Technische Universität Braunschweig

Medizinische Hochschule Hannover

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Dimensionsreduktion

Sarah Nee

Peter L. Reichertz Institut für Medizinische Informatik
der TU Braunschweig und der Medizinischen Hochschule Hannover
Sarah.Nee@plri.de, www.plri.de

Dimensionsreduktion

- Visualisierung
- Clusteranalyse
- Rauschunterdrückung
- Embedding
- Methoden
  - Feature Extraction
    - Extraktion (weniger) neuer Feature
  - Feature Selection
    - Auswahl der wichtigsten Feature

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Übersicht

- Fluch der Dimensionalität
  - Data sparsity
  - Distanzen
- (Wiederholung) PCA
  - Grundgedanke
  - Mathe
  - Vorgehen
- t-SNE
  - Grundgedanke
- Embeddings
  - UMAP
  - word2vec

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität

PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

Fluch der Dimensionalität

- Große Mengen Daten, viele Feature/Attribute/Variablen
- n:= Anzahl Samples, f := Anzahl Feature
  - f >> n
- Datensparsamkeit
  - Nur wenige mögliche Ausprägungskombinationen im Datensatz
  - Generalisierbarkeit, Randfälle
- Distanzen
  - Verlieren Bedeutung
  - Große Distanz zwischen allen Punkten
    -> alle gleich weit voneinander entfernt

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

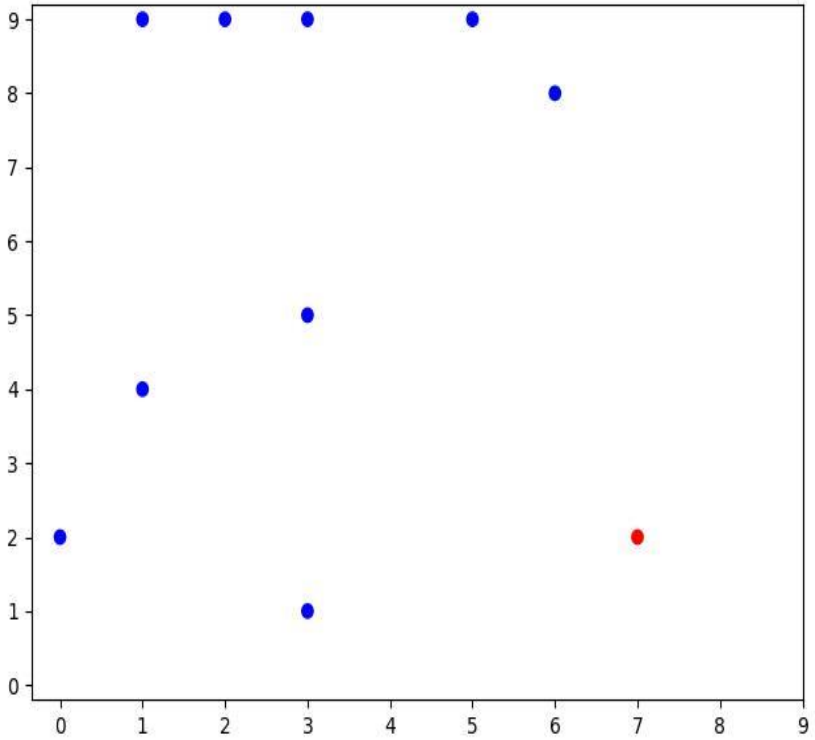
Fluch der Dimensionalität - Data sparsity (1)

- Beispiel:
  - Eine Dimension
  - $x \in [0, 9]$
  - Gleichverteilung

PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

Fluch der Dimensionalität - Data sparsity (2)

- Beispiel:
  - Zwei Dimensionen
  - $x \in [0, 9]^2$
  - Gleichverteilung

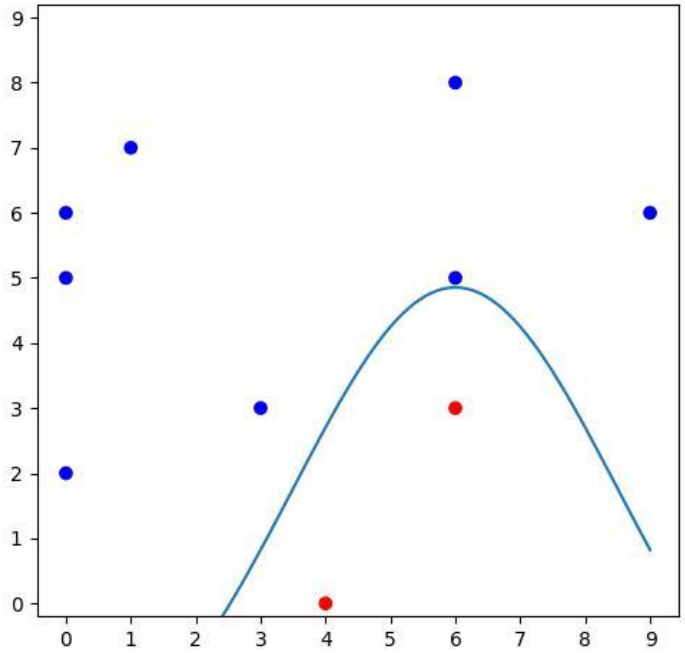


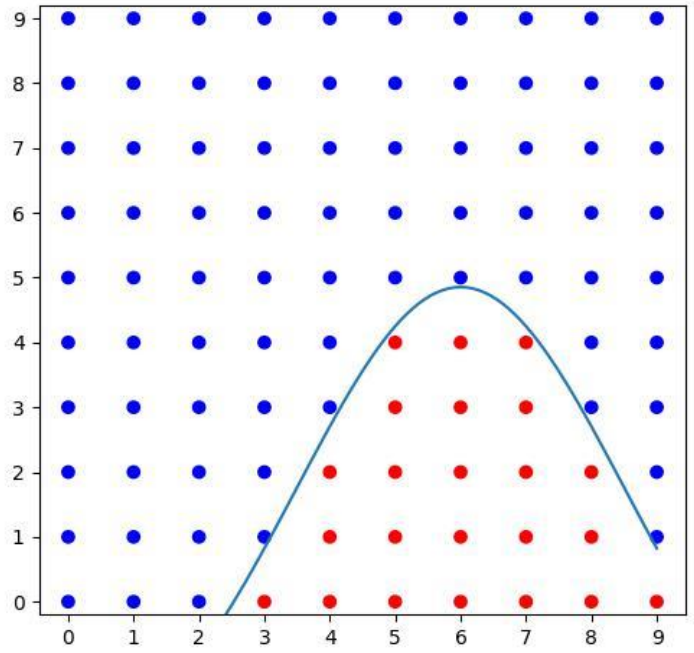
PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

Fluch der Dimensionalität - Data sparsity (3)

- Wo verläuft die Grenze?
- Wie viele Punkte notwendig für ähnliche "Abdeckung" wie in einer Dimension?

  -> n^d

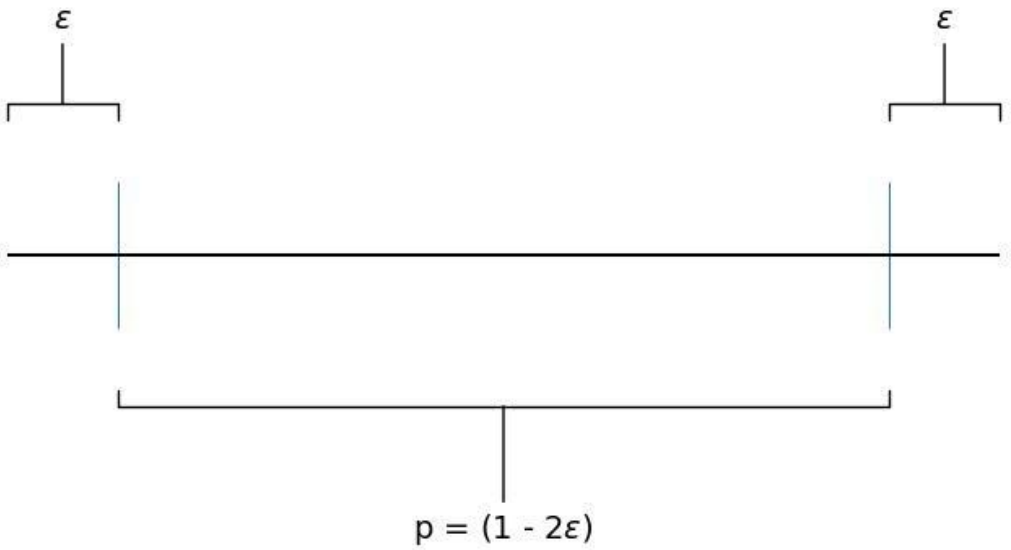




PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität - Distanzen (1)

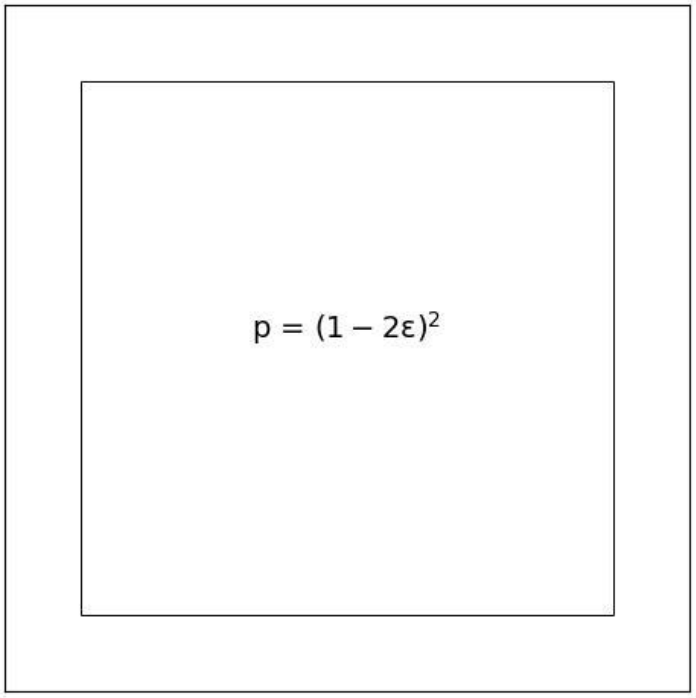
- Annahme: Gleichverteilung, Linie der Länge 1
- p := Wahrscheinlichkeit, **nicht** am Rand zu liegen



