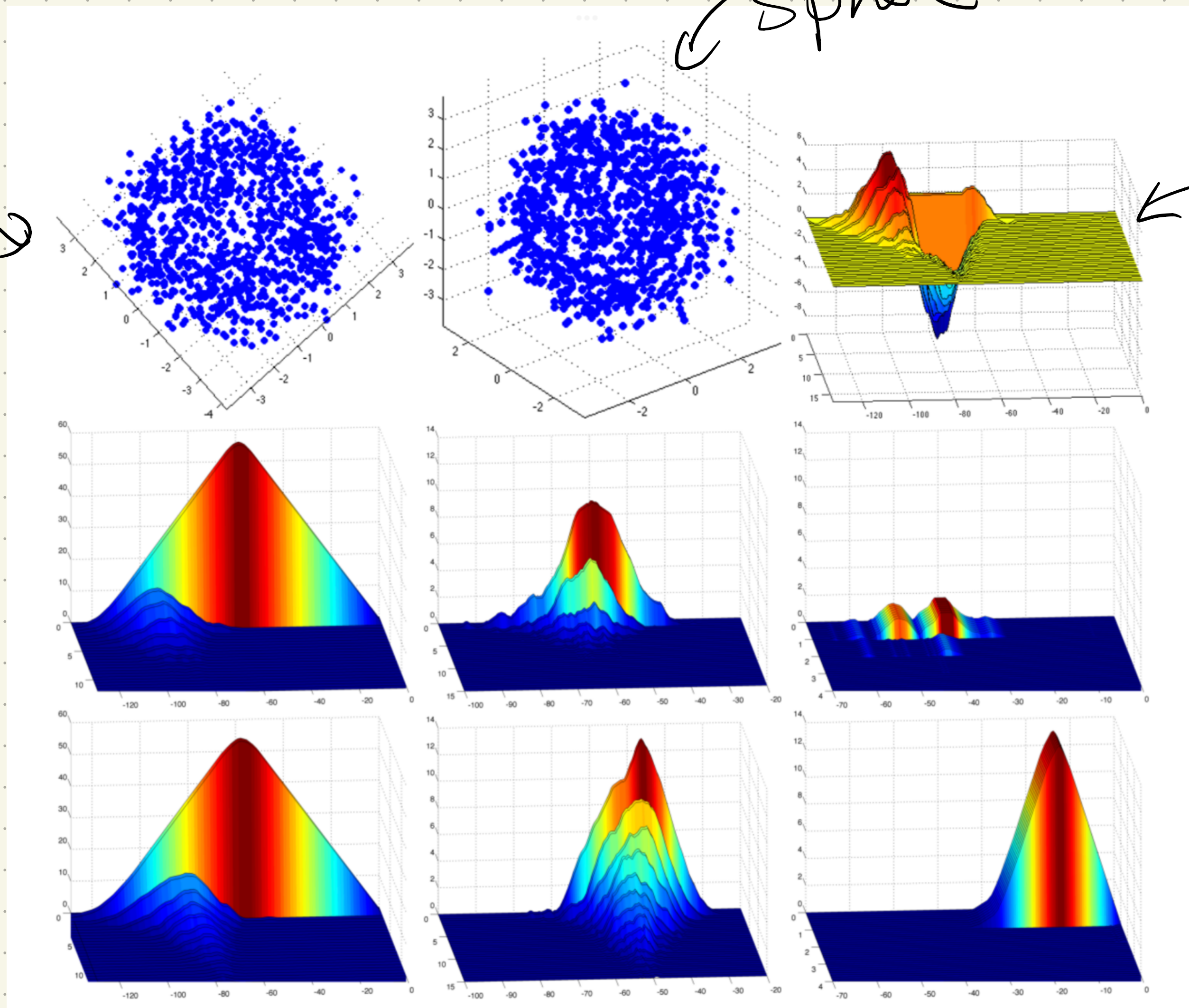
dim 0 dim 1 dim 2

And with noise, can still distinguish!

Sphere

Torus

Gaussian noise



Torus →

Sphere →

10,000 repetition average

dim 0 dim 1 dim 2

Pros / Cons

- Implemented in many libraries, + seems to do well!
  - ↳ can take averages, has notion of stability, & useful in ML
    - Barnes, Polanco, Perea 2021
- Averge is not itself a Persistence diagram
  - ↳ interpretability then suffers

