Algorithmische Topologie und ihre Anwendungen: Persistente Homologie und Topologische Datenanalyse

Mikael Vejdemo-Johansson

16. November 2010

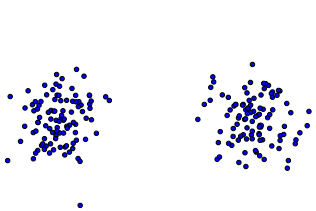
Datenmengen haben Form

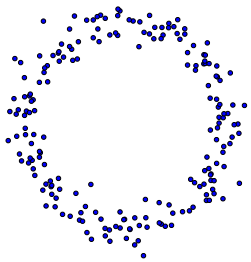
Was sind Daten?

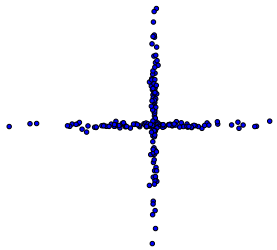
Gruppen von Messwerten: z.B. physiologische Messungen – Körpergewicht, Größe, Blutdruck, . . . von Patienten.

Wird als (endliche) Mengen von Vektoren im $\mathbb{R}^d$ aufgefasst.

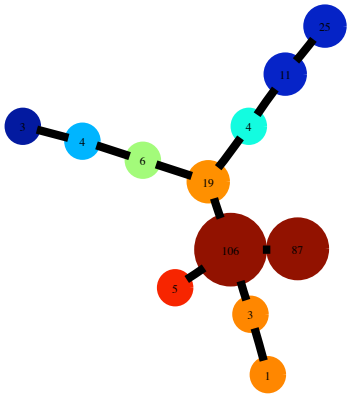
Was ist Form?

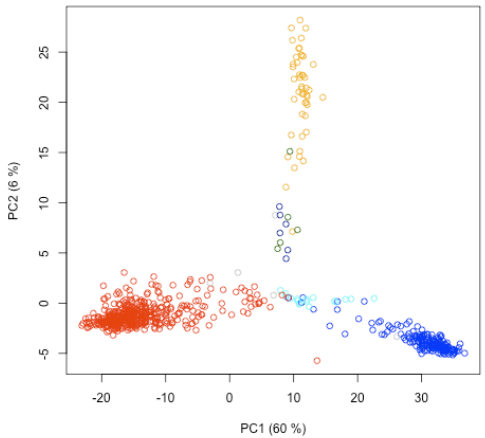


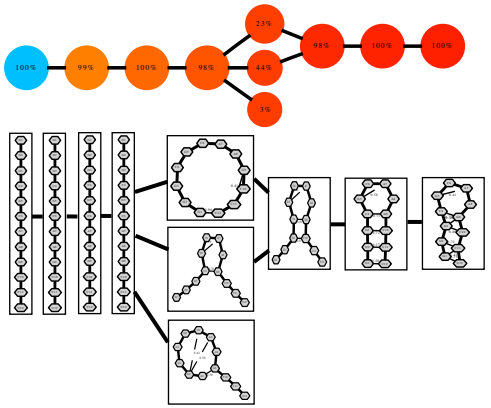




Form ist wichtig







Homologie

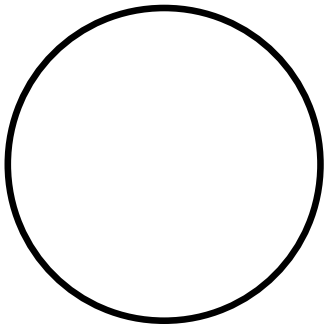
Unser Hauptwerkzeug, um diese Formen zu beschreiben, kommt aus der Topologie.

Homologie weist (für einen Körper k und eine nichtnegative ganze Zahl i) jedem topologischen Raum X einen Vektorraum $H_i(X; k)$ zu.

Einfachste Beschreibung: $\beta_i = \dim_k H_i(X; k)$ – Bettizahl. Zählt die Anzahl der i-dimensionalen Hohlräume.

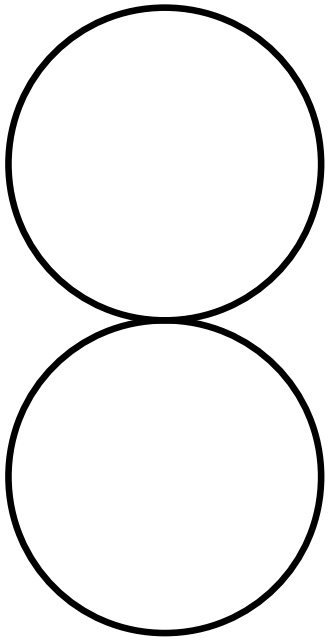
Von triangulierbaren Räumen X lässt sich die Homologie durch Matrixoperationen berechnen.

Homologie – intuitiv



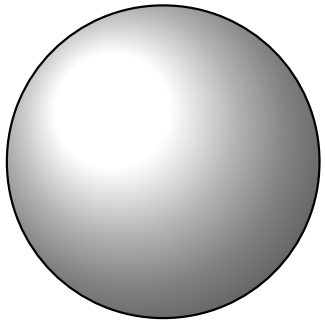
$\beta_0 = 1$

$\beta_1 = 1$



$\beta_0 = 1$

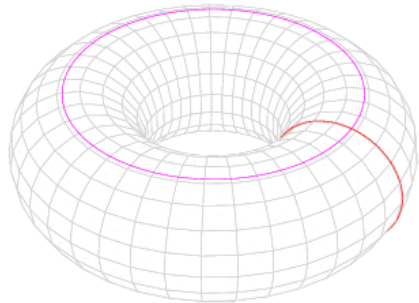
$\beta_1 = 2$



$\beta_0 = 1$

$\beta_1 = 0$

$\beta_2 = 1$



$\beta_0 = 1$

$\beta_1 = 2$

$\beta_2 = 1$

Homologie – warum algebraisch?

Auch wenn wir am liebsten nur mit β_i arbeiten, bleibt die Algebra die durch den Vektorräumen eingeführt wird, wichtig.

Kern davon: wenn sich die Vektorräume verändern, verändern sich auch die Homologie-Vektorräume, und zwar durch lineare Abbildungen.

$$X \xrightarrow{f} Y$$

$$H_i(X;k) \xrightarrow{H_i(f;k)} H_i(Y;k)$$

Die Vektorraumstruktur trägt zusätzliche Informationen, die wir ausnutzen können.

Wir kommen bald auf die Funktorialität zurück.

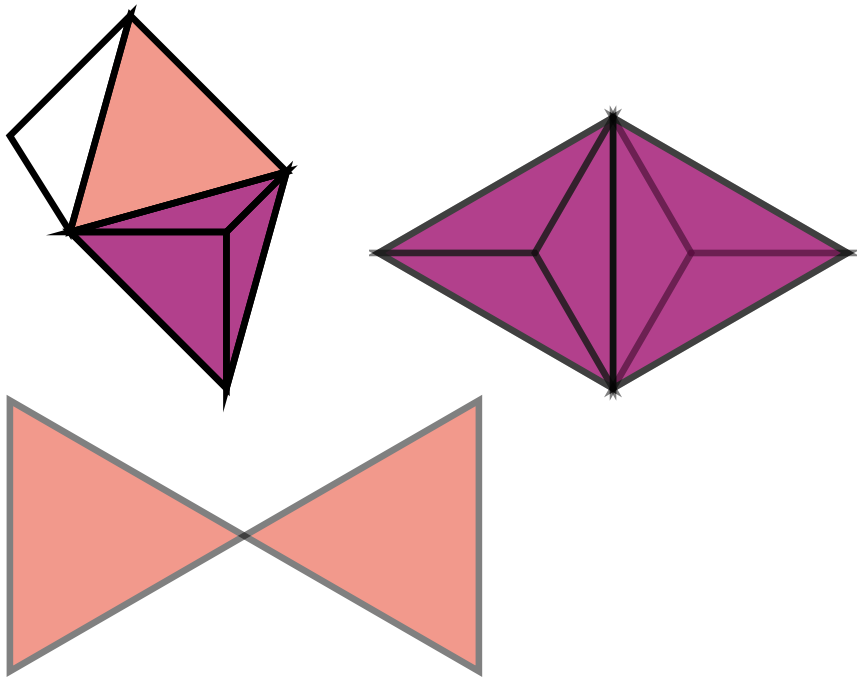
Simpliziale Topologie: aus stetig wird diskret

Definition

Ein simplizialer Komplex ist eine Familie von Simplizes – Punkt, Strecke, Dreieck, Tetrahedron, . . . – so dass alle Simplizes sich in Teilsimplizes schneiden.

Definition

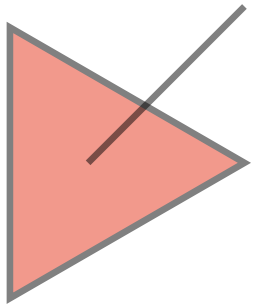
Ein abstrakter simplizialer Komplex ist eine Familie von Teilmengen einer Menge V, so dass, wenn $\sigma \in V$ und $\tau \subset \sigma$, auch $\tau \in V$ ist.

