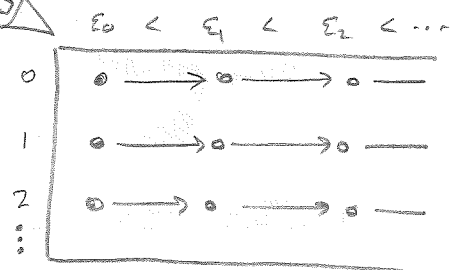The BIG picture: homological data analysis. (persistent homology)

Data
(Points in $\mathbb{R}^n$,
or a finite
metric space)

$\varepsilon > 0$

Increasing family
of simplicial
complexes
(eg $\check{C}ech_\varepsilon$)

$\varepsilon_0 < \varepsilon_1 < \varepsilon_2 < \cdots$
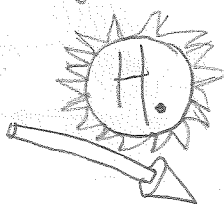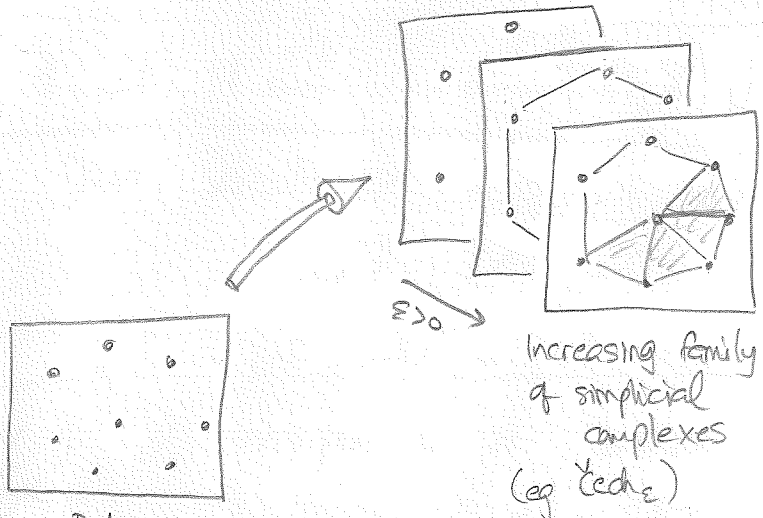
In each dimension d, an
ε-indexed sequence of
vector spaces.

$d = 0$
$d = 1$
$d = 2$
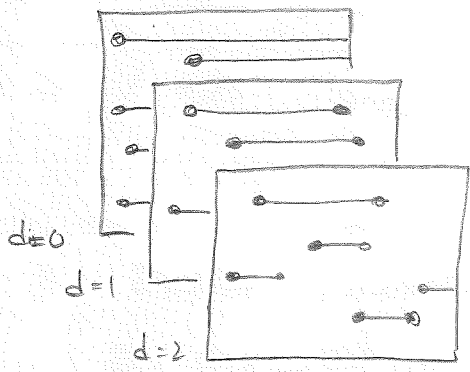
"Barcodes",
one for
each $d \geq 0$.

Today's focus

Def 1

a) A <u>FILTRATION</u> of a simplicial complex K is a sequence of subcomplexes of the form

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n.$$

[We have already seen at least one example: for a finite set $P \subseteq \mathbb{R}^n$, consider $\check{C}ech_\varepsilon(P)$, the $\check{C}ech$ complex whose homology coincides (by the Nerve Theorem) with that of $\bigcup_{p \in P} Ball(p; \varepsilon)$, i.e, union of ε-balls around $p \in P$.]