Persistence Images

Adams et al 2017, JMLR

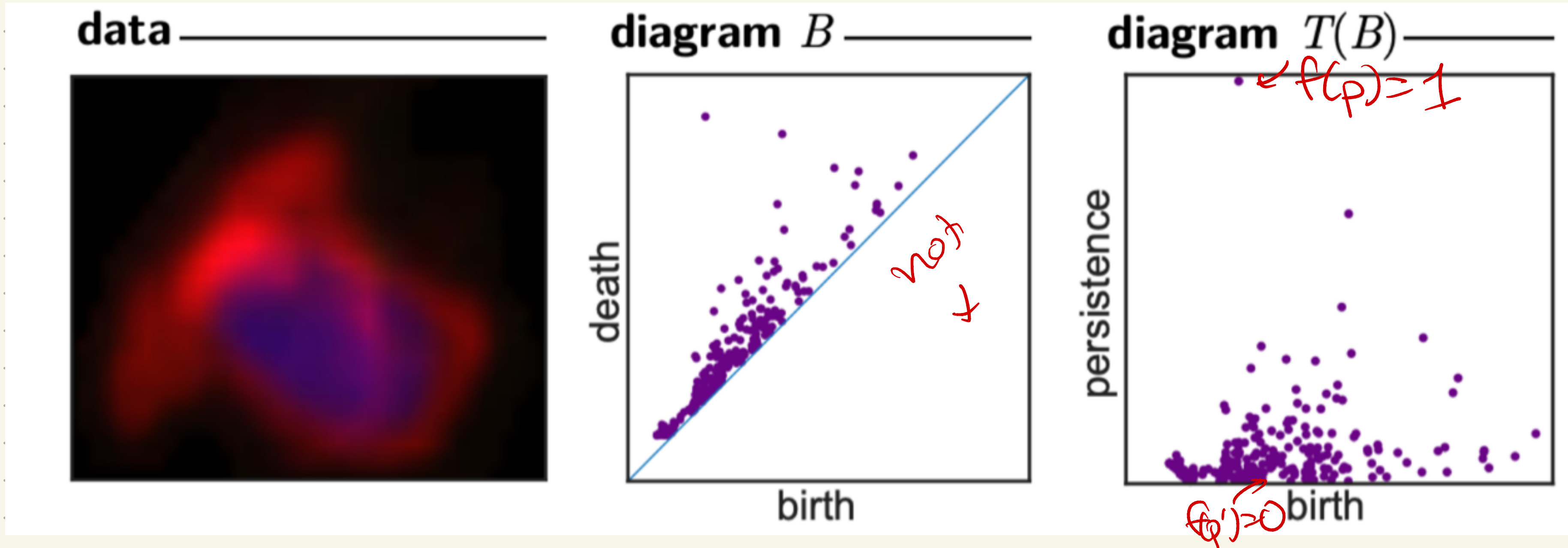
Goal: Represent a persistent diagram so that:

- Output of representation is a vector $\in \mathbb{R}^n$
- representation is stable & efficient to compute
- interpretable connection to original P.D.
- Can adjust relative importance of points in different parts of P.D.

Trade offs with persistence landscaps:

- PLs are invertible & good averages
- PIs are better for ML

Defining (again via picture) PIs



First, transform diagram B to T(B):

$$T(B) = \{(x, y-x) \mid (x,y) \in X\}$$

So: $(3,7) \longmapsto (3,4)$

↳ now have (birth, lifespan) pairs

Next, weight the points: for example

$$f(\vec{p}) = p_Y / (\text{max persistence of any point})$$

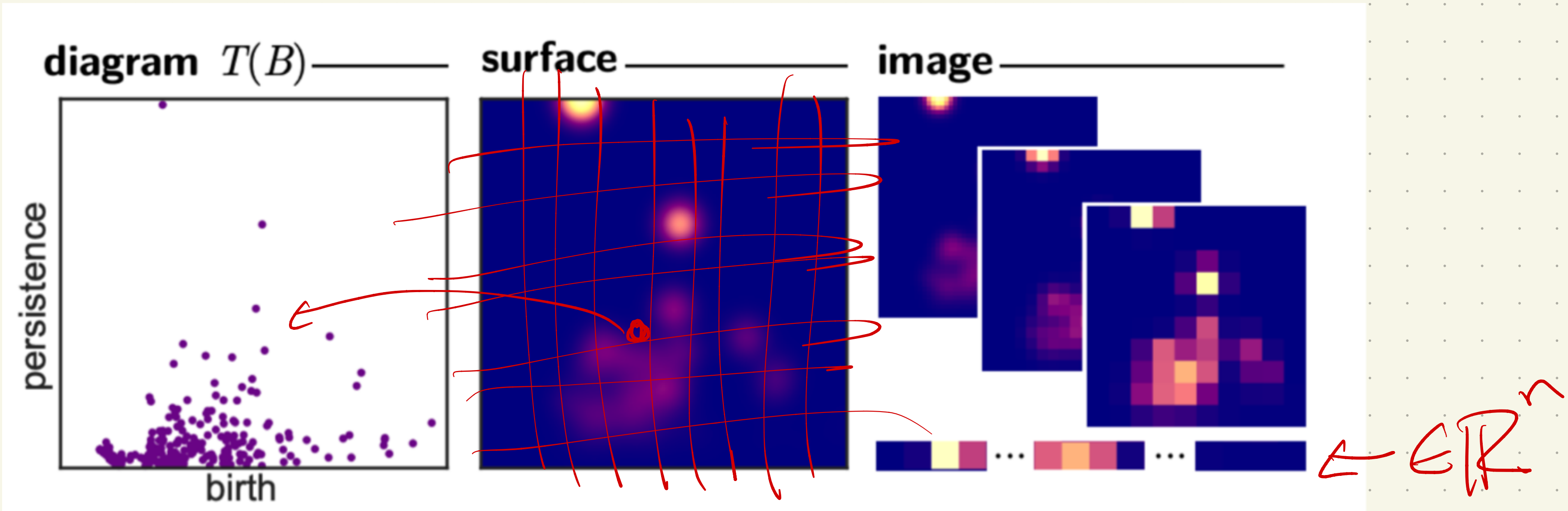
(or could choose other weighting!)

Add Gaussian around points
(or some differentiable probability distribution)

$$e_p(z) = \frac{1}{2\pi\sigma^2} \cdot e^{-((x-p_x)^2 + (y-p_y)^2)/2\sigma^2}$$

(where σ is usual variance)

Then, transform the diagram to a scalar function on $\mathbb{R}^2$



Let $\mu_B : \mathbb{R}^2 \longrightarrow \mathbb{R}$

$$x \longmapsto \sum_{p \in T(B)} f(p) \cdot e_p(x)$$

This gives a "surface".

Then, reduce to finite dimensional vector by "boxing" with value = integral of μ in boxes

Why?

- Stable!
- Lots of choice in parameters & weights & distribution
- Vector output can just plug & play into most ML pipelines
- Seems useful even with noise, & can infer what region influenced results (to some extent)

Aside:

Code tutorial

→ using GUDHI