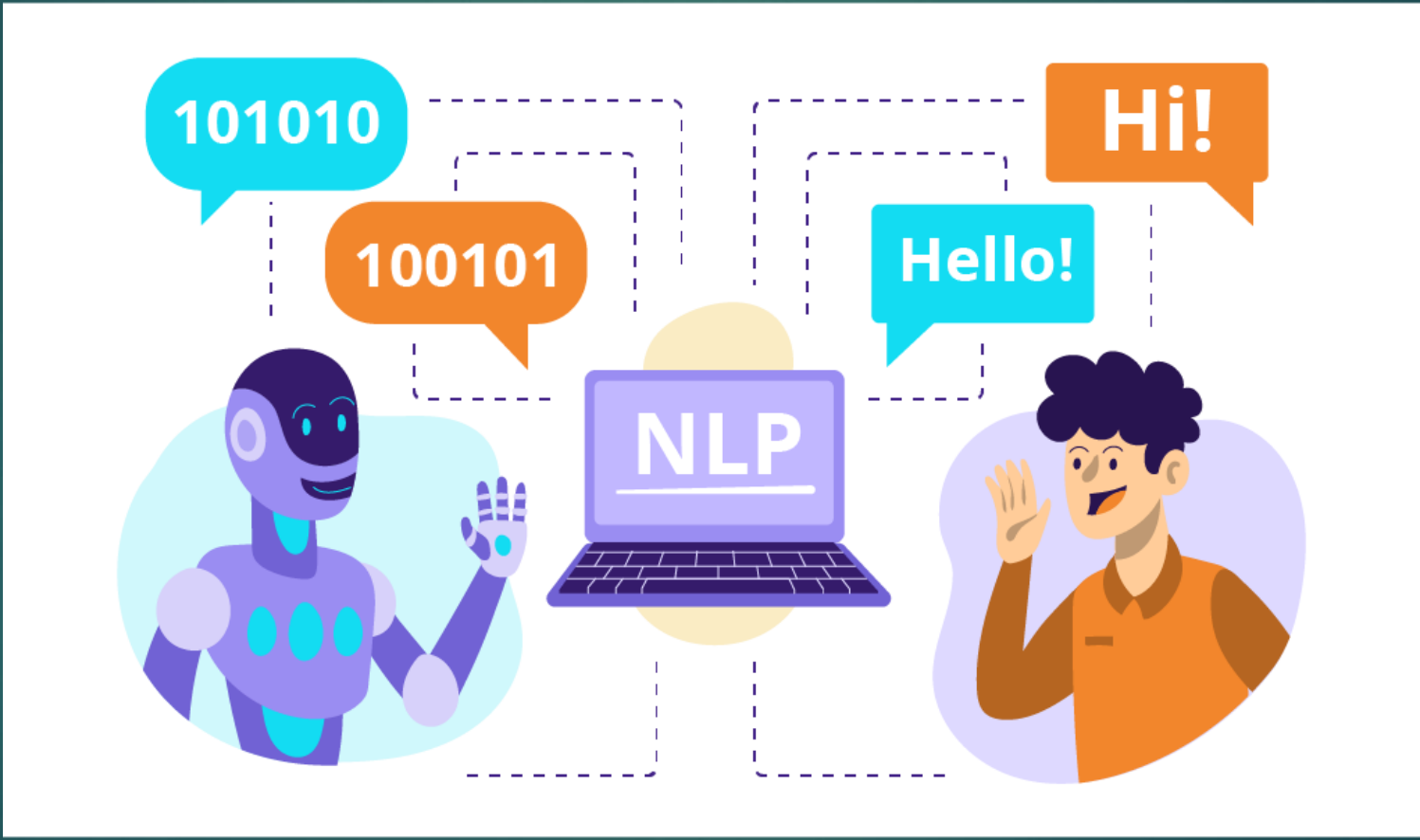
11- Deep learning for text

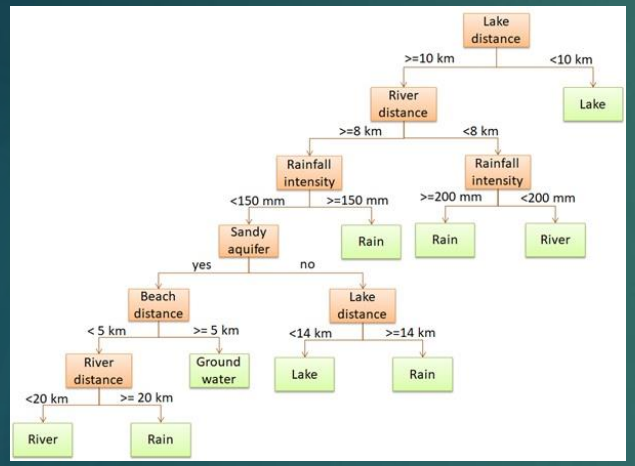
AMIN HEYDARI & FARIBA TAVAKOLI