Here's another example:

b)
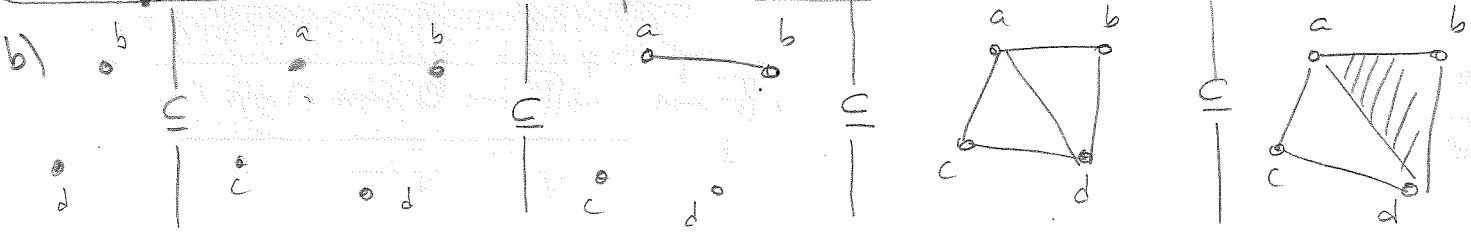
Note 2

Applying "d-dimensional homology with coefficients in a field $\mathbb{F}$, i.e., $H_d(\ ;\mathbb{F})$, to a filtration gives a sequence of vector spaces & linear maps:

$$\boxed{H_d(K_0;\mathbb{F}) \xrightarrow{H_d f_0} H_d(K_1;\mathbb{F}) \xrightarrow{H_d f_1} \dots \xrightarrow{H_d f_{n-1}} H_d(K_n;\mathbb{F})}$$

where f_0 is the <u>inclusion</u> simplicial map sending K_0 to the next K_{0+1}.

(Note that this need NOT be a chain complex!)

Def 3

a) A PERSISTENCE MODULE is a sequence V_0 of vector spaces and linear maps, i.e.,

$$\boxed{V_0 \xrightarrow{\varphi_0} V_1 \xrightarrow{\varphi_1} V_2 \xrightarrow{\varphi_2} \dots \xrightarrow{\varphi_{n-1}} V_n}$$

In our heads, these are the homology of some filtration

[We usually assume $\dim V_i < \infty$ for all i]

---

b) The PERSISTENT HOMOLOGY GROUPS of V_0 are
$$H_i^j(V_0) = \text{im}\left[\varphi_{j-1} \circ \dots \circ \varphi_{i+1} \circ \varphi_i\right], \qquad [i \le j]$$
note that this composite goes from V_i to V_j. For convenience, let's write it as $\varphi_i^j : V_i \to V_j$

---

c) An element $\alpha \in V_i$
- is <u>BORN</u> at i if $\alpha \notin \text{im } \varphi_{i-1}$.
- <u>DIES</u> at $j > i$ if $\varphi_i^j(\alpha) = 0$ but $\varphi_i^{j-1}(\alpha) \ne 0$ say "dies at ∞" if $\exists$ no such j.

Q4 How to organize all this information?

Def 5

Nicest possible persistence modules.

Given $i \le j$, the <u>INTERVAL MODULE</u> $\mathbb{I}_0^{[i,j)}$ is:

$$\dots \to 0 \to \dots \; 0 \to \underbrace{\mathbb{F}}_{\substack{\| \\ V_i}} \xrightarrow{1} \underbrace{\mathbb{F}}_{\substack{\| \\ V_{i+1}}} \xrightarrow{1} \dots \xrightarrow{1} \underbrace{\mathbb{F}}_{\substack{\| \\ V_{j-1}}} \to \underbrace{0}_{\substack{\| \\ V_j}} \to 0 \to \dots$$

To each persistence module $V_\bullet$ of vector spaces over $\mathbb{F}$, one can associate a multi-set of intervals

$$\mathrm{Bar}(V_\bullet) = \left\{ [i,j) \mid 0 \le i < j \le \infty \right\} \leftarrow \text{"Barcode" of } V_\bullet$$

So that $V_\bullet$ is isomorphic to a direct sum of interval modules.

$$V_\bullet \cong \bigoplus_{[i,j) \in \mathrm{Bar}(V)} \mathbb{I}_\bullet^{[i,j)}$$

"Pf"

