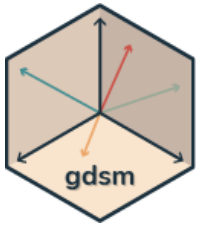


Session 03:
Statistik semantischer Vektoren

Viktoria Schneider & Dominic Schmitz

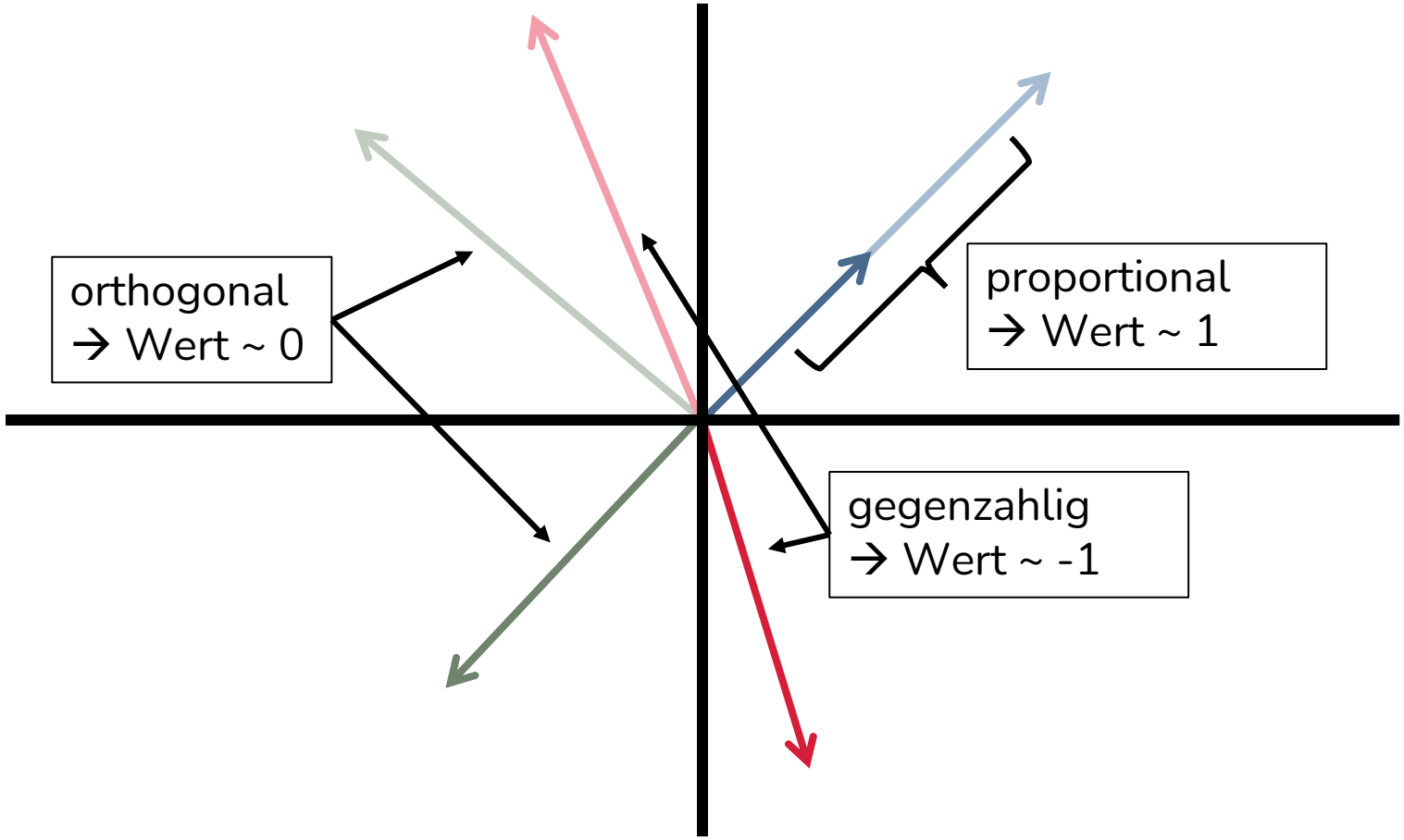
Verein für Diversität in der Linguistik

Statistische Maße

- Einige statistische Maße werden regelmäßig zur statistischen Analyse semantischer Vektoren genutzt
  - Kosinus-Ähnlichkeit *cosine similarity*
  - Euklidischer Abstand *Euclidean distance*
  - Manhattan-Distanz *Manhattan distance*
  - Shannon Entropie *Shannon entropy*
  - Nächste Nachbarn *nearest neighbours*
  - Nachbarschaftsdichte *neighbourhood density*