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität - Distanzen (2)

- Annahme: Gleichverteilung, Würfel mit Seitenlängen 1
- p := Wahrscheinlichkeit, **nicht** am Rand zu liegen



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität - Distanzen (3)

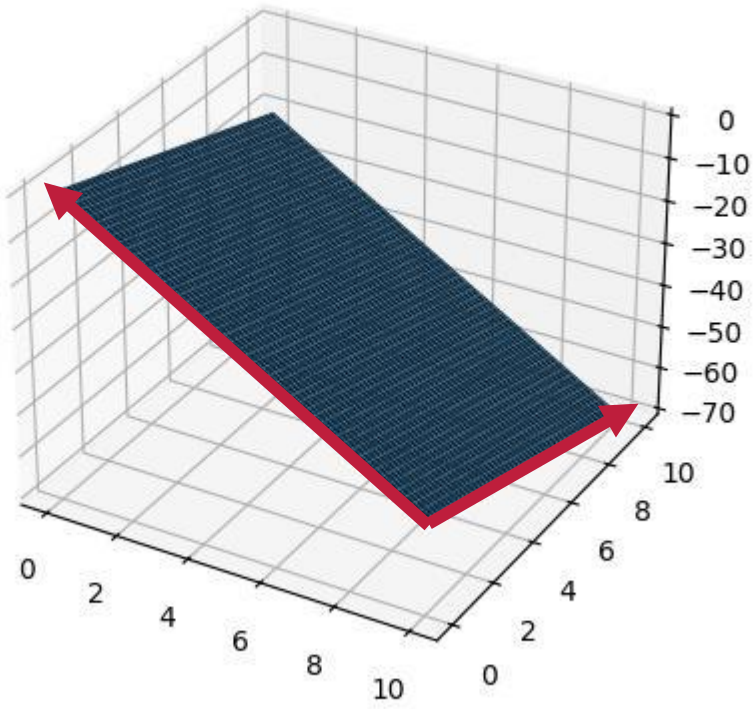
- Wahrscheinlichkeit im "Inneren" zu liegen für ε = 0.1 und Dimensionen d:

d	p
1	0.8000
2	0.6400
3	0.5120
4	0.4096
5	0.3277
6	0.2621
7	0.2097
8	0.1678
9	0.1342
10	0.1074

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität - Unterräume

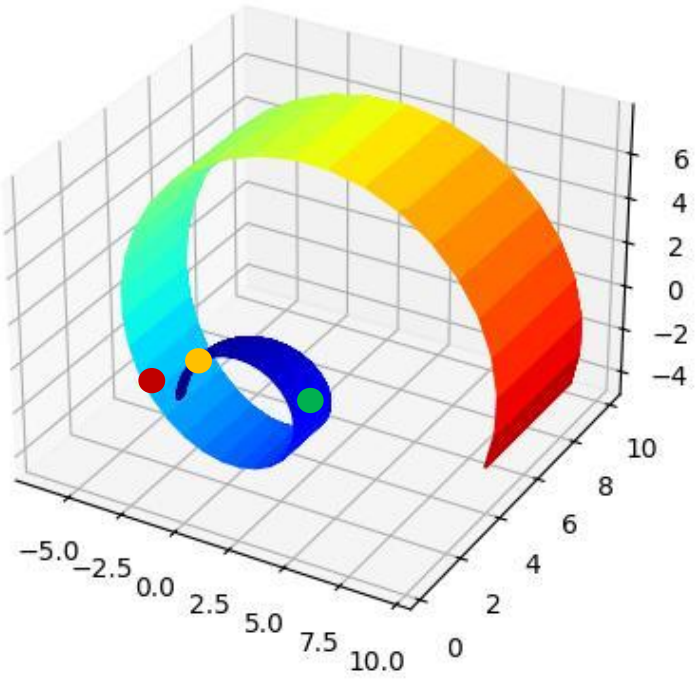
- Daten können in niedriger dimensionale Unterräume liegen
- Im Ursprungsraum alle Dimensionen notwendig
- Neue Achsen beschreiben Punkte eindeutig mit weniger Dimensionen



PLRI
PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität - Mannigfaltigkeit

- Daten liegen auf Oberfläche, die sich durch/über alle Dimensionen "schlängelt"
- Lokal euklidisch, global nicht



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Fluch der Dimensionalität - Zusammenfassung

- Data sparsity
  - Generalisierbarkeit
- Distanzen
  - Distanzmetriken
- Unterräume, Mannigfaltigkeiten
  - Daten oft lokal niedrig-dimensional
- Dimensionsreduktion

Hauptkomponentenanalyse

PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

Hauptkomponentenanalyse

- Principal Component Analysis (PCA)
- Annahme
  - Variablen mit große Varianz enthalten die meiste Information
  - Korrelierte Variablen messen Ähnliches
- Hauptkomponenten
  - Linearkombinationen der ursprünglichen Variablen
  - Sortiert nach Varianz ("Informationsgehalt")
- Dimensionsreduktion
  - Nur die wichtigsten Hauptkomponenten behalten

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Kovarianz

- Varianz
  - mittlere quadratische Abweichung vom Mittelwert

$$Var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$$

- Kovarianz
  - Maß für linearen Zusammenhang
  - Vom Maßstab der Daten abhängig
  - Korrelation: standardisiertes Maß, ρ ∈ [-1, 1]

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Kovarianzmatrix:

$$\begin{pmatrix} Var(x_1) & Cov(x_1,x_1) & \dots & Cov(x_1,x_m) \\ Cov(x_2,x_1) & Var(x_2) & \dots & Cov(x_2,x_m) \\ \dots & \dots & \dots & \dots \\ Cov(x_m,x_1) & Cov(x_m,x_2) & \dots & Var(x_m) \end{pmatrix}$$

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Eigenwerte und Eigenvektoren

- Eigenwertproblem: $Mv = \lambda v, v \neq 0$
  - v wird nur skaliert, Richtung ändert sich nicht!
  - Eigenvektoren v
  - Eigenwerte λ

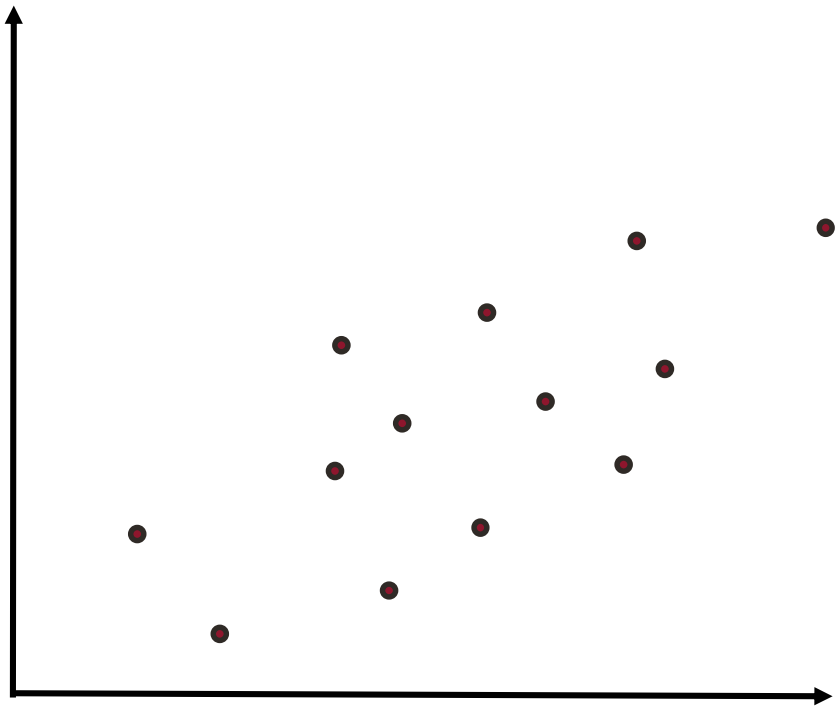
- Berechnung der Eigenwerte und –Vektoren

$$Mv = \lambda v$$
$$Mv = \lambda I v$$
$$(M - \lambda I)v = 0$$
$$\det(M - \lambda I) = 0$$

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (1)