Every persistence module is an "honest module" in the sense of abstract algebra over the polynomial ring $\mathbb{F}[t]$, where t acts on $\alpha \in V_i$ by pushing it to $\varphi_i(\alpha) \in V_{i+1}$. These "finitely generated modules" over $\mathbb{F}[t]$, a "principal ideal domain", decompose as follows: $\exists \{\alpha_i, i \in I\}$ and $\{\beta_j < \delta_j, j \in J\}$ so that

$$V_\bullet \cong \bigoplus_{\alpha_i} t^{\alpha_i} \cdot \mathbb{F}[t] \oplus \bigoplus_{\beta_j < \delta_j} t^{\beta_j} \cdot \mathbb{F}[t] \big/ (t^{\delta_j})$$

$$\underbrace{\hspace{5cm}}_{\text{"free part"}} \quad \underbrace{\hspace{5cm}}_{\substack{\text{"torsion part"} \\ \text{quotient by ideal}}}$$

and in this case

$$\mathrm{Bar}(V_\bullet) = \coprod_i \left\{ [\alpha_i, \infty) \right\} \amalg \coprod_j \left\{ [\beta_j, \delta_j) \right\}$$

$$\underbrace{\hspace{3cm}}_{\text{infinite bars}} \qquad \underbrace{\hspace{3cm}}_{\text{finite bars}}$$

Returning to our example: (see Def 1b)
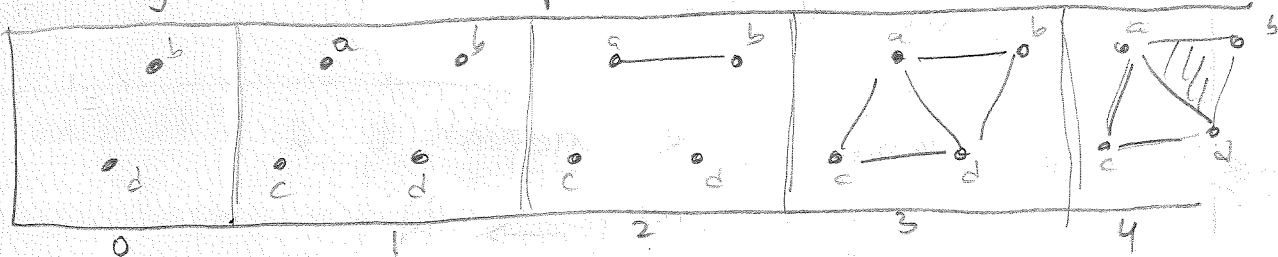
0d barcode $\quad [0,\infty), \; [0,3), \; [1,2), \; [1,3)$

$\qquad\qquad\qquad\qquad\;\; b \qquad\quad d \qquad\quad a \qquad\quad c$

1d- barcode $\quad [3,\infty), \; [3,4)$

all other barcodes: $\quad \emptyset$

Fig 8 | Let's draw these:

dim 0   dim 1.
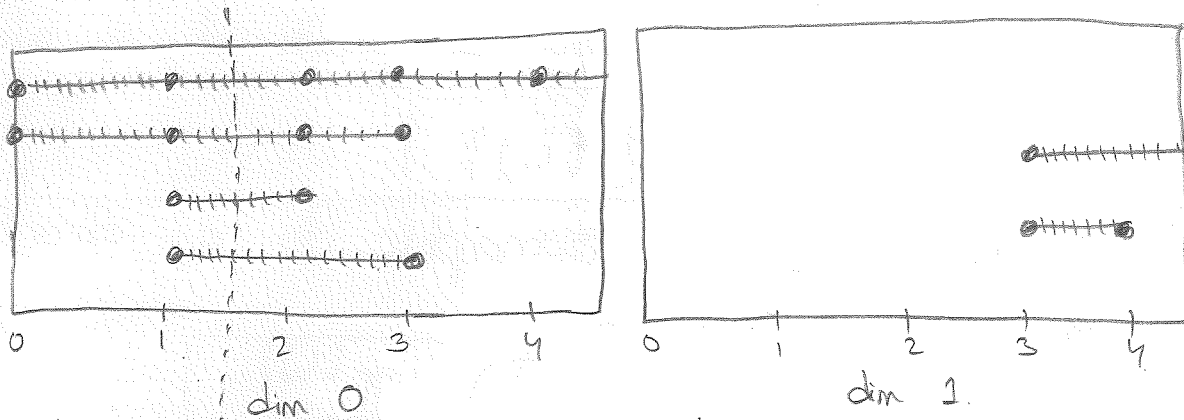
intersections gives
Betti number! $\beta_0 = 4$.
between "1" and "2"

ALGORITHM

Note 9

a) For a one-step filtration $K_0 = K$, computing persistent homology is the same as computing ordinary homology $H_d(K;F)$: all barcode-intervals are of the form $[0,\infty)$, and their number is just $\dim H_d(K;F)$. So, we will need the usual row-and-column operations on boundary matrices.

b) BUT for a general filtration
$$K_0 \subseteq K_1 \subseteq \ldots \subseteq K_{n-1} \subseteq K_n = K,$$
there is a nice & simple ordering of the simplices which allows us to compute the barcode all at once using ONLY COLUMN OPERATIONS.