Kosinus-Ähnlichkeit

- Maß für die Ähnlichkeit zweier Vektoren



Kosinus-Ähnlichkeit

- Maß für die Ähnlichkeit zweier Vektoren
- in R:

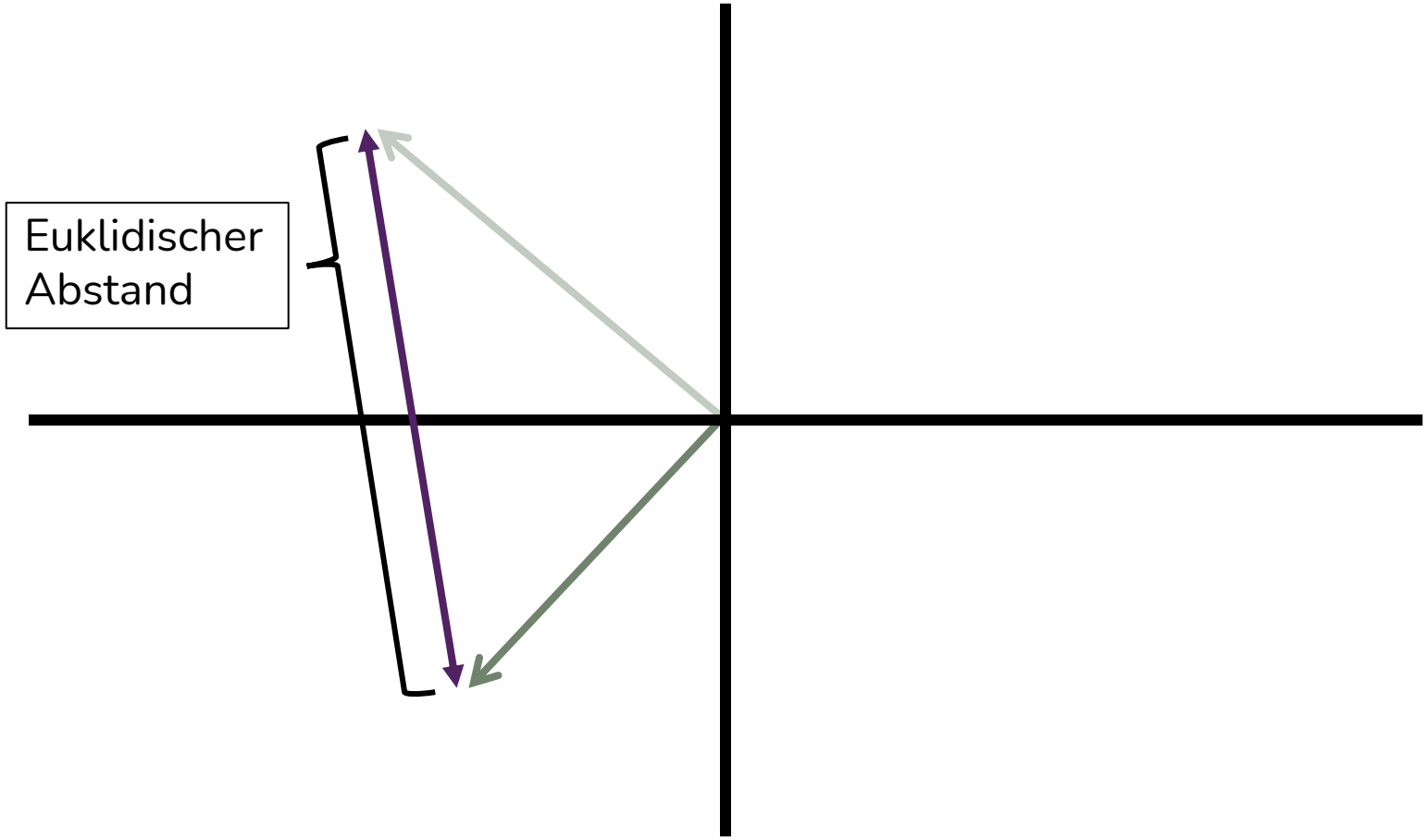
```
library("gdsm")

# Vektoren aus einem Vektorraum
cosim("var01", "var02", gdsm_mat)

# Vektoren, die individuell hinterlegt sind
cosim(var01, var02)
```

Euklidischer Abstand

- Maß für den Abstand zweier Vektoren



Euklidischer Abstand

- Maß für den Abstand zweier Vektoren
- in R:

