Vorlesung Knowledge Discovery,
M. Thrun

Hochdimensionale Daten

Databionics Research Group

Philipps Universität Marburg

Daten im R^n

Geg:

d Datensätze der Dimension n, x_i aus R^n

Sagt sich leicht, aber welche Eigenschaften hat der R^n?

s.a.: Verleysen, Werz et. Al. *On the effects of dimensionaility on data analysis*, 2003

Eigenschaften

1. Empty space phenomenom
2. Concetration of measure phenomenem
3. Curse of dimensionality

Volumen im R^n: „empty space“

Eine Eigenart hochdimensionaler Räume ist , dass das

Volumen einer Hyperkugel

sich praktisch vollständig

in einer nahezu beliebig dünnen Schale "unterhalb" der Oberfläche befindet.

„Concetration of measure“

Für n-> unendlich geht

$$\frac{\max(d(x,y)) - \min(d(x,y))}{\min(d(x,y))} \text{-> } 0$$

- Es gibt kaum einen Unterschied zwischen kleinster und größter euklidischer Distanz!

Folgerung: „Concetration of measure“

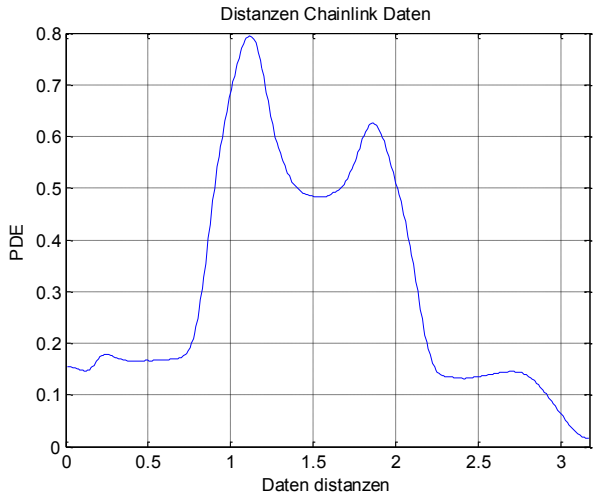
- Mit höher werdender Dimension, tendieren euklidische Distanzen zwischen beliebigen Punkten gegen eine Konstante

=> Hochdimensionale Punkte sind alle in etwa gleich weit von einander entfernt

Der R^n ist leer

Um die Verteilung 1-Dim Daten auszuloten sollte man wie viele Daten besitzen?

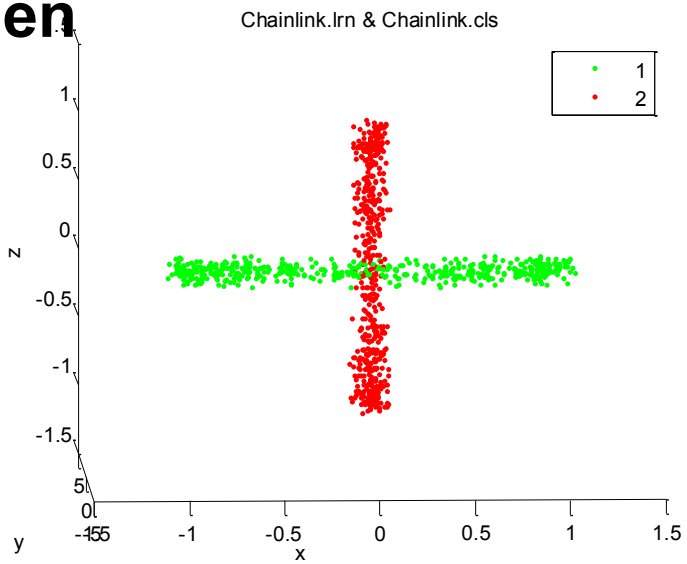
Beispiel:



Um diese Verteilung auszuloten wären sicherlich mindestens 100 Datensätze wichtig.

Kombinationen

Jede Dimension kann mit jeder anderen interessante Kombinationen bilden



d.h. idealerweise wären 100^n Daten nötig

Dimension	Gew. Datensätze	100	1000	10000	100000
n	100^n	Datensätze s	Datensätze s	Datensätze s	Datensätze si
1	100	100%	1000%	10000%	100000%
2	10000	1%	10%	100%	1000%
3	1.00E+06	0.01%	0.10%	1.00%	10.00%
4	1.00E+08	0.00%	0.00%	0.01%	0.10%
5	1.00E+10	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
6	1.00E+12	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
7	1.00E+14	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
8	1.00E+16	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
9	1.00E+18	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
10	1.00E+20	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
11	1.00E+22	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
12	1.00E+24	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
13	1.00E+26	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
14	1.00E+28	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
15	1.00E+30	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
16	1.00E+32	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
17	1.00E+34	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
18	1.00E+36	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
19	1.00E+38	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill
20	1.00E+40	<0.1 Promill	<0.1 Promill	<0.1 Promill	<0.1 Promill