Note 10   Every filtration of K corresponds to a "FILTERING FUNCTION"
$$\underline{c} : \left\{ \begin{array}{c} \text{simplices} \\ \text{of } K \end{array} \right\} \longrightarrow \mathbb{N}$$
where $\underline{c}(\sigma) \leq \underline{c}(\tau)$ whenever σ is a face of τ, as follows.
$$K_i = \{ \sigma \in K \mid \underline{c}(\sigma) \leq i \} \quad \text{or} \quad \text{(to go back)}$$
$$\underline{c}(\sigma) = \min \{ i \in \mathbb{N} \mid \sigma \in K_i \}$$

Algo 11

a) Assume that simplices of K have been totally ordered into $(a_1, a_2, \ldots, a_N)$ so that
- if $\underline{c}(a) \leq \underline{c}(\tau)$ then a precedes τ, and
- if $\underline{c}(a) = \underline{c}(\tau)$ and $\dim a < \dim \tau$, then a precedes τ.

i.e., order first by $\underline{c}$-value, then by dimension [you can break ties arbitrarily.

- -

b) The input is the ENTIRE boundary matrix, which will be block-upper triangular.

$$\Delta = \begin{array}{c} \\ a_1 \\ a_2 \\ \\ \\ a_j \\ a_N \end{array} \overset{\displaystyle a_1 \ a_2 \ \cdots \ a_i \ \cdots \ a_N}{\left[\begin{array}{c} \\ \\ \\ \boxed{m_{ij}} \\ \\ \\ \end{array} \right]}$$

Here,
$$m_{ij} = \begin{cases} \pm 1 & \text{if } a_j \leq a_i \text{ with } \dim a_i - \dim a_j = 1 \\ 0 & \text{else.} \end{cases}$$

The sign depends on orientation as usual.

- -

c) For each $i \in \{1, \ldots, N\}$, let $\underline{low}(i)$ be the largest row-index j so that $m_{ij} \neq 0$ [i.e, index of the last-born codimension-one face of a_i]

- -

PSEUDO-CODE ! | **Set P = empty list of simplex-pairs**

1. <u>For</u> $i = 1$ to N
2. Set $k = low(i)$ [$= 0$ if col(i) is zero].
3. <u>While</u> $\exists\, j < i$ with $low(j) = k \neq 0$,
4. | <u>add</u> $[-m_{ik}/m_{jk}] \cdot col(j)$ to col(i)
5. <u>end</u> while
6. <u>end</u> for

This uses the previous lows to clear out the current low.

- -

Namely, IF two columns have their lowest nonzero entry in the same row, THEN add a multiple of the left one to zero out the lowest entry of the right one.

$$k\left[\begin{array}{cc} \overset{j}{\vdots} & \overset{i}{\vdots} \\ 0 & 0 \end{array} \right] \Rightarrow col\, i \leftarrow col\, i + \underline{\alpha}\, col\, j$$

Prop 12

The trinity

When the above algorithm terminates, there are three possibilities in each column i, labelled ①, ②, ③:

START → is $col(i) = 0$?

No → Set $k = low(i)$ → ① → σ_i contributes a bar $[c(\sigma_k), c(\sigma_i))$ in dim σ_k

Yes ↓ → is $i = low(j)$ for some j ?