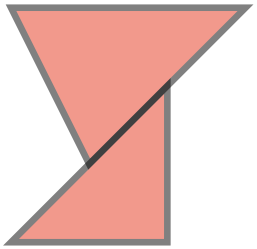


Simpliziale Topologie: aus stetig wird diskret

Definition

Ein simplizialer Komplex ist eine Familie von Simplizes – Punkt, Strecke, Dreieck, Tetrahedron, ... – so dass alle Simplizes sich in Teilsimplizes schneiden.





Definition

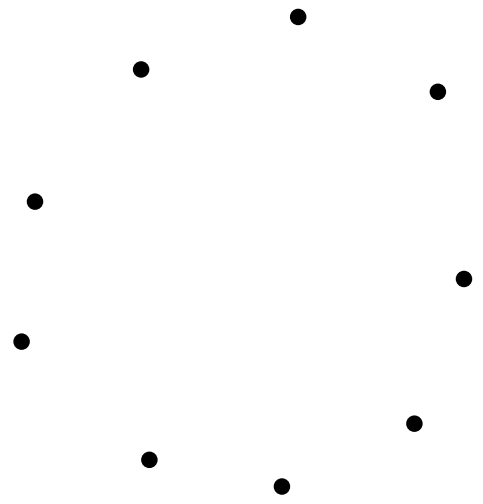
Ein abstrakter simplizialer Komplex ist eine Familie von Teilmengen einer Menge V, so dass, wenn $\sigma \in V$ und $\tau \subset \sigma$, auch $\tau \in V$ ist.

Simplizielle Komplexe: aus diskret wird stetig

Definition

Der Vietoris-Rips-Komplex ist ein abstrakter simplizialer Komplex $VR_\epsilon(X)$ für $\epsilon \in \mathbb{R}_+$ und X:

- Enthält 0-Simplizes (Ecken) entsprechend X
- Enthält $(x_0, \ldots, x_k)$ genau dann, wenn $d(x_i, x_j) < \epsilon$ für alle verschiedenen $0 \leq i, j \leq k$.

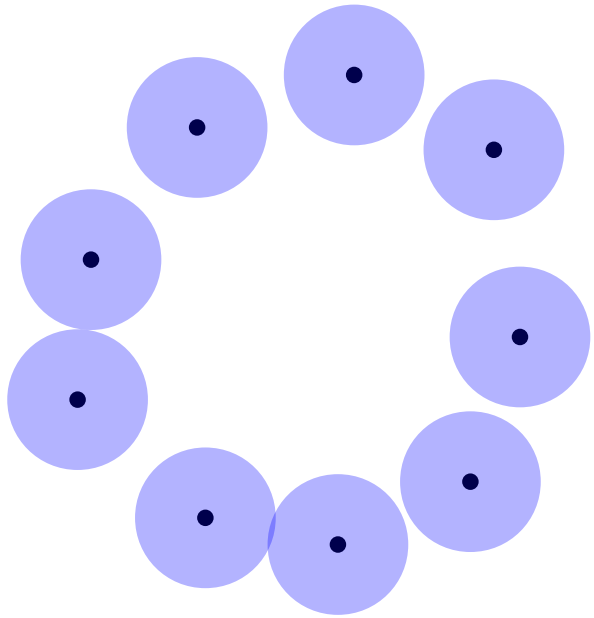


Simplizielle Komplexe: aus diskret wird stetig

Definition

Der Vietoris-Rips-Komplex ist ein abstrakter simplizialer Komplex $VR_\epsilon(X)$ für $\epsilon \in \mathbb{R}_+$ und X:

- Enthält 0-Simplizes (Ecken) entsprechend X
- Enthält $(x_0, \ldots, x_k)$ genau dann, wenn $d(x_i, x_j) < \epsilon$ für alle verschiedenen $0 \leq i, j \leq k$.

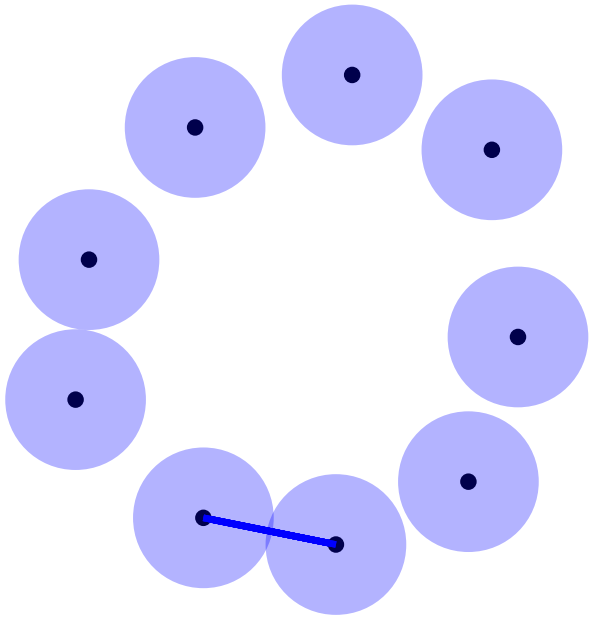


Simplizielle Komplexe: aus diskret wird stetig

Definition

Der Vietoris-Rips-Komplex ist ein abstrakter simplizialer Komplex $VR_\epsilon(X)$ für $\epsilon \in \mathbb{R}_+$ und X:

- Enthält 0-Simplizes (Ecken) entsprechend X
- Enthält $(x_0, \ldots, x_k)$ genau dann, wenn $d(x_i, x_j) < \epsilon$ für alle verschiedenen $0 \leq i, j \leq k$.

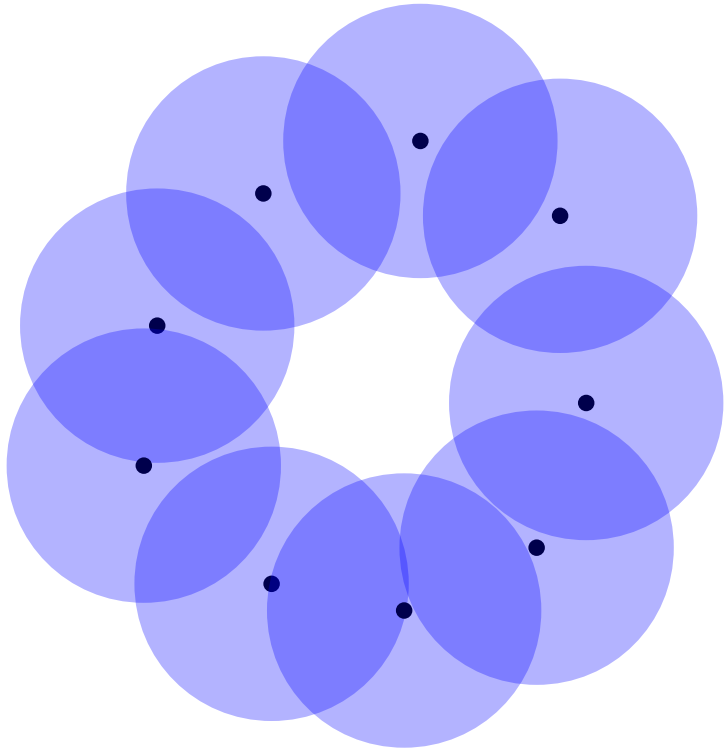


Simplizielle Komplexe: aus diskret wird stetig

Definition

Der Vietoris-Rips-Komplex ist ein abstrakter simplizialer Komplex $VR_\epsilon(X)$ für $\epsilon \in \mathbb{R}_+$ und X:

- Enthält 0-Simplizes (Ecken) entsprechend X
- Enthält $(x_0, \ldots, x_k)$ genau dann, wenn $d(x_i, x_j) < \epsilon$ für alle verschiedenen $0 \leq i, j \leq k$.