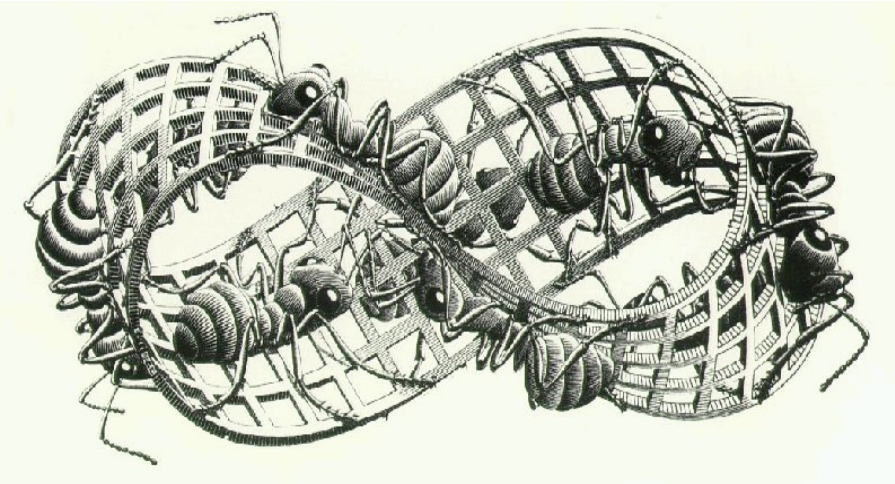
„curse of dimensionality“

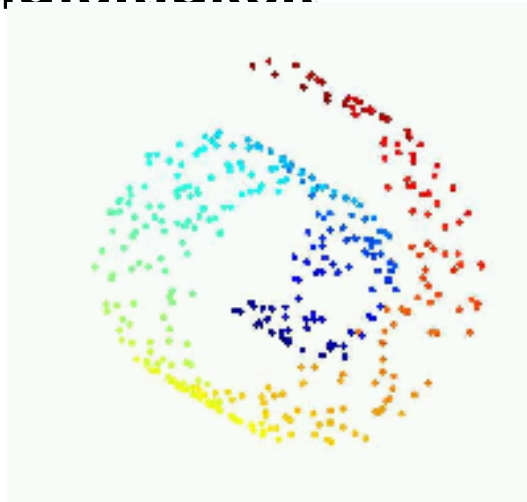
- Um genügend genau zu sein, müssten wir O(exp(n))
- Daten haben
- Wir haben fast immer (ab n=4) zu wenig Daten!

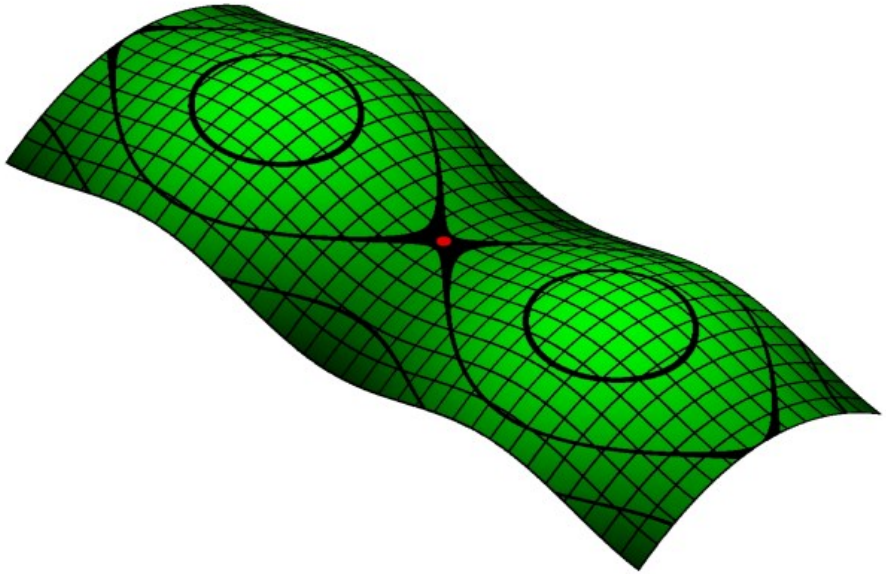
- Der R^n ist in der Regel leer

Hoffnung: Daten liegen auf einer Unter-Mannigfaltigkeit

Mannigfaltikeit (Manifold)Mannigfaltigkeit







„swiss roll“

Ein endlicher und hoffentlich begrenzter Unterraum, auf dem die Daten liegen

Mögliche Ansätze

- Eine kluge Wahl einer nicht euklidischen Distanz
  - s. Verleysen 2003 et al.
- Dimension der Untermannigfaltigkeit abschätzen:
  - s. Vorlesung über intrinsische Dimension
- **Daten projizieren R^n -> R^2**
  - **Problem: Es ist nicht möglich alle Distanzen in einem Projektionsverfahren zu erhalten**