No → ② → σ_i contributes a bar $[c(\sigma_i), \infty)$ in dim σ_i

Yes → ③ → σ_i does not contribute a new bar beyond what has already happened in ① for column j
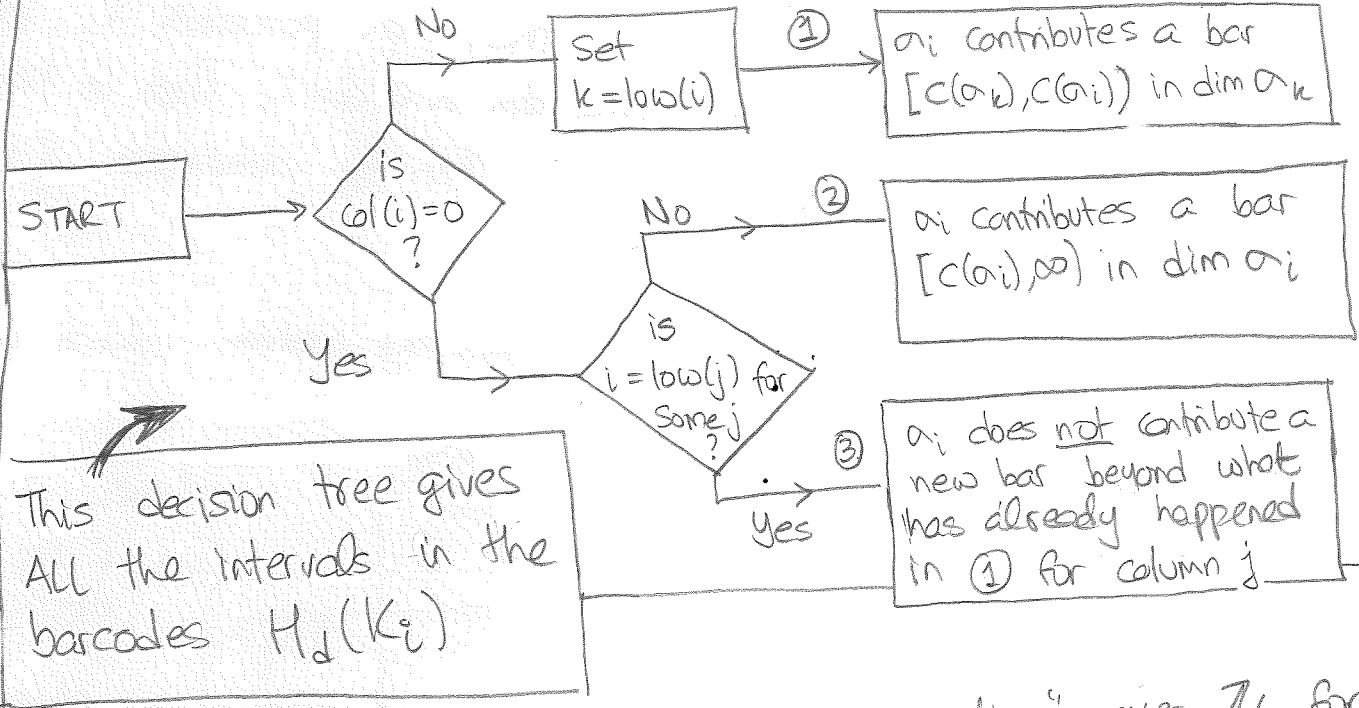
This decision tree gives ALL the intervals in the barcodes $H_d(K_i)$

Note 13

Let's see a few iterations "in practice" over $\mathbb{Z}/2$ for

K_0 K_1 K_2 K_3

$$C = $$

	a	b	c	d	ac	ab	ad	bd	cd	acd
	0		1			2				3

$$\Delta = $$
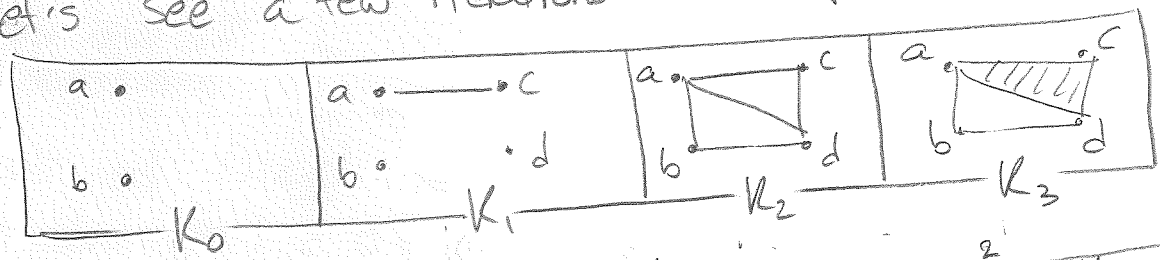
rows: a, b, c, d, ac, ab, ad, bd, cd, acd

[All unwritten entries are zero. In general this matrix is VERY sparse!]