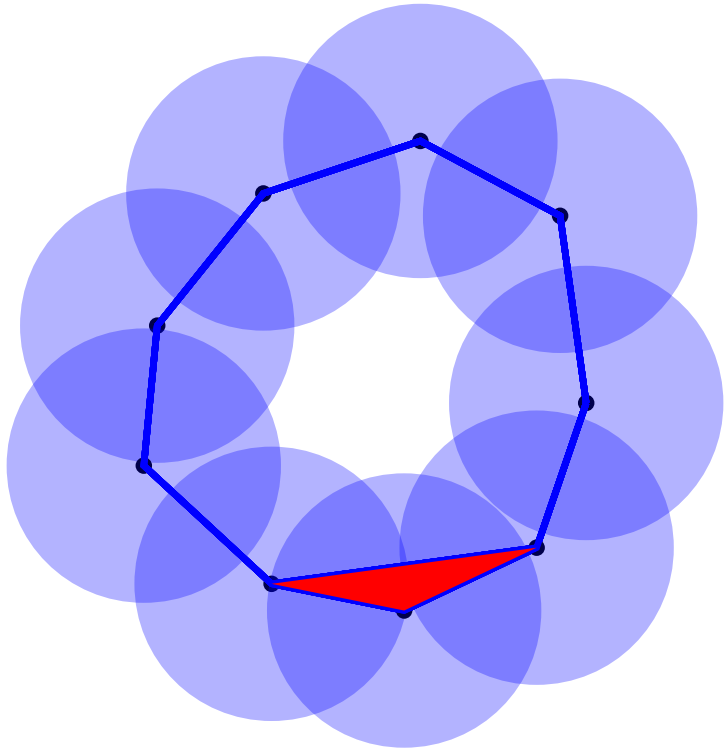


Simplizielle Komplexe: aus diskret wird stetig

Definition

Der Vietoris-Rips-Komplex ist ein abstrakter simplizialer Komplex $VR_\epsilon(X)$ für $\epsilon \in \mathbb{R}_+$ und X:

- Enthält 0-Simplizes (Ecken) entsprechend X
- Enthält $(x_0, \ldots, x_k)$ genau dann, wenn $d(x_i, x_j) < \epsilon$ für alle verschiedenen $0 \leq i, j \leq k$.



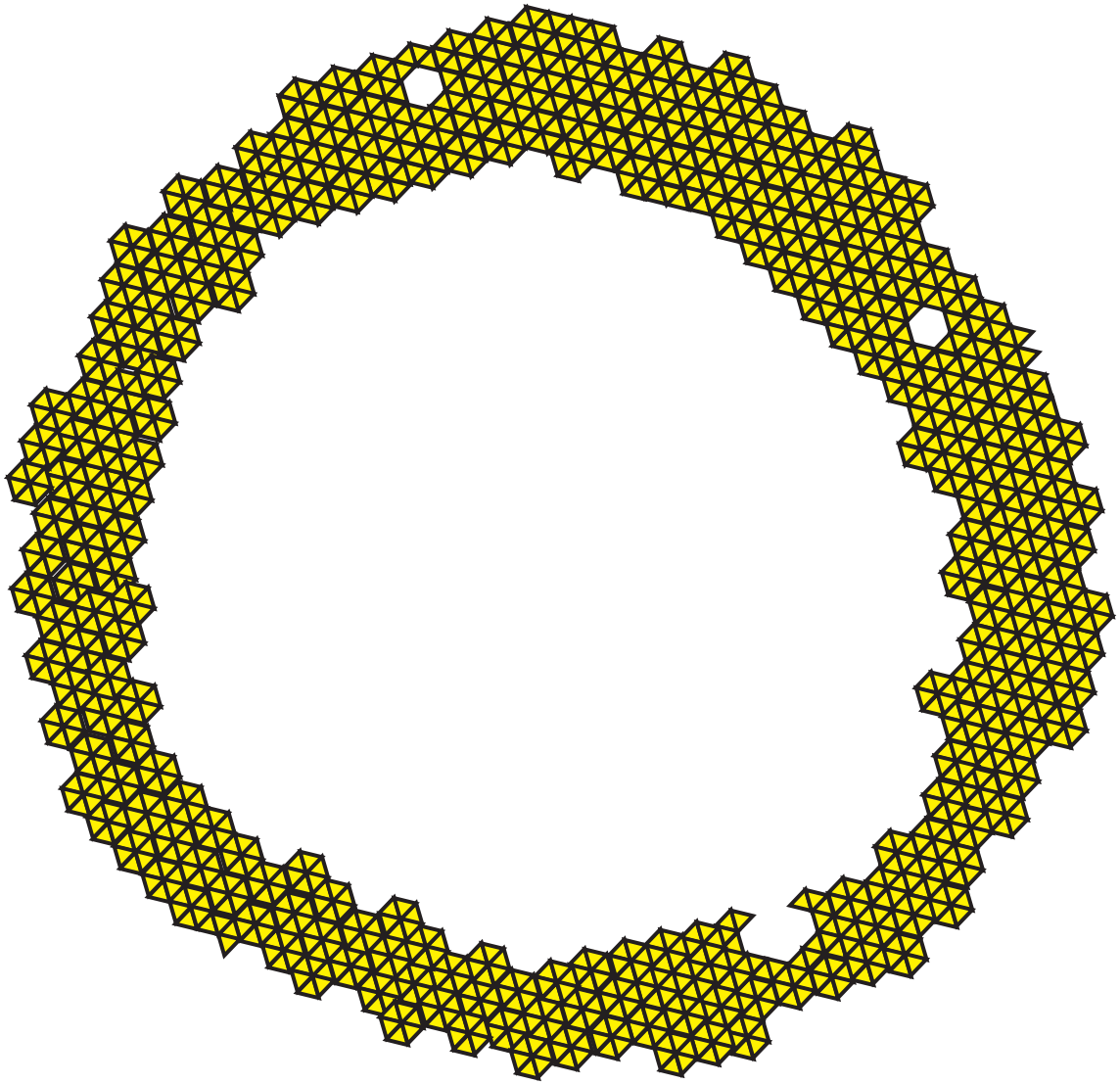
Berechnung von Homologie

Von einem simplizialen Komplex S können wir die Homologie durch Matrixalgebra berechnen:

- S ordnen wir einen Vektorraum $CS = \bigoplus_{\sigma \in S} \sigma \cdot k$ zu.
- CS ordnen wir einen linearen Randabbildung $\partial : CS \to CS$ zu. Auf Simplizes wird ∂ als alternierende Summe der Teilsimplizes auf dem Rand des Simplexes definiert.
- Aus der Algebra (und der Geometrie) folgt $\partial(\partial\sigma) = 0$ für alle Simplizes σ. Daher gilt $\partial(S) \subseteq \ker \partial$.
- Wir definieren formal die Homologie $H(S; k) = \ker \partial / \partial(S)$.
- Beschränkung auf Simplizes der Dimension i ergibt $H_i(S; k)$.

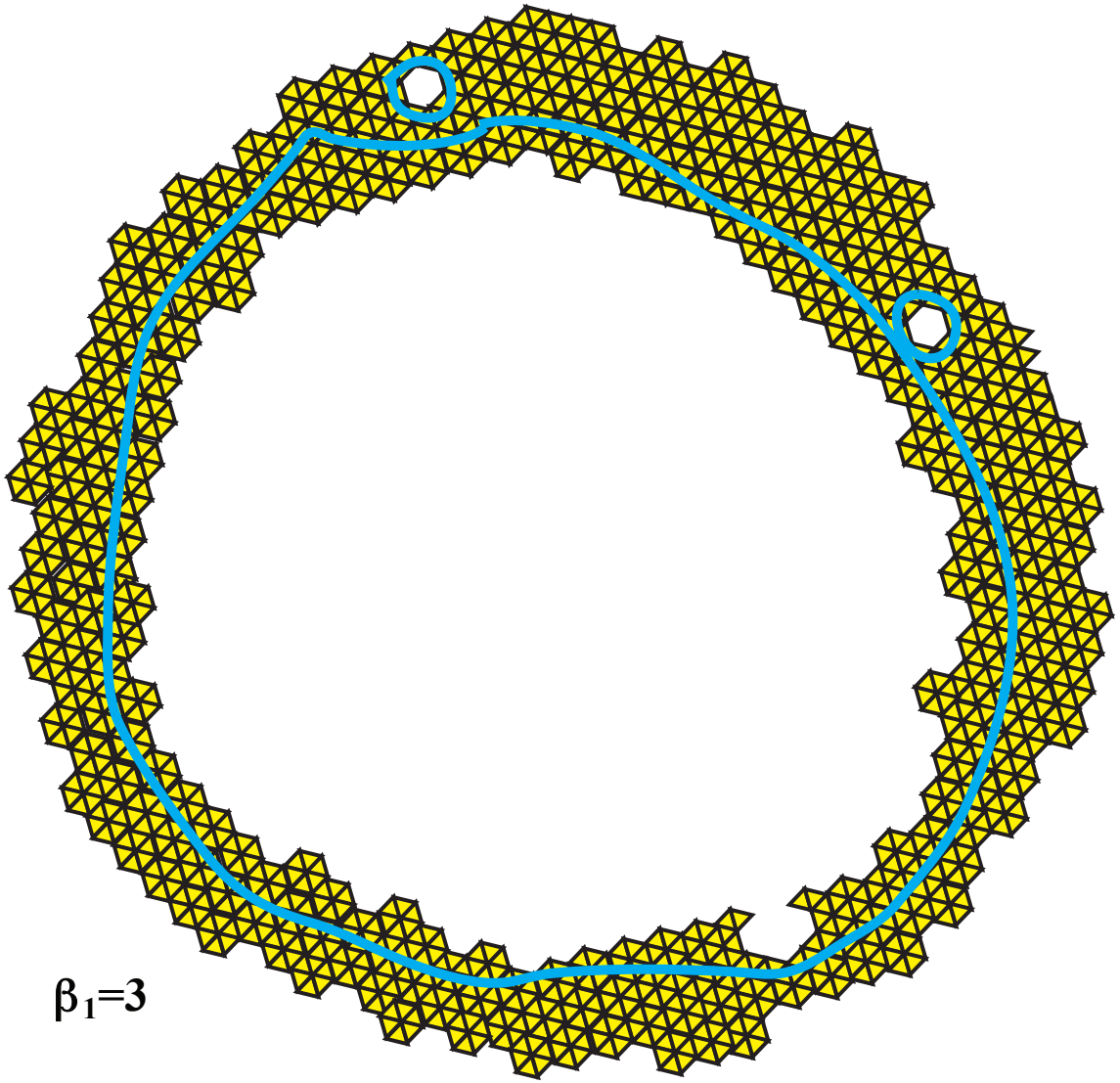
Beispiel: Punktwolke eines Kreises

Erste Idee: Finde gute Werte für ϵ um $VR_\epsilon(X)$ zu studieren.