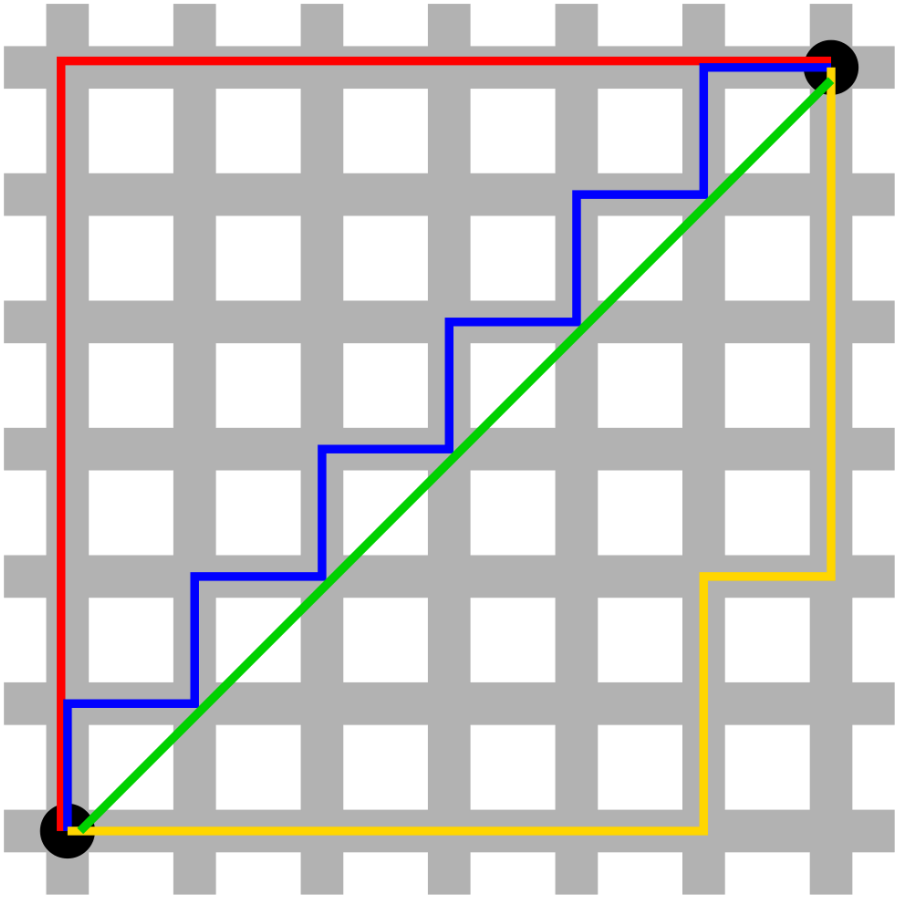
```
library("gdsm")

# Vektoren aus einem Vektorraum
euclid("var01", "var02", gdsm_mat)

# Vektoren, die individuell hinterlegt sind
euclid(var01, var02)
```

Manhattan-Distanz

- Maß für den Abstand zweier Vektoren



Manhattan-Distanz

- Maß für den Abstand zweier Vektoren
- in R:

```
library("gdsm")

# Vektoren aus einem Vektorraum
manhat("var01", "var02", gdsm_mat)

# Vektoren, die individuell hinterlegt sind
manhat(var01, var02)
```

Shannon Entropie

- Maß für die Eindeutigkeit eines Vektors (Stichwort: Polysemie)
- Achtung: Aktuell noch unbekannt, ob tatsächlich nützlich/verlässlich

```
library("gdsm")

# Vektoren aus einem Vektorraum
shannon("var01", gdsm_mat)

# Vektoren, die individuell hinterlegt sind
shannon(var01)
```

Shannon Entropie

- Maß für die Eindeutigkeit eines Vektors (Stichwort: Polysemie)
- in R:

```
library("gdsm")

# Vektor aus einem Vektorraum
shannon("var01", gdsm_mat)

# Vektor, der individuell hinterlegt ist
shannon(var01)
```

Nächste Nachbarn

- Anhand der Korrelation von einem Vektor X und allen anderen Vektoren werden die ähnlichsten Vektoren bestimmt
- in R:

```
library("gdsm")

# Vektor aus einem Vektorraum
find_nn("var01", gdsm_mat)

# Vektor, der individuell hinterlegt ist
find_nn(var01)
```

Nachbarschaftsdichte

- Die durchschnittliche Korrelation mit den *n* nächsten Nachbarn sagt aus wie dicht die direkte Nachbarschaft des Wortes ist
- in R:

```
library("gdsm")

# Vektor aus einem Vektorraum
n_density("var01", gdsm_mat)

# Vektor, der individuell hinterlegt ist
n_density(var01)
```