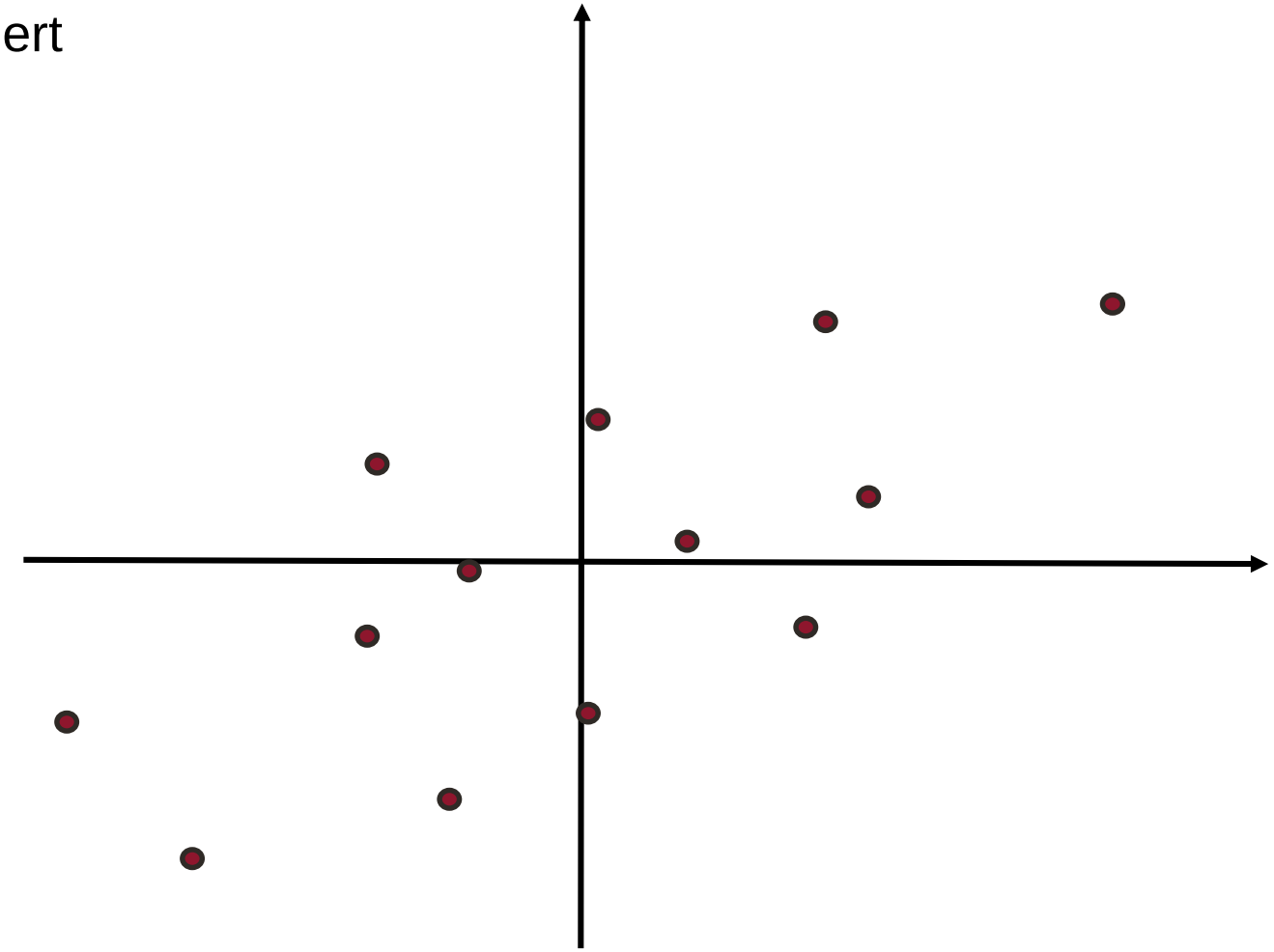
- Korrelierte Punkte



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (2)

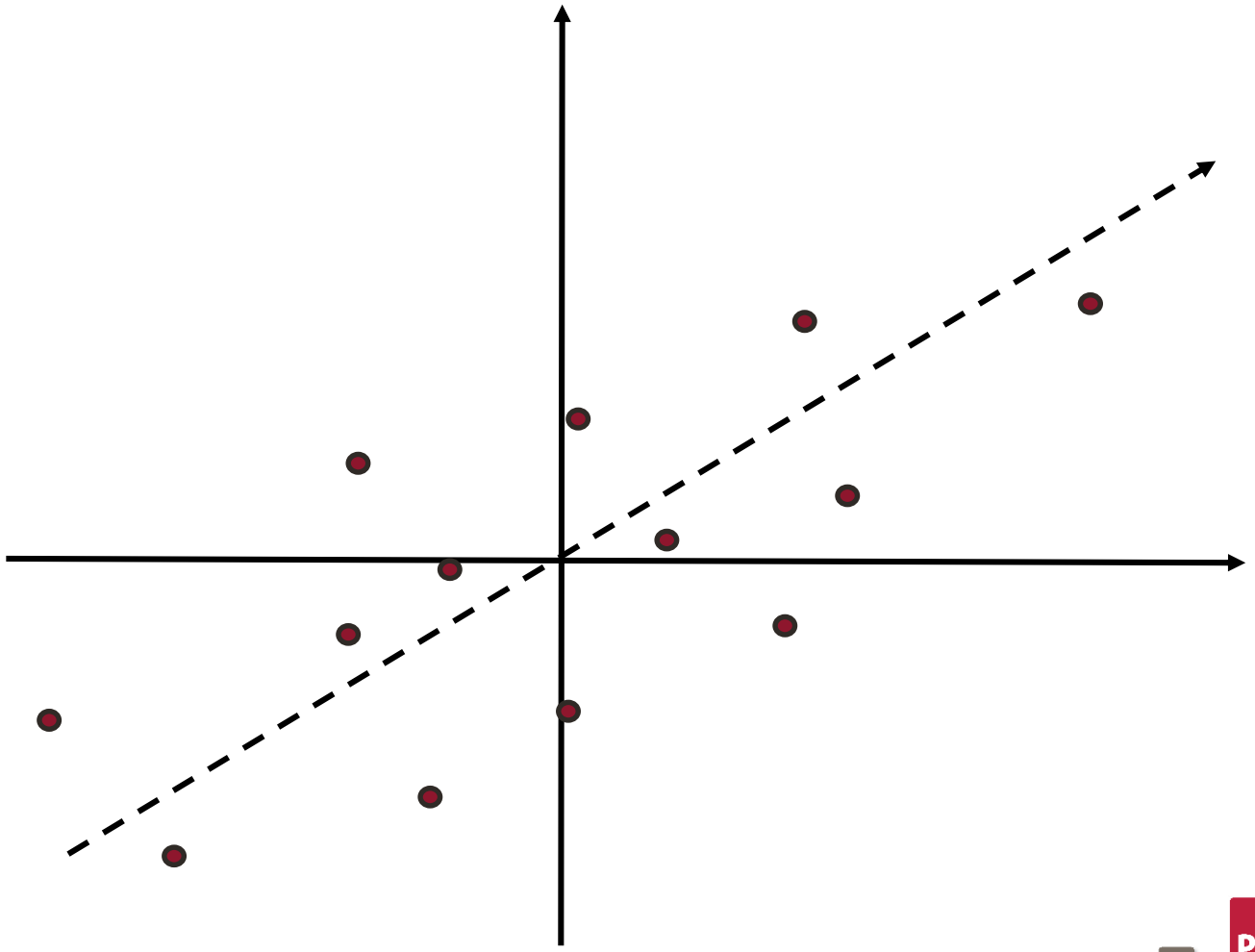
- Zentriert



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (3)

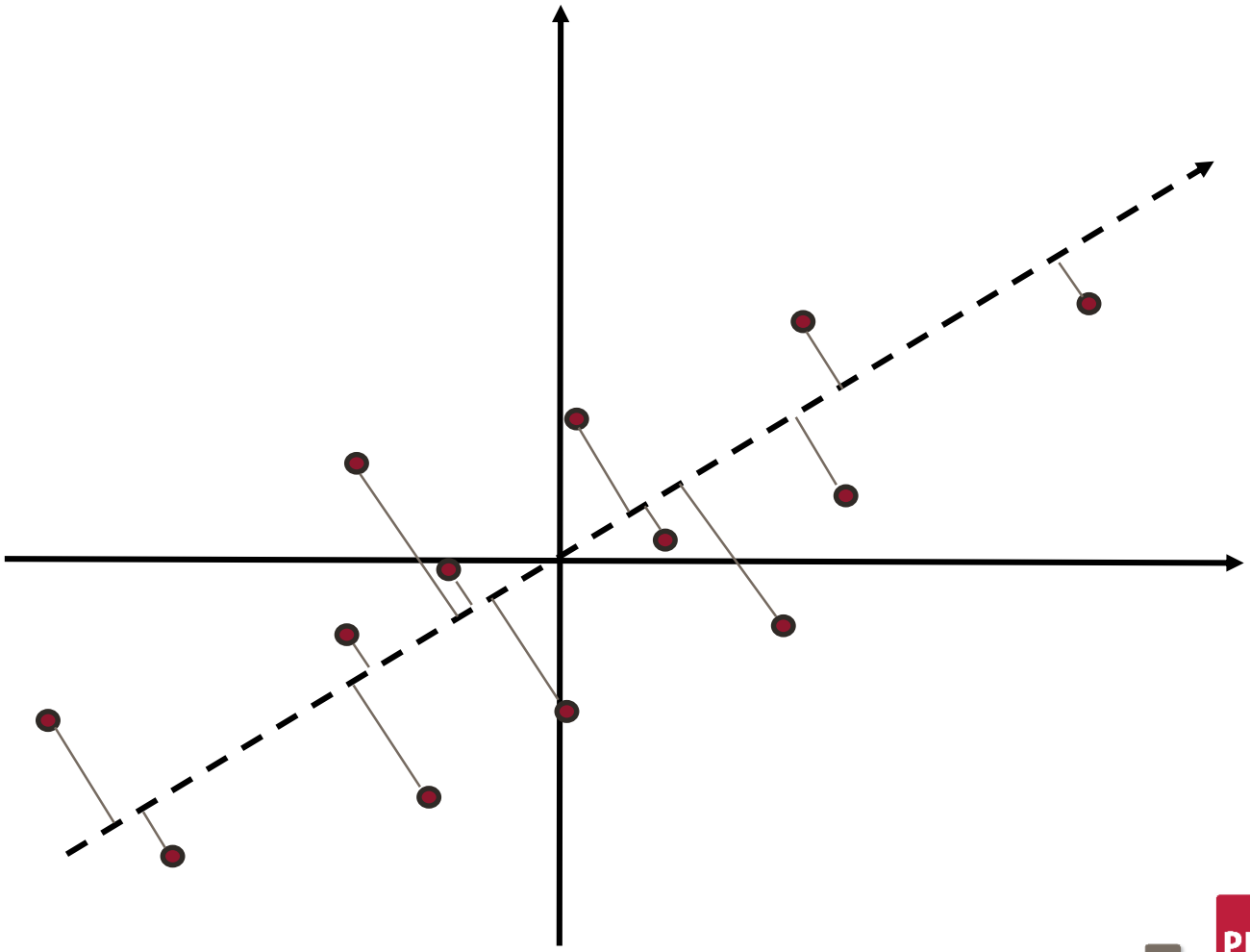
- PC1: Richtung mit der größten Varianz



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (4)