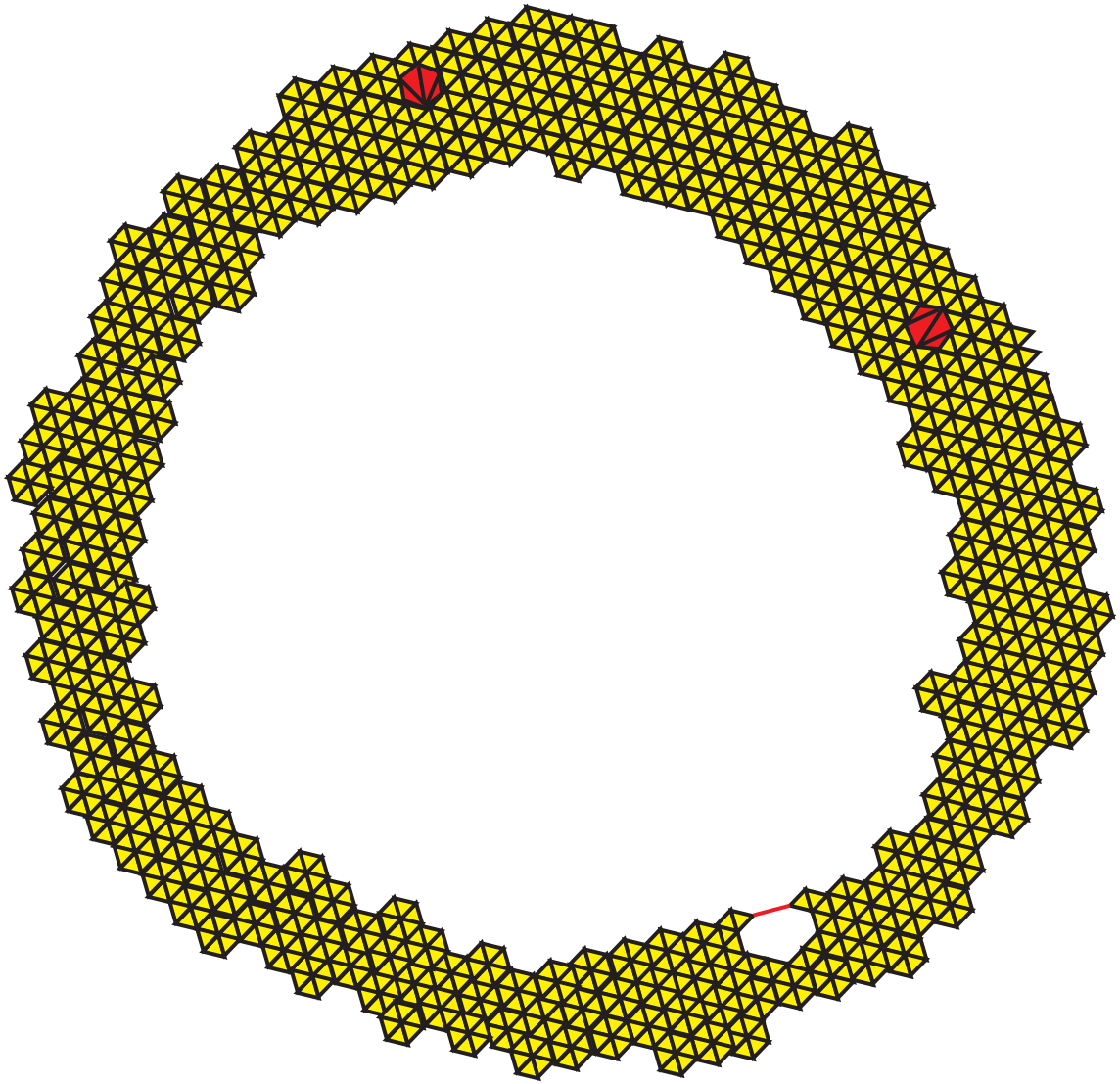
Beispiel: Punktwolke eines Kreises

Erste Idee: Finde gute Werte für ϵ um $VR_\epsilon(X)$ zu studieren.



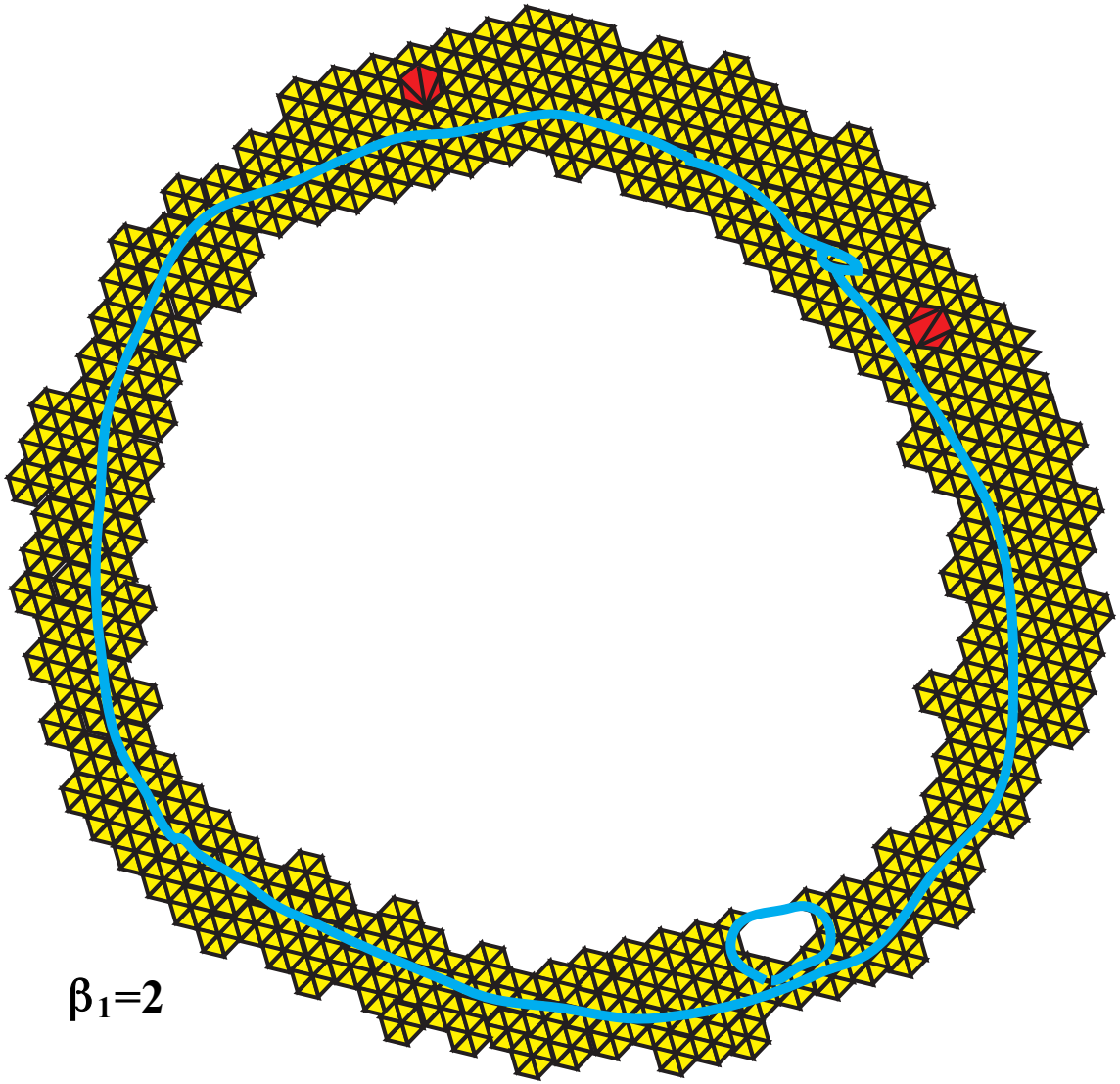
Beispiel: Punktwolke eines Kreises

Erste Idee: Finde gute Werte für ϵ um $VR_\epsilon(X)$ zu studieren.