Human langs vs Machine langs

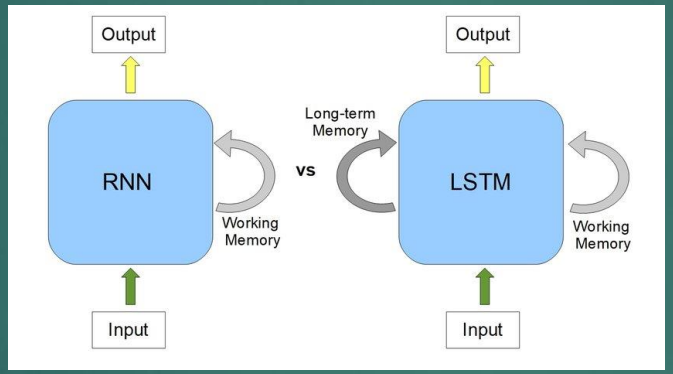


NLP

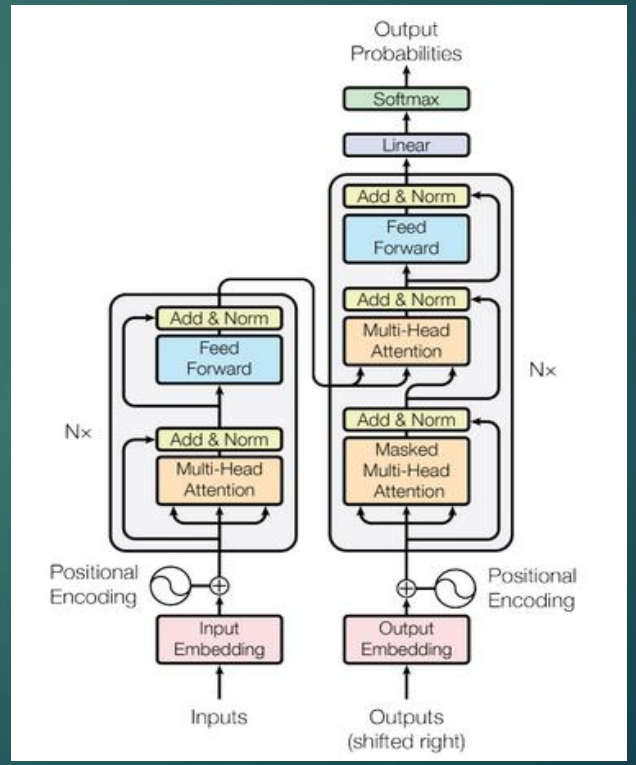
Decision tree
1990-2010



RNN & LSTM