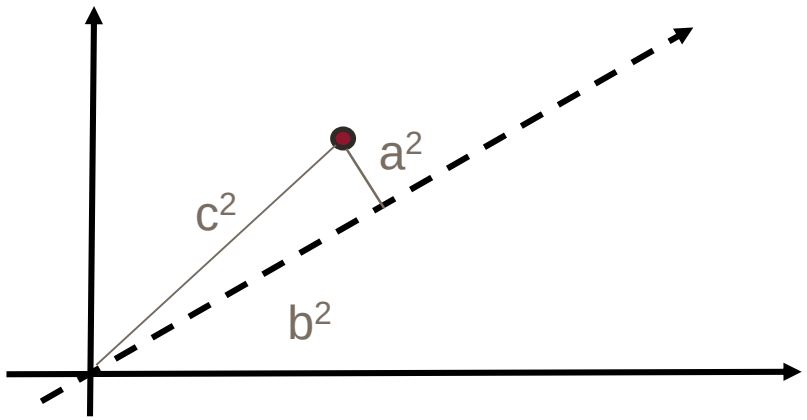
- Minimalster Abstand zur neuen Achse



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (5)

- Maximiere Distanz des projizierten Punktes zum Ursprung
- Minimiere Abstand zwischen Linie und Punkt
- Äquivalenz!
  - Pythagoras

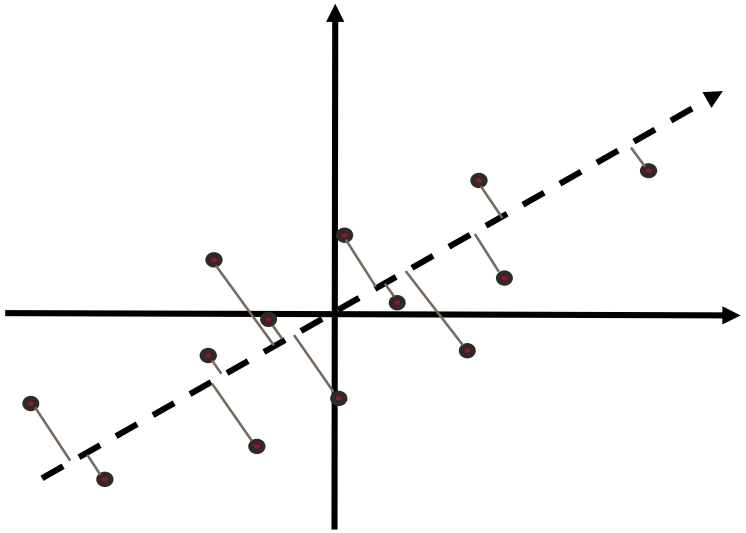


$$a^2 + b^2 = c^2$$

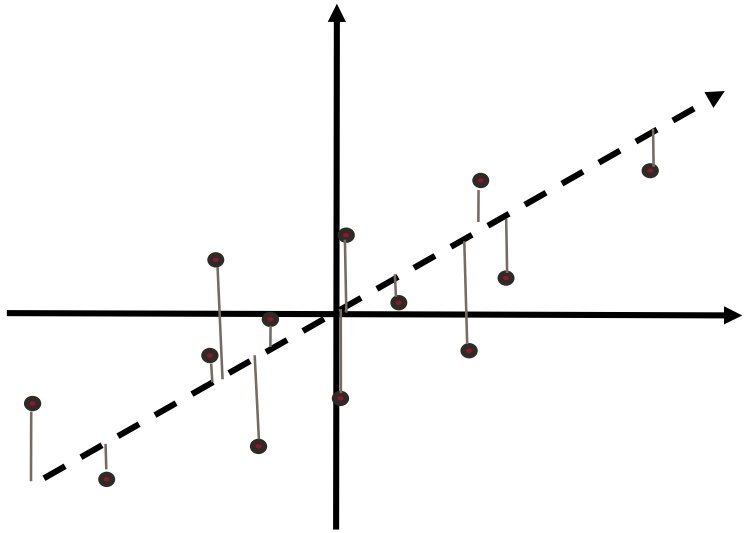
PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (6)

PCA



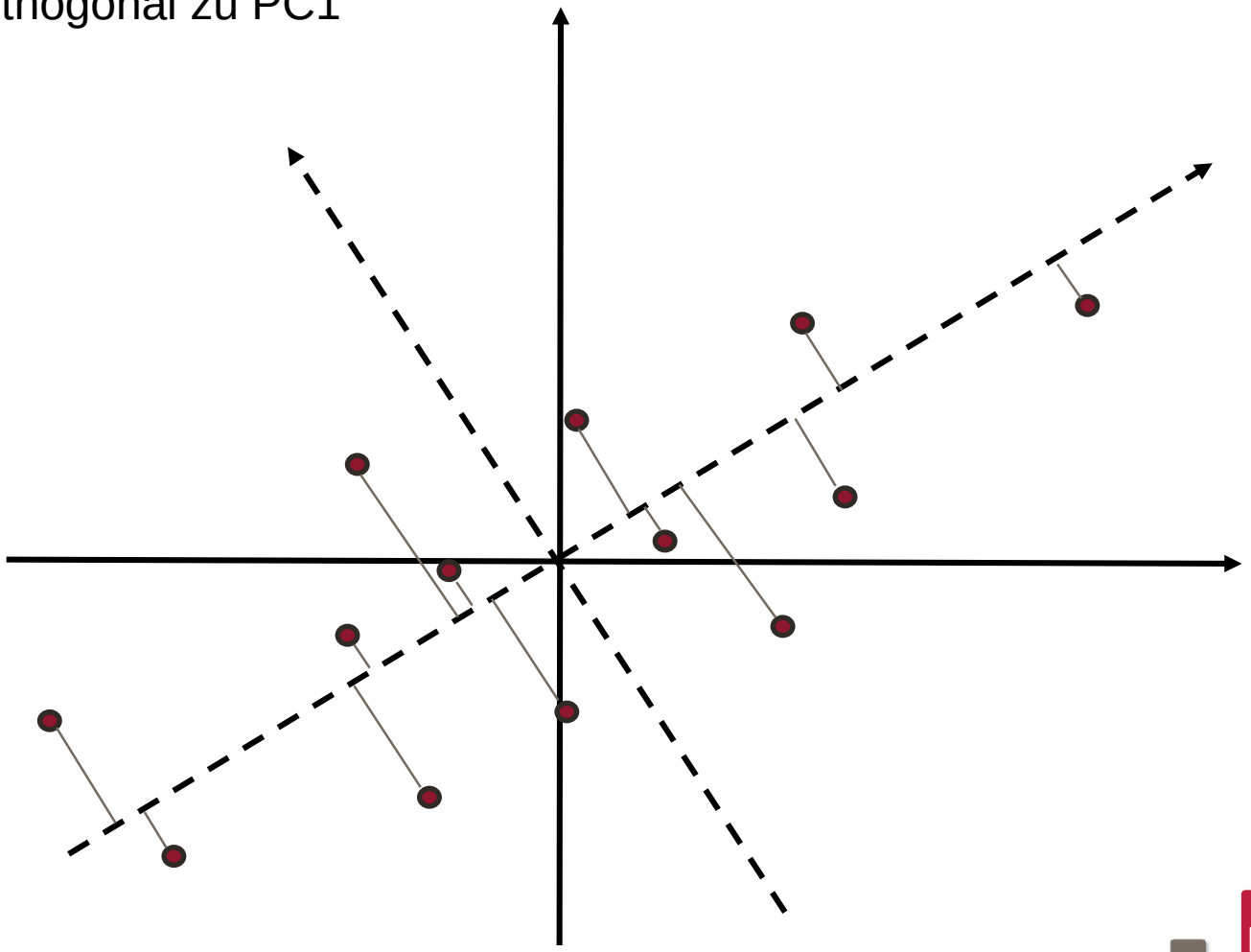
Regression



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Idee (7)