Beispiel: Punktwolke eines Kreises

Erste Idee: Finde gute Werte für ϵ um $VR_\epsilon(X)$ zu studieren.



Beispiel: Punktenwolke vom Kreis

Bessere Idee: Studiere die Veränderungen in $H_i(VR_\epsilon(X))$ für verschiedene Werte von ϵ.

Wenn $\epsilon < \epsilon'$, dann $VR_\epsilon(X) \subset VR_{\epsilon'}(X)$. Wegen Funktorialität von H_i folgt die Existenz einer Abbildung $H_i(VR_\epsilon(X)) \to H_i(VR_{\epsilon'}(X))$.

Insgesamt ergibt sich ein Diagramm von Vektorräume

$$H_i(VR_{\epsilon_0}(X)) \to H_i(VR_{\epsilon_1}(X)) \to \cdots \to H_i(VR_{\epsilon_k}(X))$$

Solch ein Diagramm nennen wir einen Persistenzraum.

Ein bisschen Algebra

Äquivalenz von Kategorien zwischen Persistenzräume und graduierten $k[t]$-Moduln.

$$V_0 \xrightarrow{\iota} V_1 \xrightarrow{\iota} \ldots \xrightarrow{\iota} V_k \qquad \Rightarrow \qquad \bigoplus_i V_i \quad =: V_*$$

Modul-Struktur (wie Vektorraum, aber Skalare sind Polynome, nicht Zahlen) ist gegeben durch die Angabe, welche Wirkung die Multiplikation mit t hat:

$$t \cdot (v_0, v_1, \ldots, v_k) = (0, \iota v_0, \iota v_1, \ldots, \iota v_{k-1})$$

Ein bisschen Algebra

Graduierte Moduln über $k[t]$ haben eindeutige Zerlegungen (ähnlich der Jordanschen Normalform):

$$V_* = \bigoplus_i t^{a_i} k[t] \oplus \bigoplus_j t^{b_j} k[t]/t^{c_j}$$

Ein bisschen Algebra

Graduierte Moduln über $k[t]$ haben eindeutige Zerlegungen (ähnlich der Jordanschen Normalform):

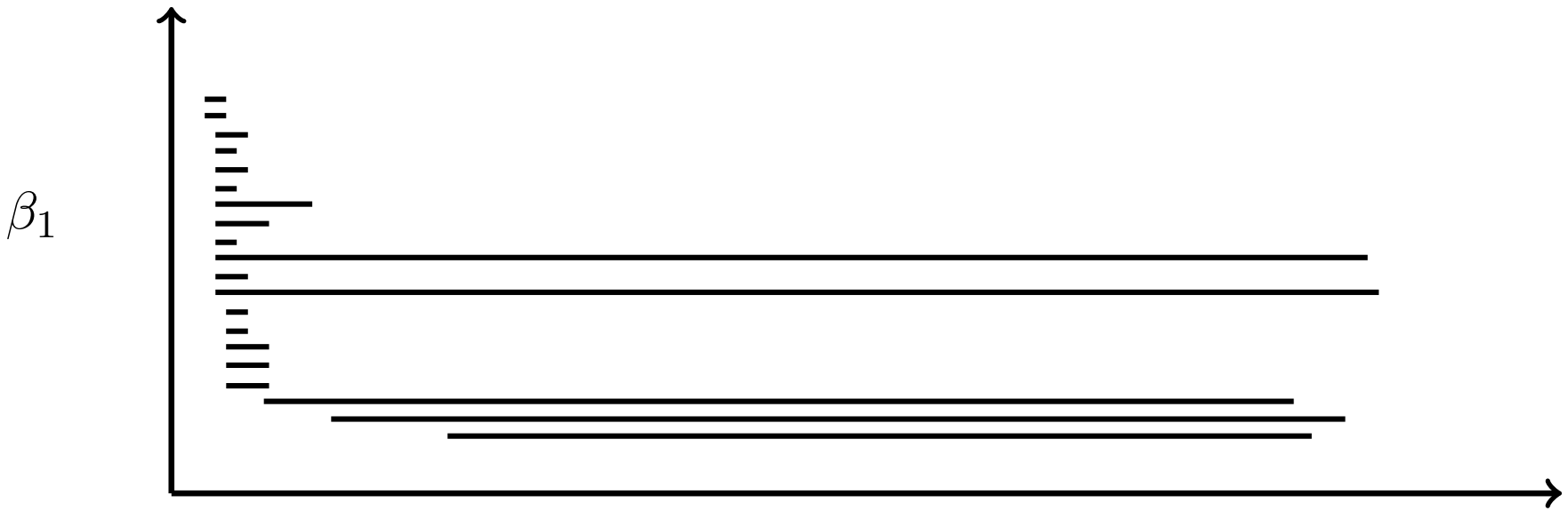
$$V_* = \bigoplus_{i}^{[a_i,\infty)} t^{a_i} k[t] \oplus \bigoplus_{j}^{[b_j,b_j+c_j)} t^{b_j} k[t]/t^{c_j}$$

Ein bisschen Algebra

Graduierte Moduln über $k[t]$ haben eindeutige Zerlegungen (ähnlich der Jordanschen Normalform):

$$V_* = \bigoplus_i t^{a_i} k[t] \oplus \bigoplus_j t^{b_j} k[t]/t^{c_j}$$

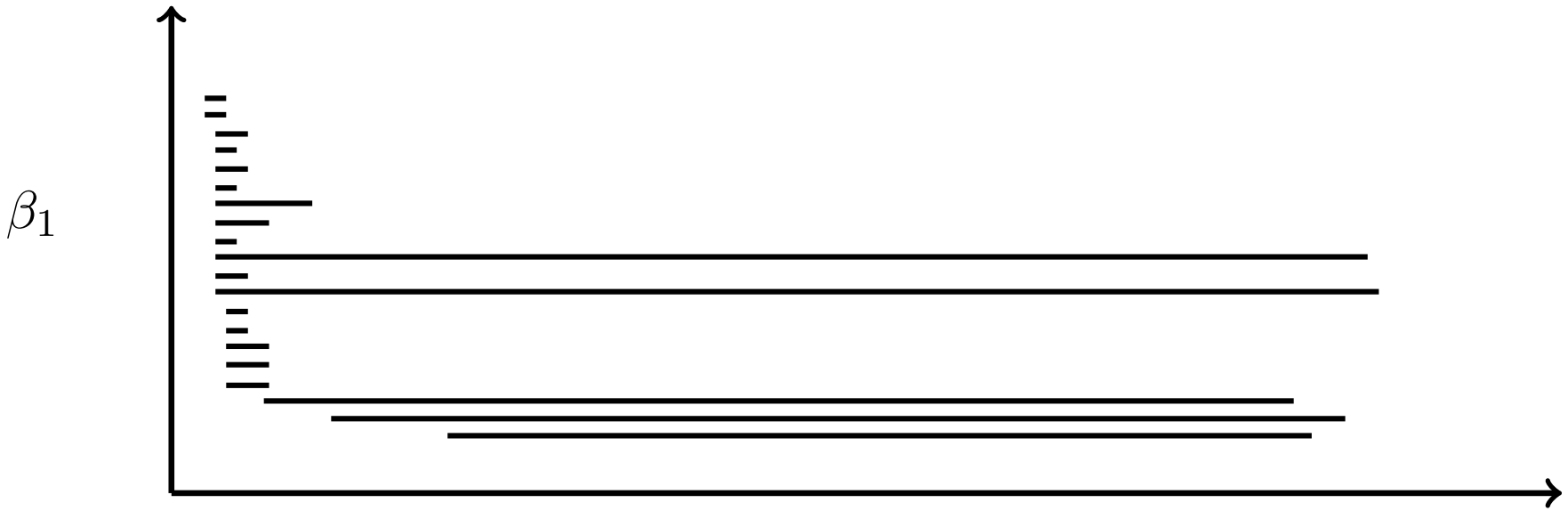
$$[a_i, \infty) \qquad [b_j, b_j + c_j)$$



Deutung der Barcodes

Die Barcodes der Bettizahlen des Vietoris-Rips-Komplex einer Punktwolke sagen uns, welche homologischen Eigenschaften signifikant sind, und welche eher von Rauschen erzeugt sind.

Die Länge eines Intervals entspricht der Größe der dazugehörigen Eigenschaft.



Algorithmik

Grundlage der Berechnung der Betti-Barcodes ist eine Variante von Smith-Normalform der Matrix der Randabbildung von simplizialen Komplexen. Sie ähnelt der Berechnung von Gröbnerbasen.

Input

Eine geordnete Folge von Simplizes, so dass $\partial\sigma$ immer vollständig erzeugt ist, bevor σ entsteht. Jedem Simplex wird eine Zeitkoordinate $t(\sigma)$ zugeordnet.