Topologie Erhaltung

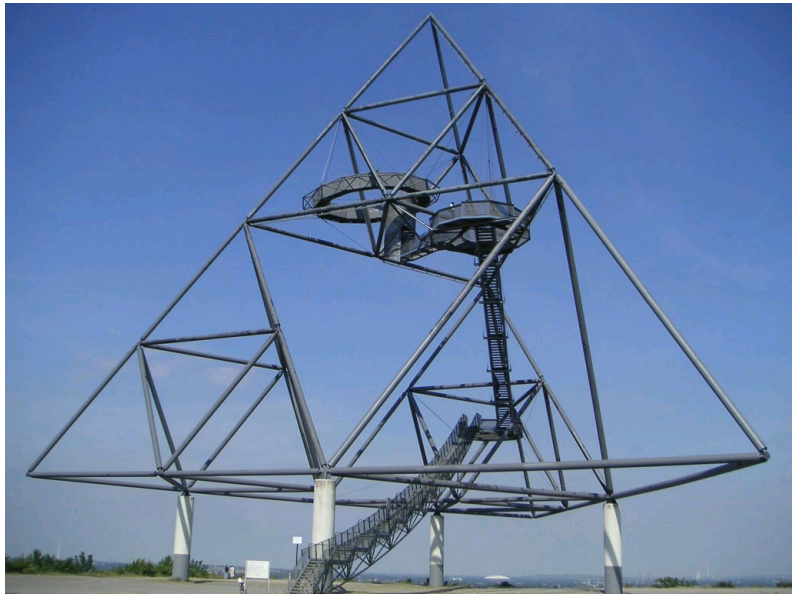
Naiv: Erhaltung der Nachbarschaften

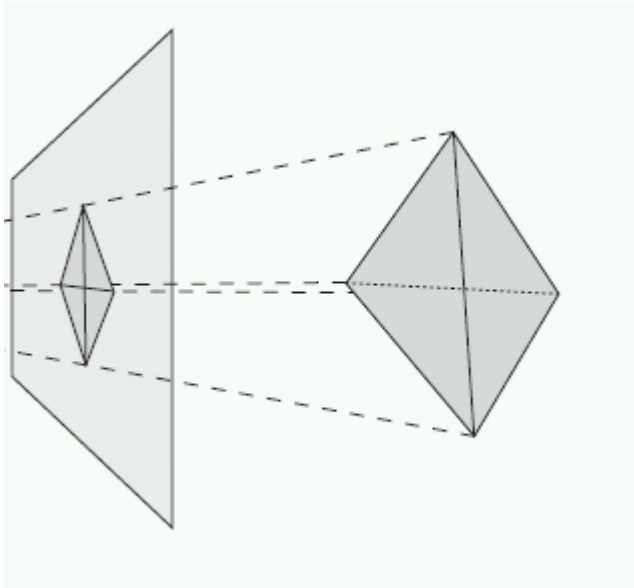
Sind Datenpunkte im hochdimensionalen Eingaberaum nahe bei einander (fern),

so liegen sie auch im niederdimensionalen Ausgaberaum nahe bei einander (fern).

Grundsätzliches

Bei einer Projektion R^N ->R^m können NIE alle Nachbarschaften perfekt erhalten werden





Masszahlen für die topograpischen Fehler einer Projektion

Appellation	Type of measurement (locus of preservation)	Topology	Errors specified
C measure	distance, (global)	Euklid graph	BPE, FPE
Kaski's Trustworthiness and Discontinuity (T&D)	ranks of distances, (local)	KNN graph	BPE, FPE
Force Approach Error	distance, (global)	Euklid graph	BPE, FPE, Gaps
Local Continuity Meta Criterion (LCMC)	surrounding, (global)	KNN graph	BPE, FPE
Minimal Pathlength (C measure)	distance, (local)	Euklid graph, KNN graph, knn=1	BPE
Minimal Wiring (C measure)	distance, (local)	Euklid-Graph, KNN-graph, knn=1	FPE
Mean Relative Rank Errors (MRRE)	ranks of distances, (local)	KNN graph	BPE, FPE
Overall Correlation: Topological Correlation	surrounding, (global)	Delaunay graph	No distinction between FPE and BPE
Overall Correlations: MTP/TC	ranks of distances, (global)	Euklid graph	No distinction between FPE and BPE
Stress (nonmetric MDS)	ranks of distances, (global)	Euklid graph	BPE, FPE
Topographic Function (TF)	surrounding, (local)	Delaunay graph	BPE, FPE, Gaps
U ranking	surrounding, (local)	Delaunay graph, Euklid graph	FPE, Gaps
Zrehen's measurement	surrounding, (local)	Gabriel graph	BPE, Gaps

Table 1: An overview of common quality measurements. – Topographic Product, Topographic Error and Quantization Error are omitted

Still missing is K measure