- PC2: Richtung mit der größter verbleibender Varianz
  - Orthogonal zu PC1



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

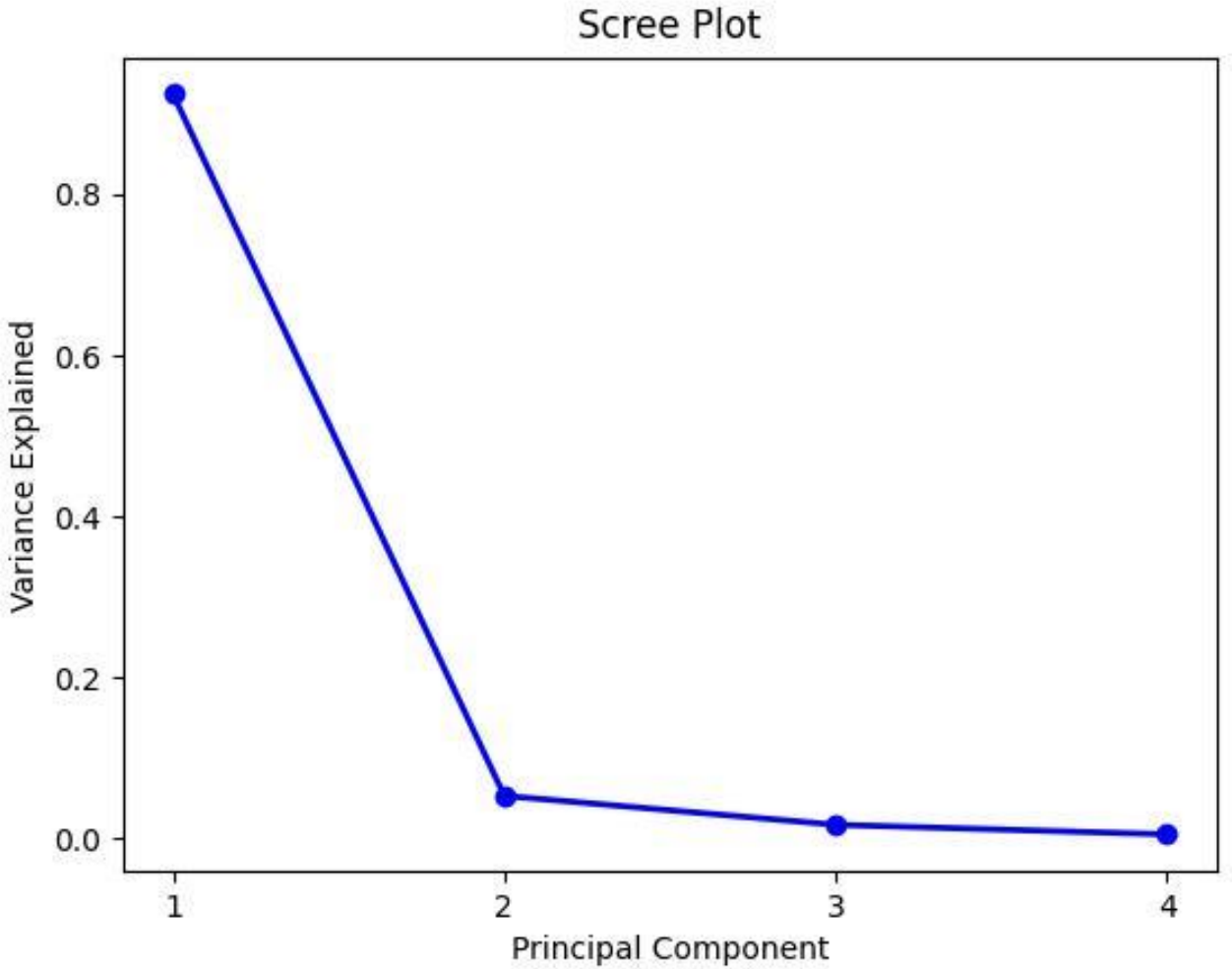
PCA – Mathematisches Vorgehen

- Daten zentrieren $X_{i,j} = X_{i,j} - \frac{\sum_{j=1}^{n} X_{i,j}}{n-1}$
- Kovarianzmatrix berechnen $Cov(X) = \frac{1}{n-1} X^\intercal X$
- Eigenwerte und Eigenvektoren der Kovarianzmatrix bestimmen $\lambda, v = eig(Cov(X))$
- Eigenwerte und Eigenvektoren nach Varianz sortieren $\lambda, v = \lambda[argsort(\lambda)], v[i, argsort(\lambda)]$

  (Die ersten r Vektoren mit größtem Eigenwert behalten) $\lambda, v = \lambda[:r], v[i, :r]$
- Daten transformieren $X_{pca} = v^\intercal X$

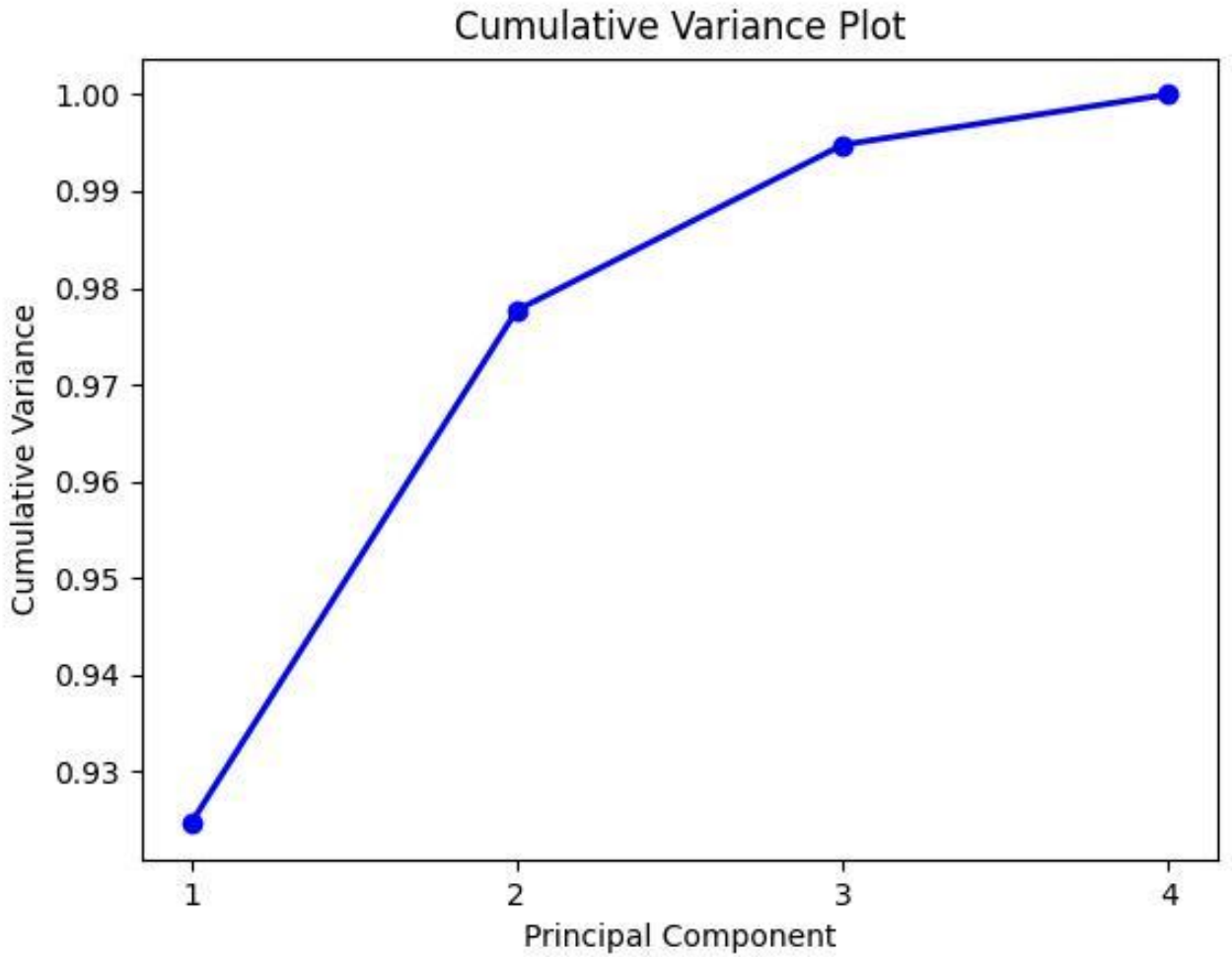
PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Scree Plot (Iris Datensatz)



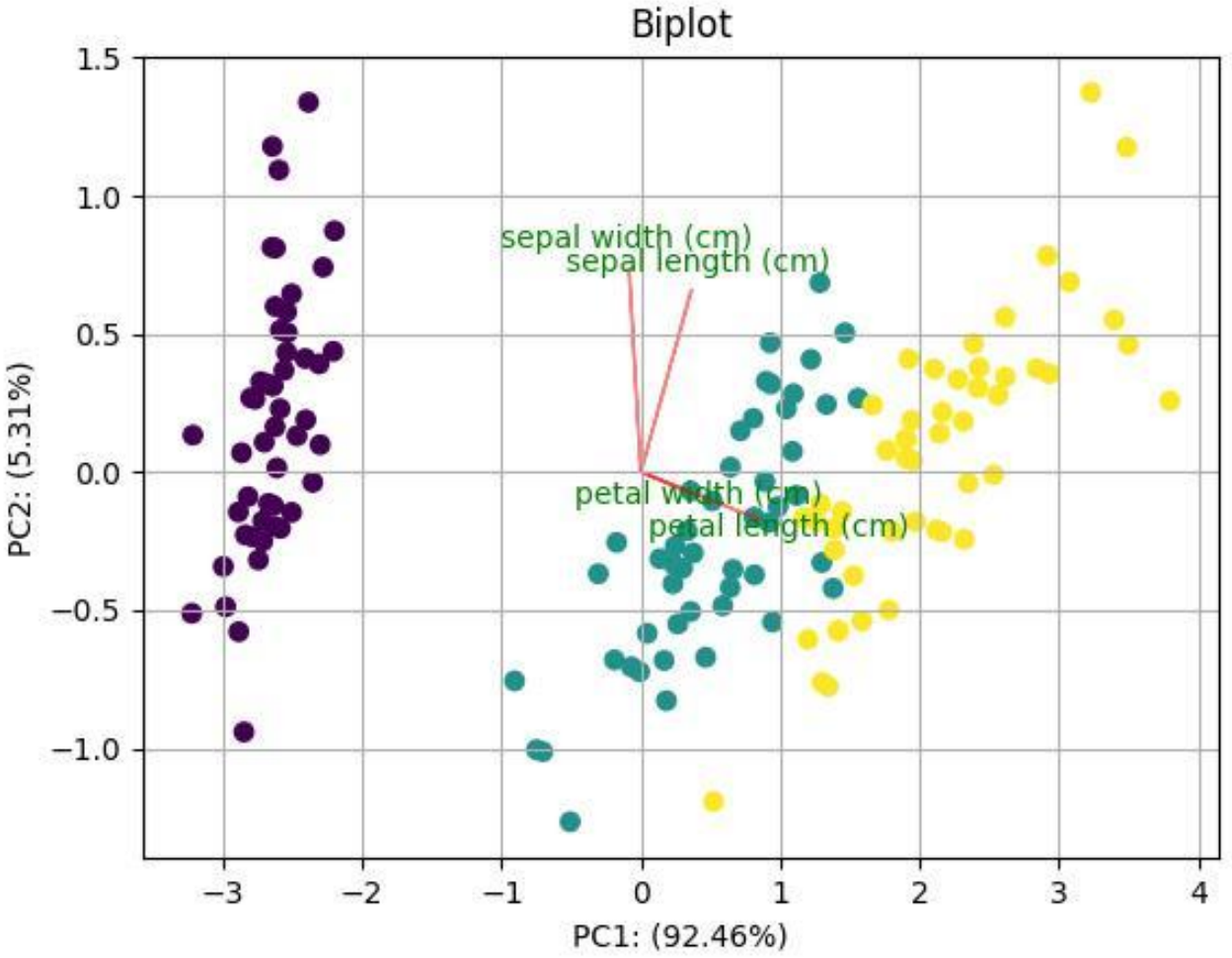
PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