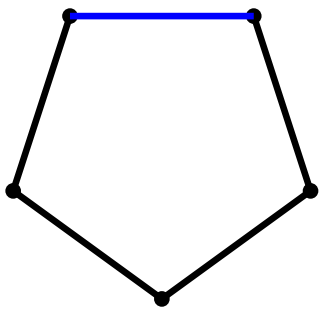
Output

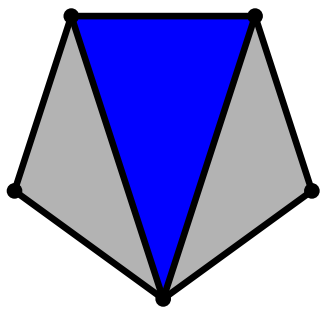
Der Betti-Barcode des Komplexes, d.h. Paare (t_b, t_d), die den Lebenszeiten der Homologieklassen entsprechen.

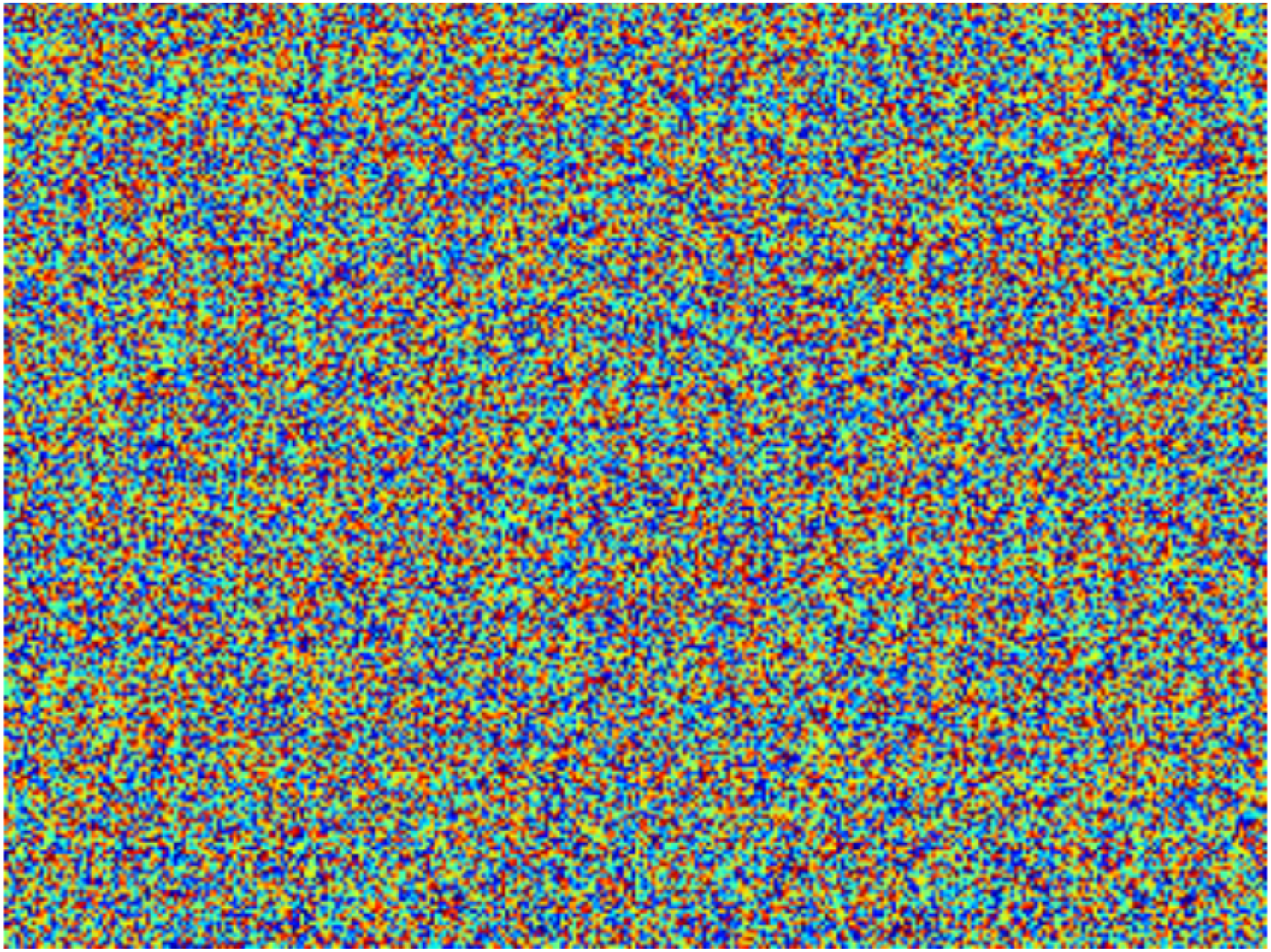
Algorithmik

Grundschritt

1. Neues Simplex σ wird eingeführt.
2. Berechne $\partial\sigma$.
   - 2.1 Ist $\partial\sigma$ ein Zykel (d.h. aufgespannt von den Erzeugern der bisherigen Zykeln), dann repräsentiert dieser Zykel jetzt nicht mehr eine Homologieklasse. Die Zeit $t(f(\partial\sigma))$ des führenden Terms von $\partial\sigma$ bildet zusammen mit $t(\sigma)$ ein neues Intervall.
   - 2.2 Andernfalls ist σ ein neuer Zykel, und wird zur Basis der Zykel hinzugefügt.







Beispiel: Natürliche Bilder

Lee-Mumford-Pedersen fragten, ob es einen statistisch messbaren Unterschied zwischen natürlichen und zufälligen Bildern gibt.

Natürliche Bilder; Teilraum des Raums aller Bilder. Dimension z.B. $640 \cdot 480 = 307\,200$.

Hohe Dimension – denn viele Bilder existieren.

Hohe Kodimension – denn zufällige Bilder sehen ganz anders aus.

Natürliche 3x3 Patches

Anstatt gesamte Bilder zu studieren betrachten wir die Verteilung von 3×3 Pixelpatches.

Die meisten sind etwa konstant. Zu viele – andere Strukturen nicht mehr sichtbar.

Lee-Mumford-Pedersen wählten $8\,500\,000$ Patches mit hohem Kontrast aus einer Sammlung natürlicher Bilder. Jeder 3×3-Patch ist ein Vektor in $\mathbb{R}^9$.

Normalisierte Helligkeit: $\mathbb{R}^9 \to \mathbb{R}^8$. Normalisierter Kontrastwerte: $\mathbb{R}^8 \to S^7$.

Pixelpatches in S^7

Dicht in S^7 – betrachten stattdessen dichte Umgebungen.

Wir betrachten die 25% dichtesten Punkte. Parametrisierte Methode, um die Dichte zu messen:

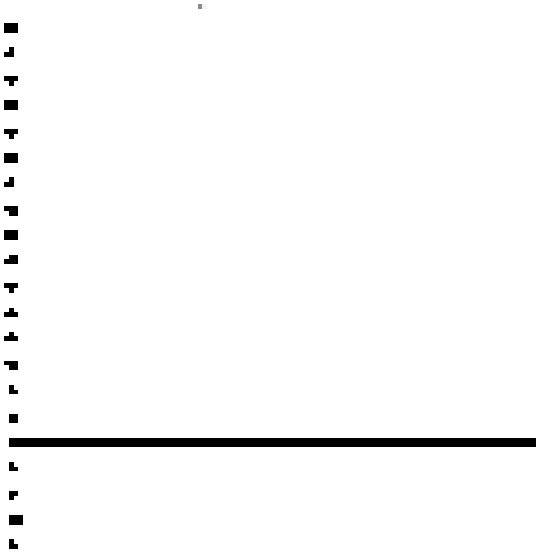
Definition

k-Kodichte $\delta_k(x)$ eines Punktes x ist der Abstand zum k-nächsten Nachbarpunkt.

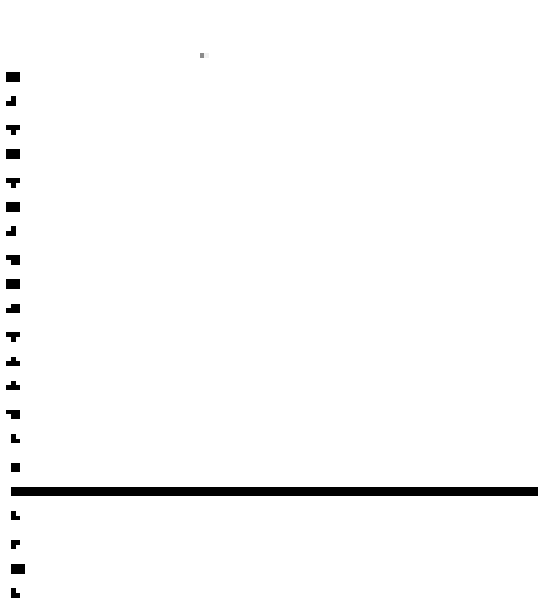
k-Dichte ist $1/\delta_k(x)$.