The first nontrivial step is $col(bd) \leftarrow col(bd) + col(ad)$, which gives $col(bd) = [1, 1, 0, \ldots]^{transpose}$ The next step is $col(bd) \leftarrow col(bd) + col(ab)$, which clears it out completely!

Can you keep going?

So far, the "main theorem" of persistent homology appears to be the STRUCTURE THEOREM (Thm 6). But there is a NEWER, DEEPER result which justifies the use of homology in data analysis.

STABILITY

Def 15

a) Let B and B' be two barcodes. An <u>ε-MATCHING</u> between them is a bijection $\mu: B_0 \xrightarrow{\sim} B_0'$ where $B_0 \subseteq B$ and $B_0' \subseteq B'$ satisfying:
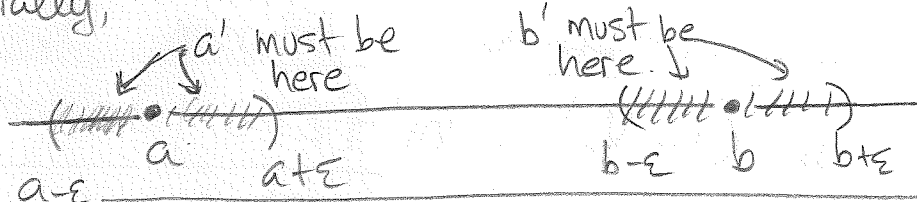
(A) All intervals in $(B - B_0)$ and $(B' - B_0')$ have length $\leq 2\varepsilon$.

(B) If $\mu([a,b]) = [a',b']$ for some $[a,b] \in B_0$, then:

$$|a - a'| < \varepsilon > |b - b'|$$

> Note: for infinite bars $[a, \infty)$, $[a', \infty)$ we don't have any constraint on the "b". Just $|a-a'| < \varepsilon$.

Pictorially,

a' must be here / b' must be here.

**b)** The BOTTLENECK DISTANCE between barcodes B and B' is given by

$$d_{BOT}(B, B') = \inf\{\varepsilon > 0 \mid \exists \varepsilon\text{-matching } B \xrightarrow{\mu} B'\}$$

Prop 16 d_{BOT} is a METRIC on the set of all barcodes with finitely many bars, EXCEPT for the fact that we allow $d_{BOT}(B, B') = \infty$. [For instance, if the number of infinite-length intervals in B is different from the number of such intervals in B', then there is no possible finite $\varepsilon > 0$ for which there exists an ε-matching!]

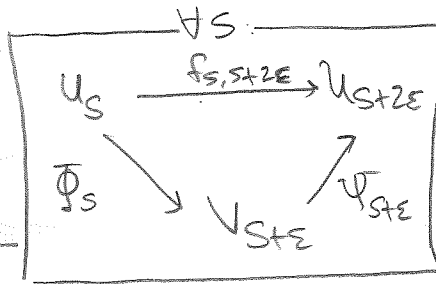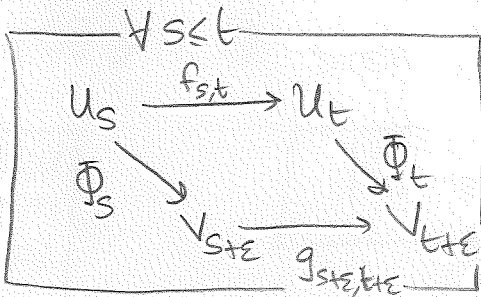Now we can compare barcodes, what about persistence modules?