PCA – Kumulative Varianz (Iris Datensatz)



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Biplot (Iris Datensatz)



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

PCA – Zusammenfassung

- Visualisierung
- Varianz ~ Informationsgehalt
- Neue Basis nach maximaler Varianz geordnet
  - Linearkombinationen der ursprünglichen Variablen
  - Linear unabhängig
- Dimensionsreduktion
  - Erste r PCs mit größter Varianz behalten
- Parametrisch
  - Abbildung/Funktion zur Projektion der Punkte wird bestimmt
- Empfindlich gegenüber Ausreißern

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE

PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

t-SNE

- T-distributed Stochastic Neighbor Embedding
  - Nicht-linear
  - Nicht-parametrisch
  - iterativ
- Grundgedanke
  - Paarweise Ähnlichkeit behalten
- Lokale Zusammenhänge
  - Keine globalen Zusammenhänge!
  - Clustergrößen verlieren Bedeutung

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

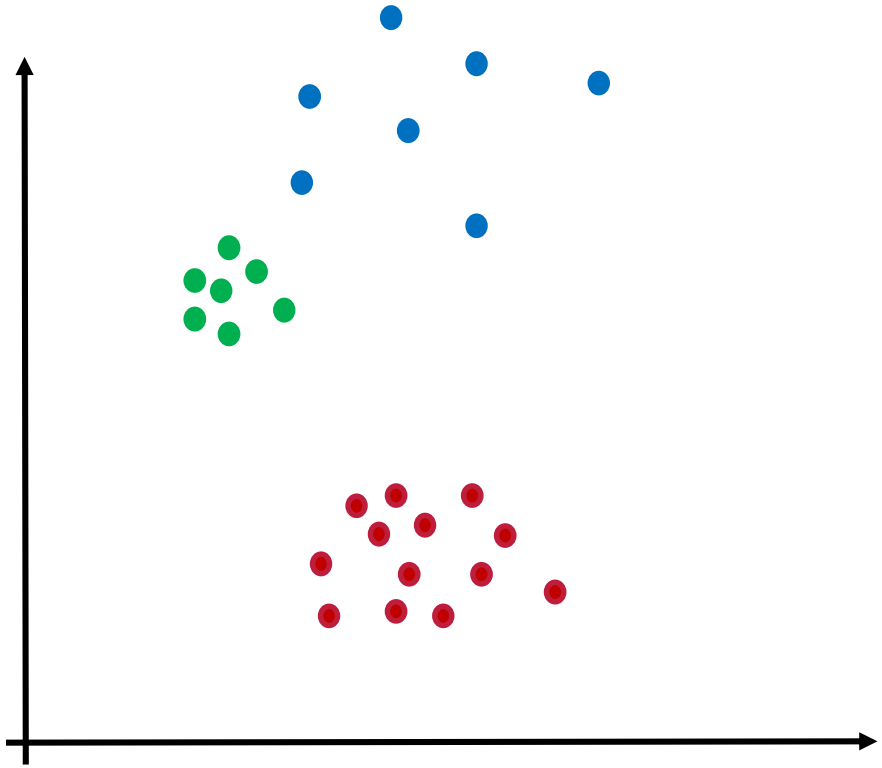
t-SNE – Grundgedanke

- Wahrscheinlichkeitverteilung über Distanzen
- Hochdimensionaler Raum
  - Gauss-Verteilung
  - Weit entfernt liegende Punkte geringere Wahrscheinlichkeit
- Niedrigdimensionaler Raum
  - t-Verteilung
  - Zufällig initialisiert
  - Paarweise Ähnlichkeit behalten
- Minimiert Divergenz zwischen beiden Verteilungen
  - Kullback-Leibler Divergenz
  - Gradient Descent

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE – Idee (1)

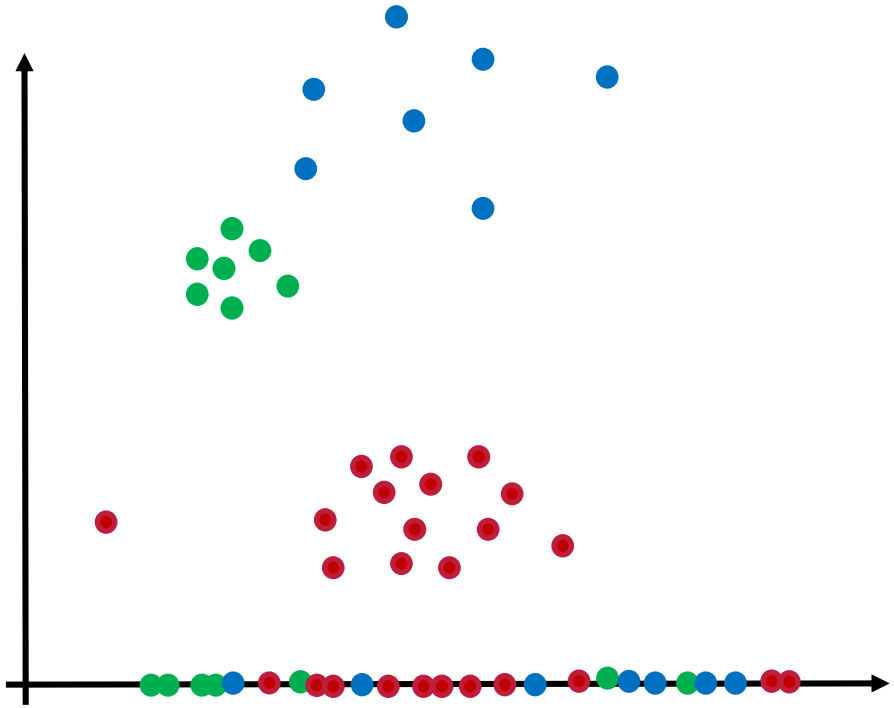
- Parameter Perplexität
  - Anzahl Nachbarn



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE – Idee (2)

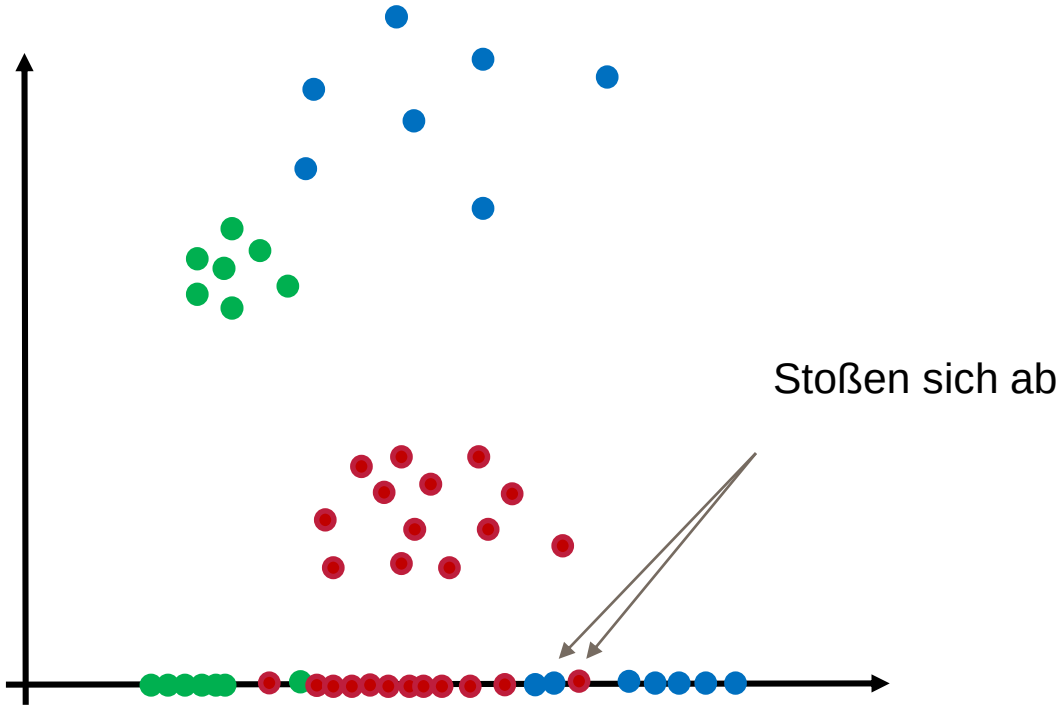
- Zufällig initialisiert



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE – Idee (3)