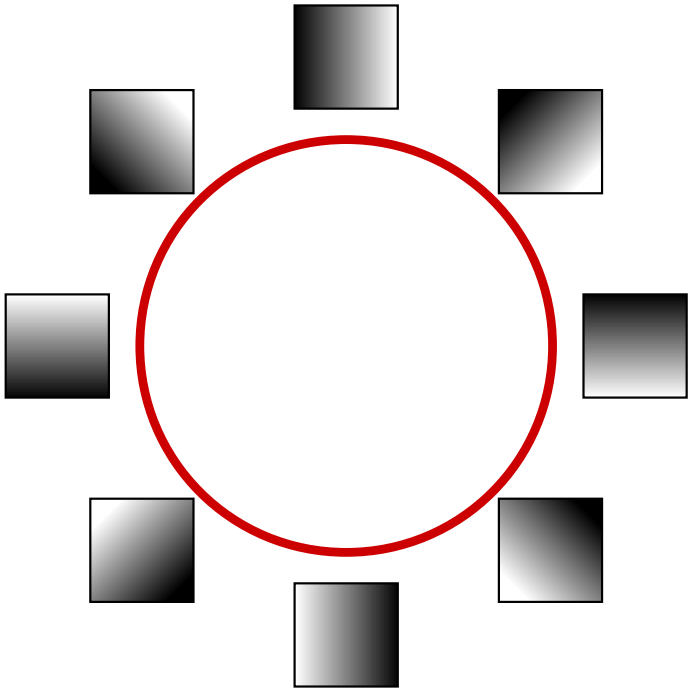
Hohes k ergibt globale Eigenschaften. Niedriges sehr viel lokale Details. Die k-Wert wirkt wie eine Art Schärfeeinstellung.