Def 18 Let $U_\bullet$ and $V_\bullet$ be two $\mathbb{R}^+$-indexed persistence modules.

a)

$$U_0 \quad U_s \xrightarrow{\;f_{s,t}\;} U_t \quad \cdots = U_\bullet$$
$$V_0 \quad V_s \xrightarrow{\;g_{s,t}\;} V_t \quad \cdots = V_\bullet$$
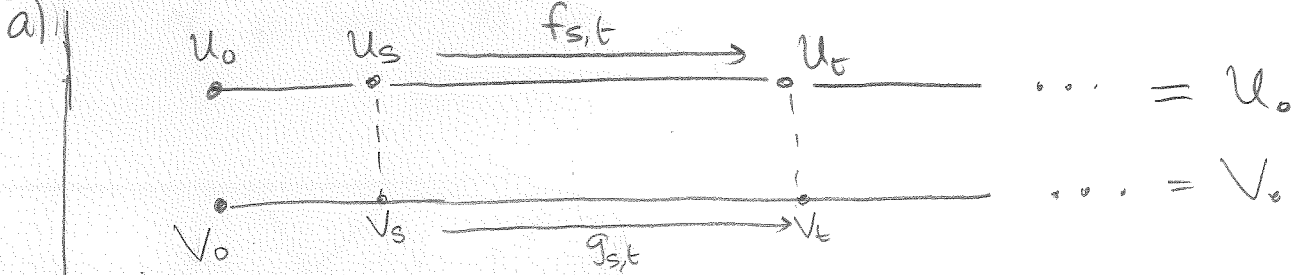
So: each U_t is a vector space for $t \geq 0$, there is a map $f_{s,t}: V_s \to V_t$ for $s \leq t$ so that $\boxed{f_{s,s} = \mathrm{id} \;\forall s}$ and $\boxed{f_{s,u} = f_{t,u} \circ f_{s,t} \;\forall s \leq t \leq u}$

(Φ_t and Ψ_t are LINEAR maps of vector spaces)

An ε-INTERLEAVING (for $\varepsilon > 0$) between $U_\bullet$ and $V_\bullet$ is a family of maps $\{\Phi_t : U_t \to V_{t+\varepsilon} \quad \forall t \in \mathbb{R}^+\}$ and $\{\Psi_t : V_t \to U_{t+\varepsilon} \quad \forall t \in \mathbb{R}^+\}$ so that all possible triangles and parallelograms commute, eg.

$\forall s \leq t$

$$U_s \xrightarrow{f_{s,t}} U_t$$
$$\Phi_s \searrow \quad \searrow \Phi_t$$
$$V_{s+\varepsilon} \xrightarrow{g_{s+\varepsilon, t+\varepsilon}} V_{t+\varepsilon}$$

$\forall s$

$$U_s \xrightarrow{f_{s,s+2\varepsilon}} U_{s+2\varepsilon}$$
$$\Phi_s \searrow \quad \nearrow \Psi_{s+\varepsilon}$$
$$V_{s+\varepsilon}$$

[+ two more!]

b) The INTERLEAVING DISTANCE between $U_\bullet$ and $V_\bullet$ is

$$\boxed{d_{INT}(U_\bullet, V_\bullet) = \inf\{\varepsilon > 0 \mid \exists \; \varepsilon\text{-interleaving} \ldots\}}$$

THM 19 (STABILITY THEOREM! IMPORTANT) If $U_\bullet$ and $V_\bullet$ are tame $\mathbb{R}^+$-indexed persistence modules then there is an isometry

$$\boxed{d_{INT}(U_\bullet, V_\bullet) = d_{BOT}(Bar(U_\bullet), Bar(V_\bullet))}$$

The most elementary proof is Lesnick + Bauer "induced matching"...

Prop 20 So what? Well, if point-sets $P, Q \subseteq \mathbb{R}^n$ have Hausdorff dist $\varepsilon > 0$, then $H_d(\check{C}_\bullet(P))$ and $H_d(\check{C}_\bullet(Q))$ are ε-interleaved [+ same for $VR_\bullet(P)$ & $VR_\bullet(Q)$] for each dimension $d \geq 0$. So,

$$\boxed{d_{Haus}(P,Q) \geq d_{INT}(H_d \check{C}_\bullet(P), H_d \check{C}_\bullet(Q)) = d_{BOT}(Bar_d(P), Bar_d(Q))}$$

So, noise in data is no more than noise in Barcode! [Here $Bar_d(P)$ is shorthand for $Bar(H_d \check{C}_\bullet(P))$, etc]