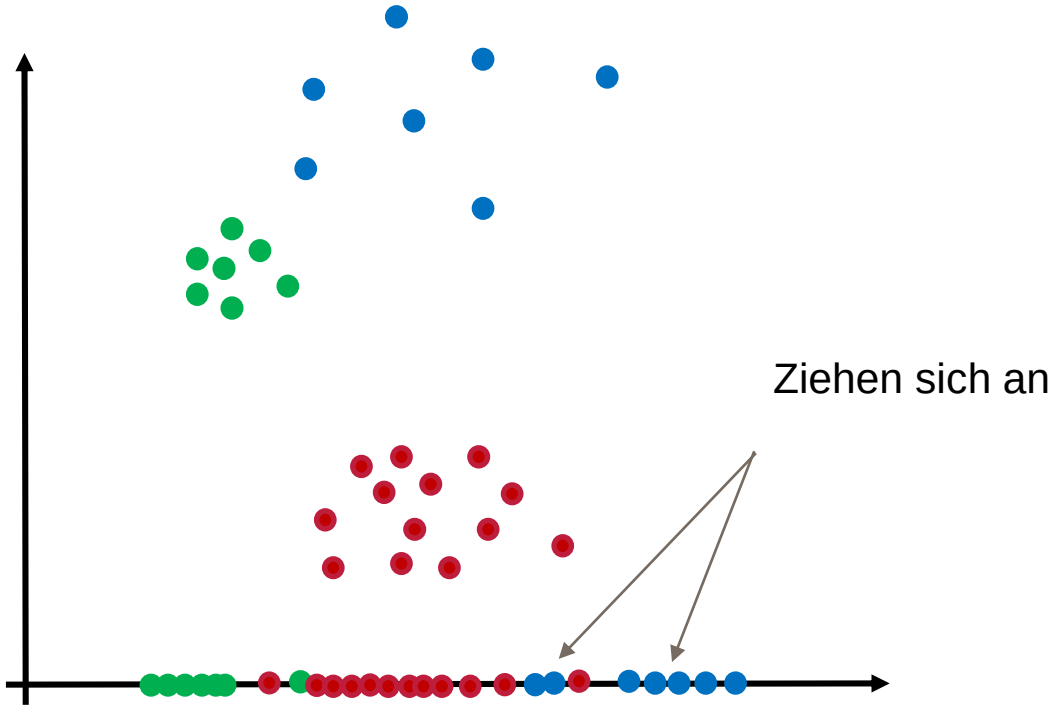
- Nachbarn ziehen sich an
- Entfernte Punkte stoßen sich ab



PLRI
PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE – Idee (4)

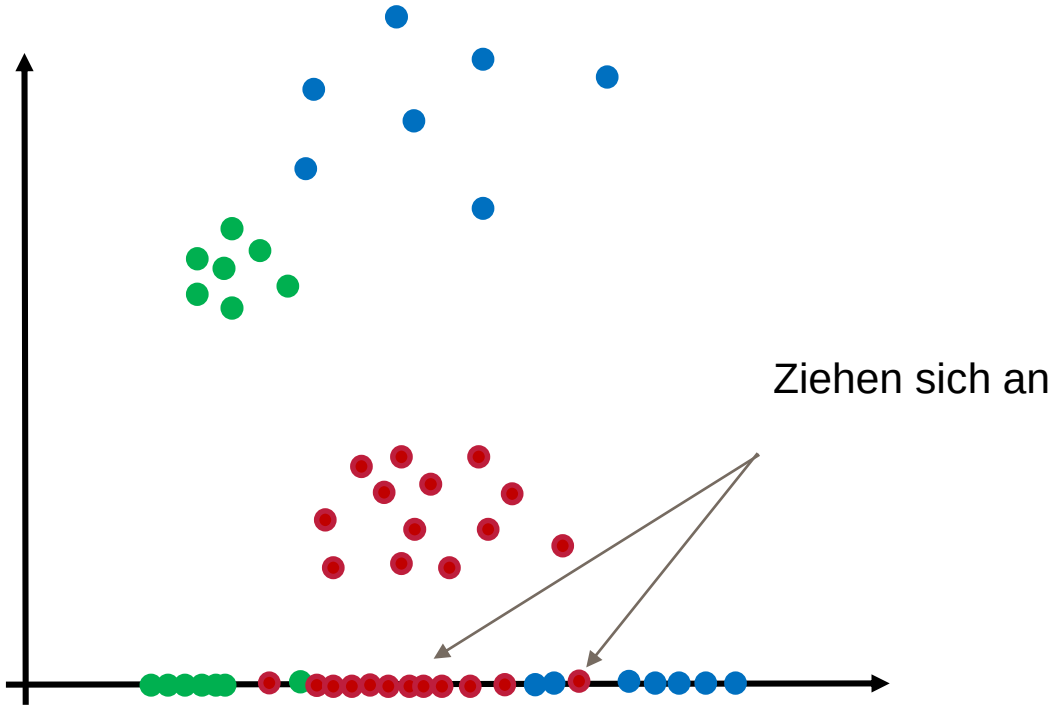
- Nachbarn ziehen sich an
- Entfernte Punkte stoßen sich ab



PLRI
PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE – Idee (5)

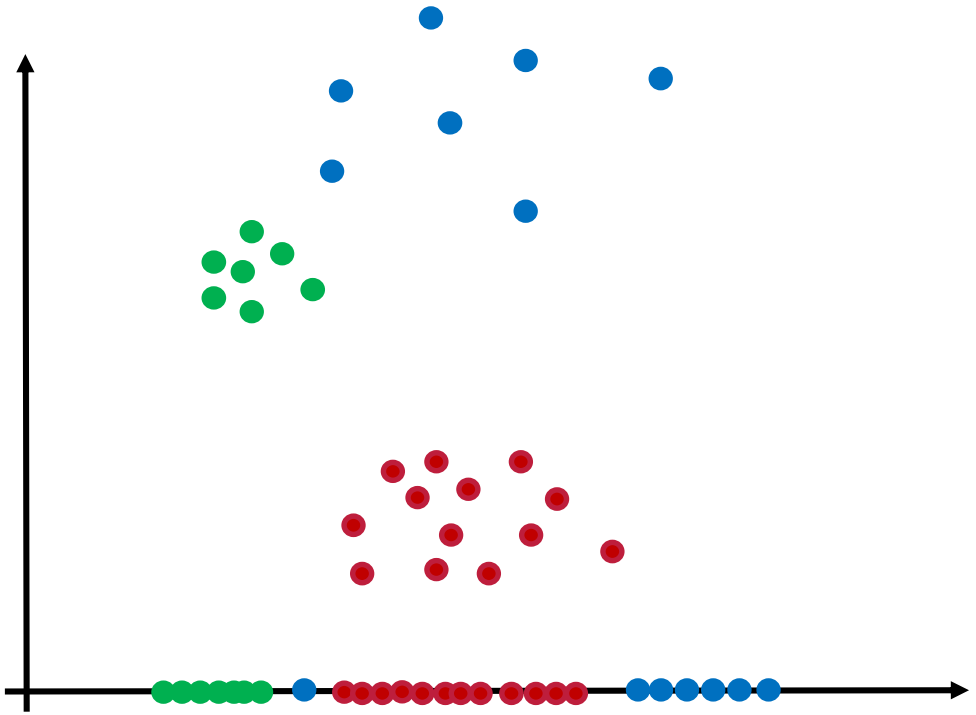
- Nachbarn ziehen sich an
- Entfernte Punkte stoßen sich ab