300-Dichte

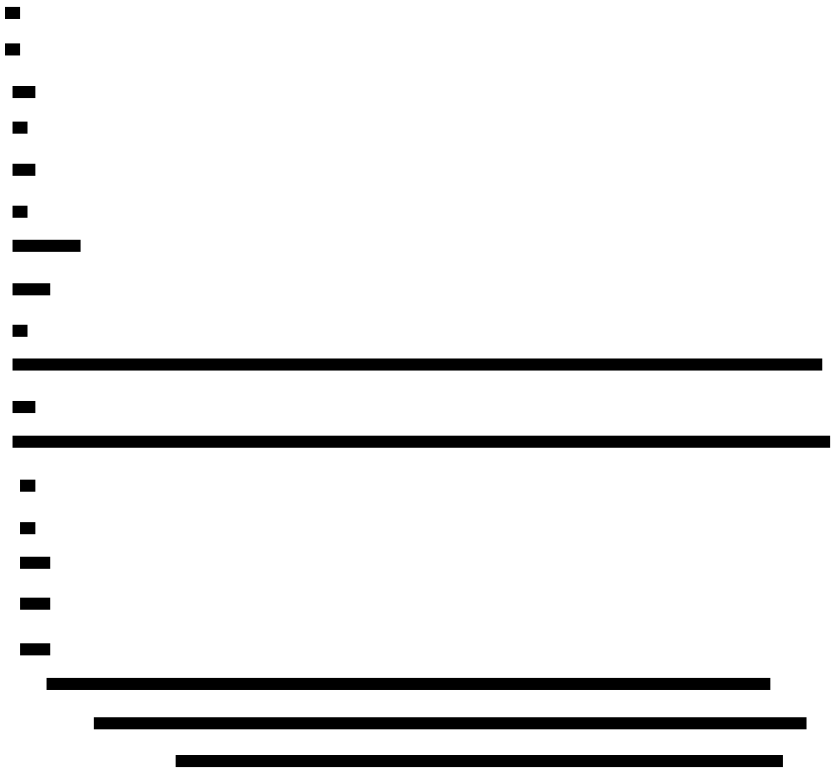


300-Dichte

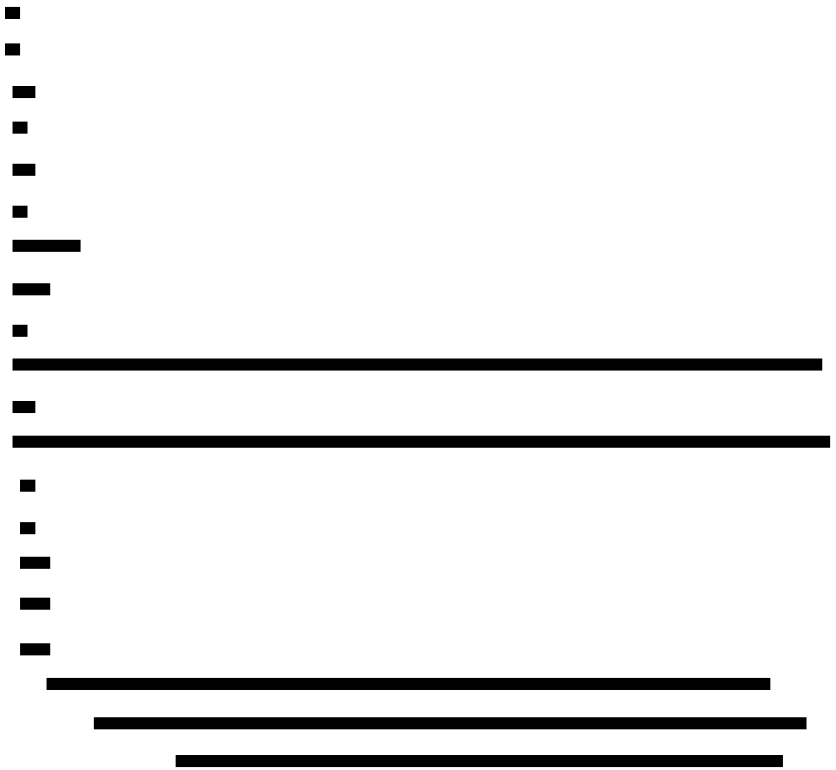


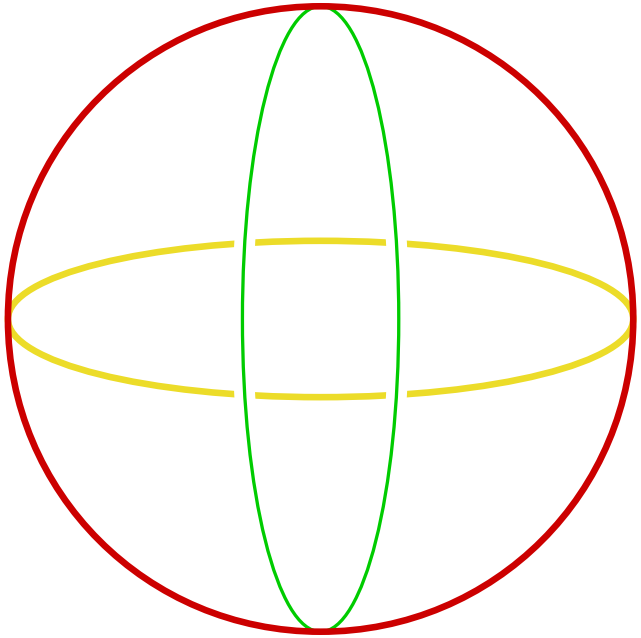


15-Dichte

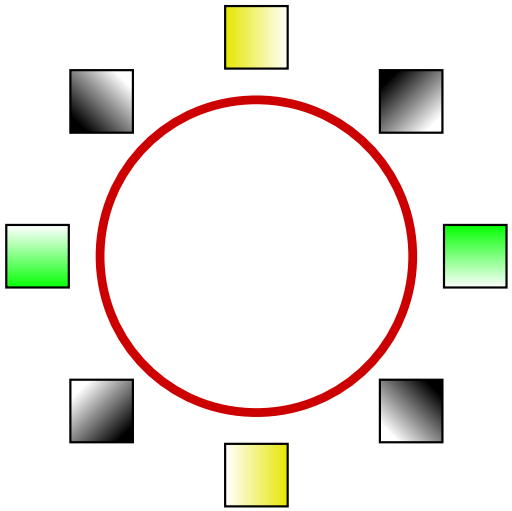


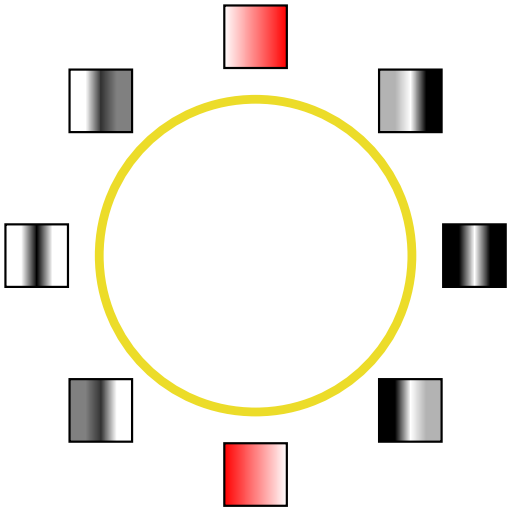
15-Dichte

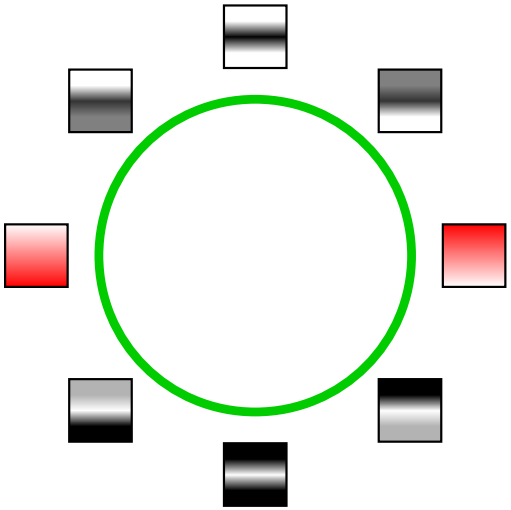




Drei Kreise





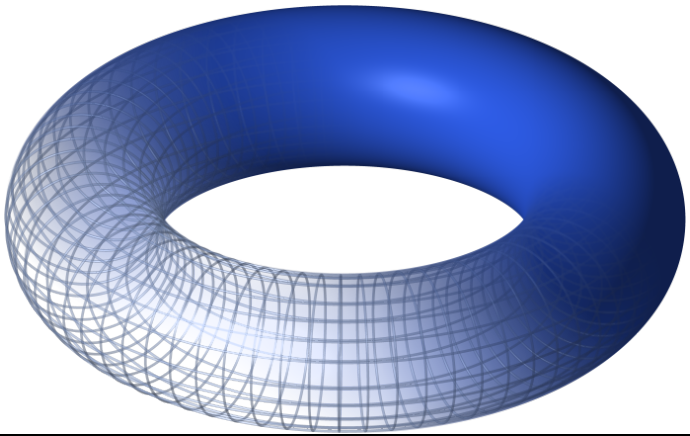


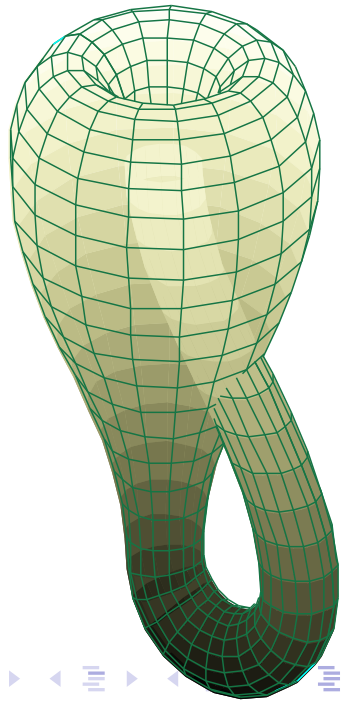
Identifikation des Teilraums der natürlichen Pixelpatches

Einbeziehen von noch mehr Punkten ergibt, mit Koeffizienten in $\mathbb{F}_2$:

$$\beta_0 = 1 \qquad \beta_1 = 2 \qquad \beta_2 = 1$$

Entspricht einem von:



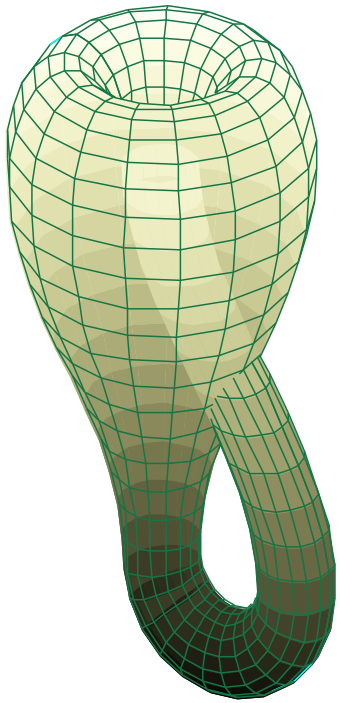


Identifikation der Teilraum der Natürlichen Pixelkluster

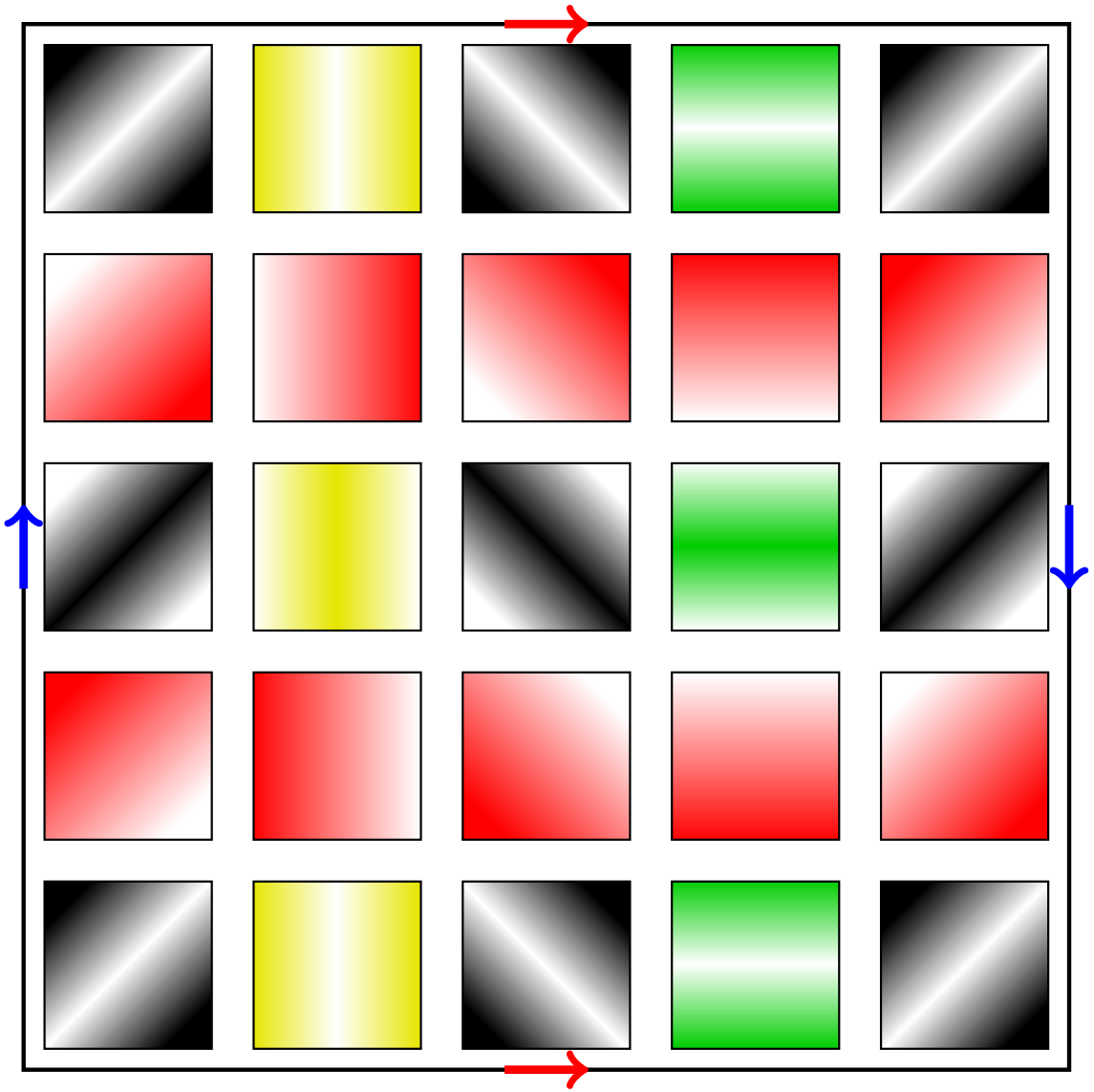
Inklusion von mehr Punkte ergibt, mit Koeffizienten in $\mathbb{F}_3$:

$$\beta_0 = 1 \qquad \beta_1 = 1$$

Entspricht insgesamt:



Kleinsche Flasche der Pixelkluster



Anwendungen dieser Analyse

Bildkomprimierung

Ein 3×3-Kluster wird ziemlich genau beschrieben durch 4 Dimensionen:

- Projektion auf die Kleinsche Flasche
- Eigentliche Helligkeit
- Eigentlicher Kontrastwert

Texturanalyse

Texturen ergeben Verteilungen auf der Kleinsche Flasche. Drehung der Textur ist Verschiebung der Verteilung.