PLRI
PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

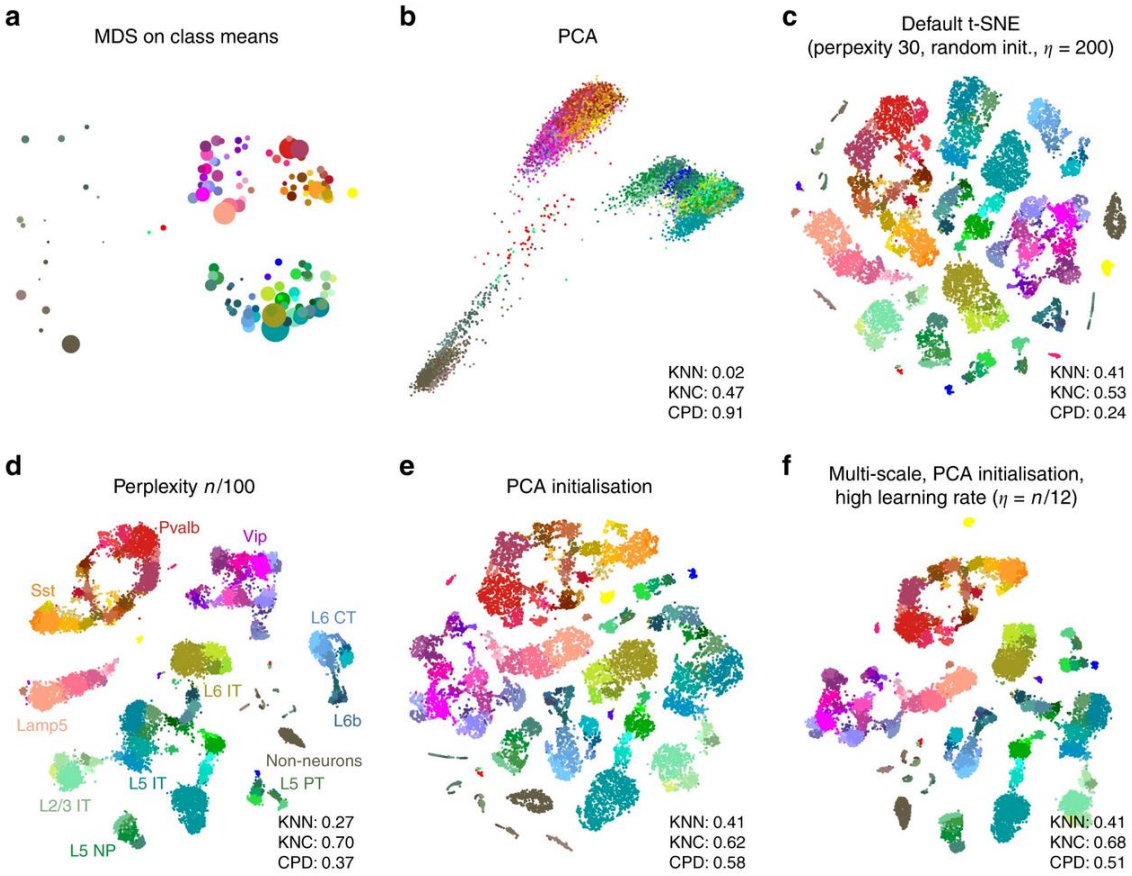
t-SNE – Early Exaggeration

- Early exaggeration
  - Am Anfang Anziehung deutlich stärker gewichten



PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

t-SNE - Beispiel



Tasic et al. data set. Sample size $n = 23,822$. Cluster assignments and cluster colours are taken from the original publication[3]. Warm colours correspond to inhibitory neurons, cold colours correspond to excitatory neurons, brown/grey colours correspond to non-neural cells. **a** MDS on class means ($n = 133$). Point sizes are proportional to the number of points per class. **b** The first two principal components of the data. KNN: 10-nearest neighbour preservation, KNC: 10-nearest classes preservation, CPD: Spearman correlation between pairwise distances. **c** Default t-SNE with perplexity 30, random initialisation, and learning rate 200. **d** T-SNE with perplexity $n/100 = 238$. Labels denote large groups of clusters. **e** T-SNE with PCA initialisation. **f** T-SNE with multi-scale similarities (perplexity combination of 30 and $n/100 = 238$, PCA initialisation, and learning rate $n/12 \approx 2000$.

Kobak, Dmitry, and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics." *Nature communications* 10.1 (2019): 5416.

t-SNE – Zusammenfassung

- Nicht-lineare Mannigfaltigkeiten
- Lernt keine Funktion
- Optimierung zwischen Anziehung und Abstoßung
- Lokale Zusammenhänge
- Clustergrößen, Dichte, Abstände gehen verloren

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Embeddings

PLRI
PETER L.
REICHERTZ INSTITUT
FÜR MEDIZINISCHE
INFORMATIK

UMAP

- Uniform Manifold Approximation & Projection
- Ähnlicher Gedanke wie t-SNE

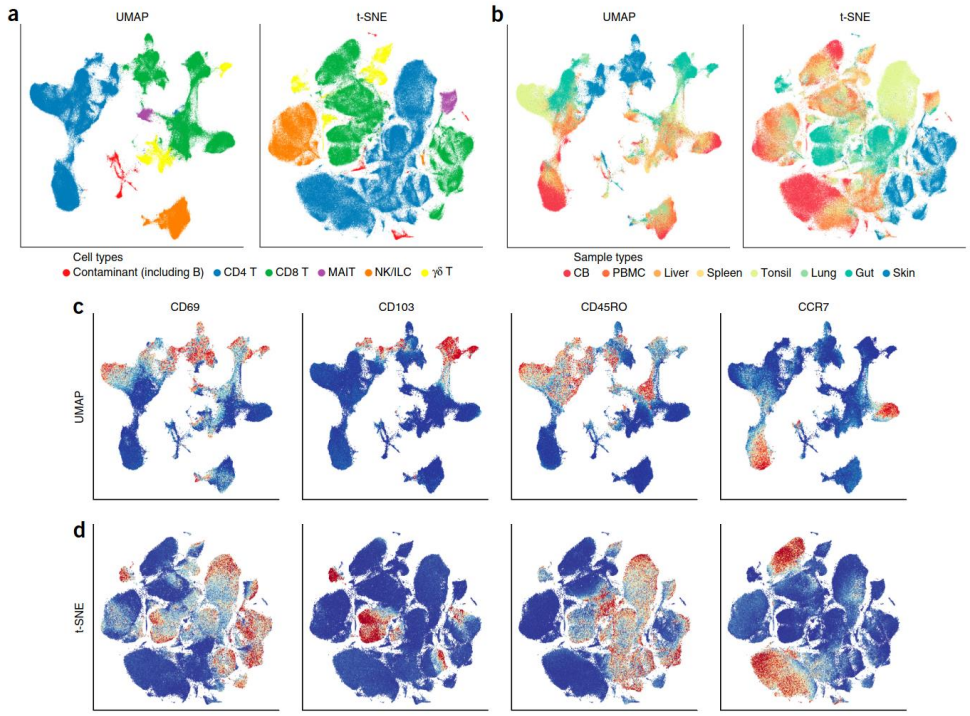


Figure 1 UMAP embeds local and large-scale structure of the data. UMAP and t-SNE projections of the Wong *et al.* dataset colored according to (**a**) broad cell lineages, (**b**) tissue of origin, and for (**c**) UMAP and (**d**) t-SNE, the expression of CD69, CD103, CD45RO and CCR7. For **c** and **d**, blue denotes minimal expression, beige intermediate and red high. MAIT, mucosal-associated invariant T cell; ILC, innate lymphoid cell; CB, cord blood; PBMC, peripheral blood mononuclear cell.

McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).

word2vec

- Word embedding
- Wörter durch ihren Kontext beschreiben
  - Ähnliche Wörter kommen oft im gleichen Kontext vor

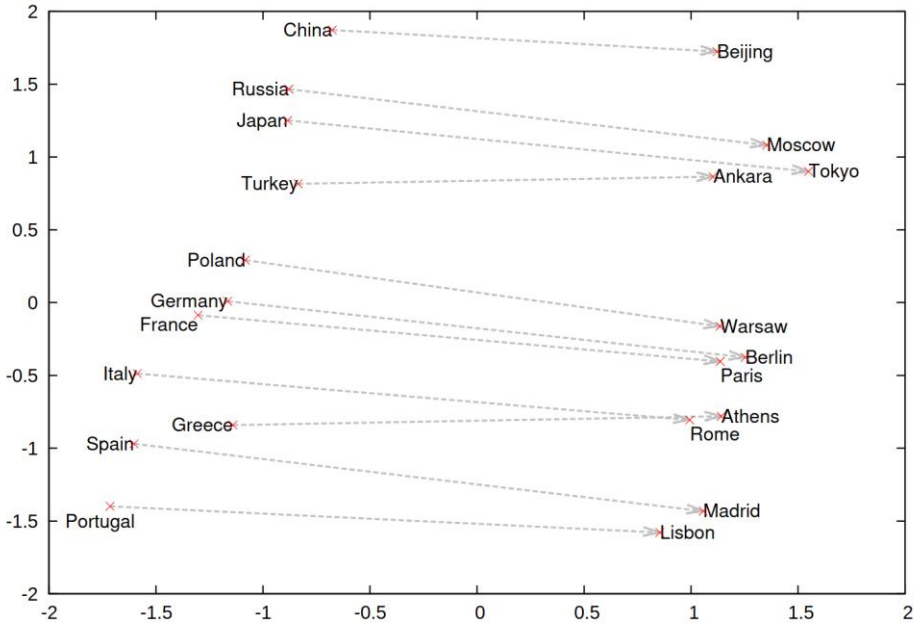


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Image embedding

- Bilder und Videos sehr hochdimensional
- Embedding reduziert Dimensionen
- Versucht, wichtige (übergeordnete) Strukturen und Informationen zu erhalten
  - Farbe
  - Textur, Formen
- Beispiel: ResNet-18
  - Convolutional Neural Network

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK

Embeddings – Zusammenfassung

- Niedrigdimensionale Vektorrepräsentationen
- Embeddings versuchen Ähnlichkeiten zu bewahren
- Global vs lokal
- Unterschiedliche Ansätze
- Parametrisch oder nicht-parametrisch
- Domänenspezifisch

Literatur

- Verleysen, Michel, and Damien François. "The curse of dimensionality in data mining and time series prediction." *International work-conference on artificial neural networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.
- Berisha, Visar, et al. "Digital medicine and the curse of dimensionality." *NPJ digital medicine* 4.1 (2021): 153.
- Fischer, Gerd. *Lernbuch Lineare Algebra und Analytische Geometrie*. Vieweg+ Teubner, 2011.
- Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010): 433-459.
- Kobak, Dmitry, and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics." Nature communications 10.1 (2019): 5416.
- McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).
- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).

Videos & Weiteres

- Lecture 4 "Curse of Dimensionality / Perceptron" -Cornell CS4780 SP17, Kilian Weinberger: https://www.youtube.com/watch?v=BbYV8UfMJSA&list=PLl8OlHZGYOQ7bkVbuRthEsaLr7bONzbXS&index=4
- Eigenvectors and eigenvalues | Chapter 14, Essence of linear algebra, 3Blue1Brown: https://www.youtube.com/watch?v=PFDu9oVAE-g&pp=ygUMZWlnZW52ZWN0b3Jz
- Singular Value Decomposition [Data-Driven Science and Engineering], Steve Brunton: https://www.youtube.com/watch?v=gXbThCXjZFM&list=PLMrJAkhIeNNSVjnsviglFoY2nXildDCcv
- Introduction to Machine Learning - 11 - Manifold learning and t-SNE, Dmitry Kobak: https://www.youtube.com/watch?v=MnRskV3NY1k
- https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

PLRI PETER L. REICHERTZ INSTITUT FÜR MEDIZINISCHE INFORMATIK