2014-2017

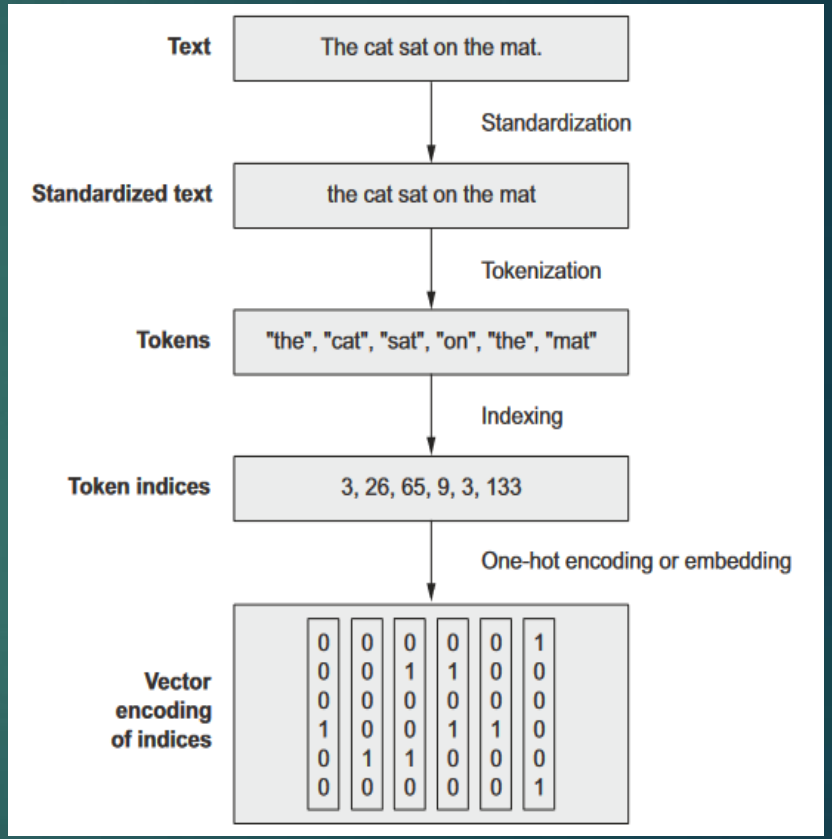


Transformers
2017 - Now

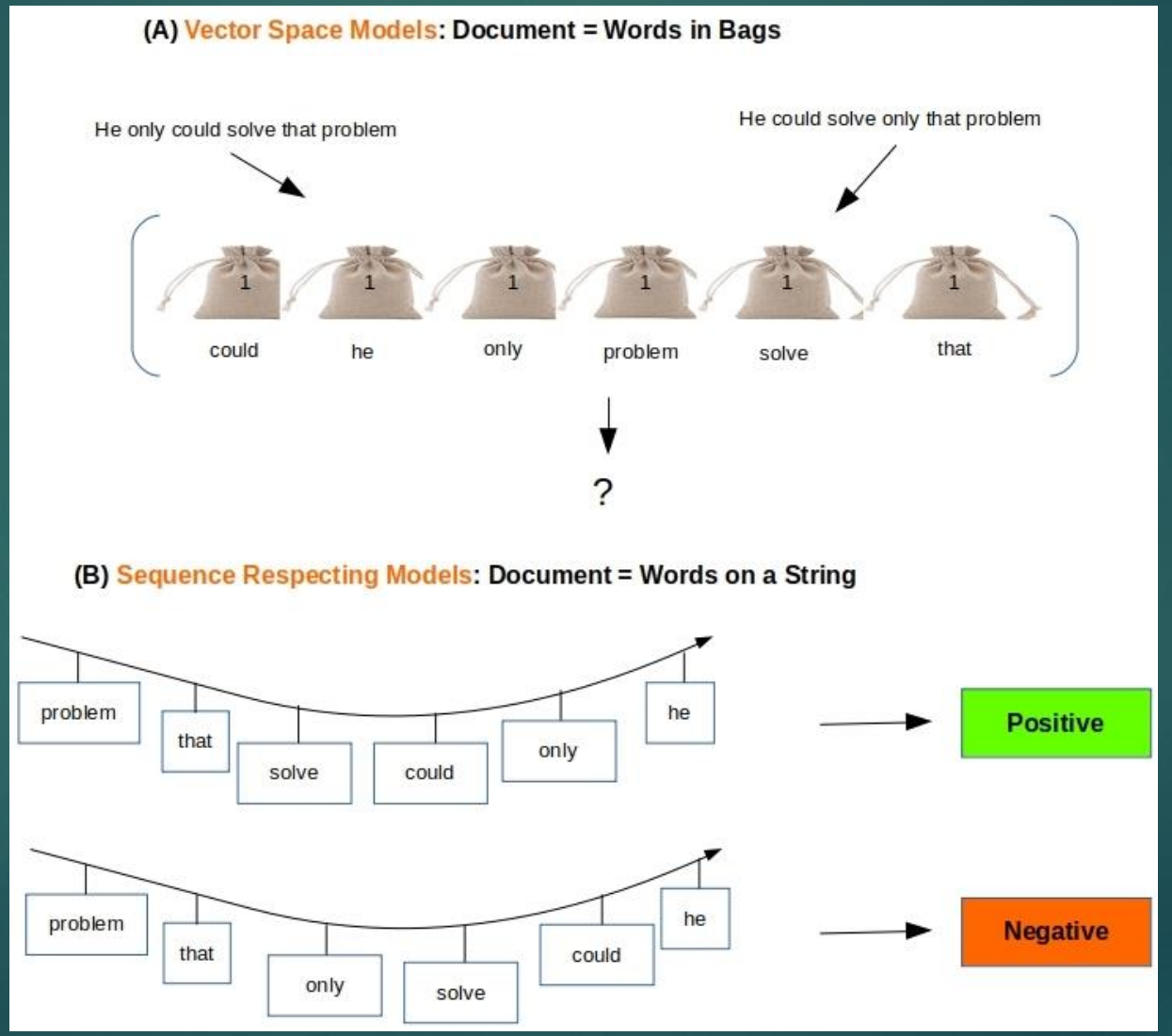


Preparing text data

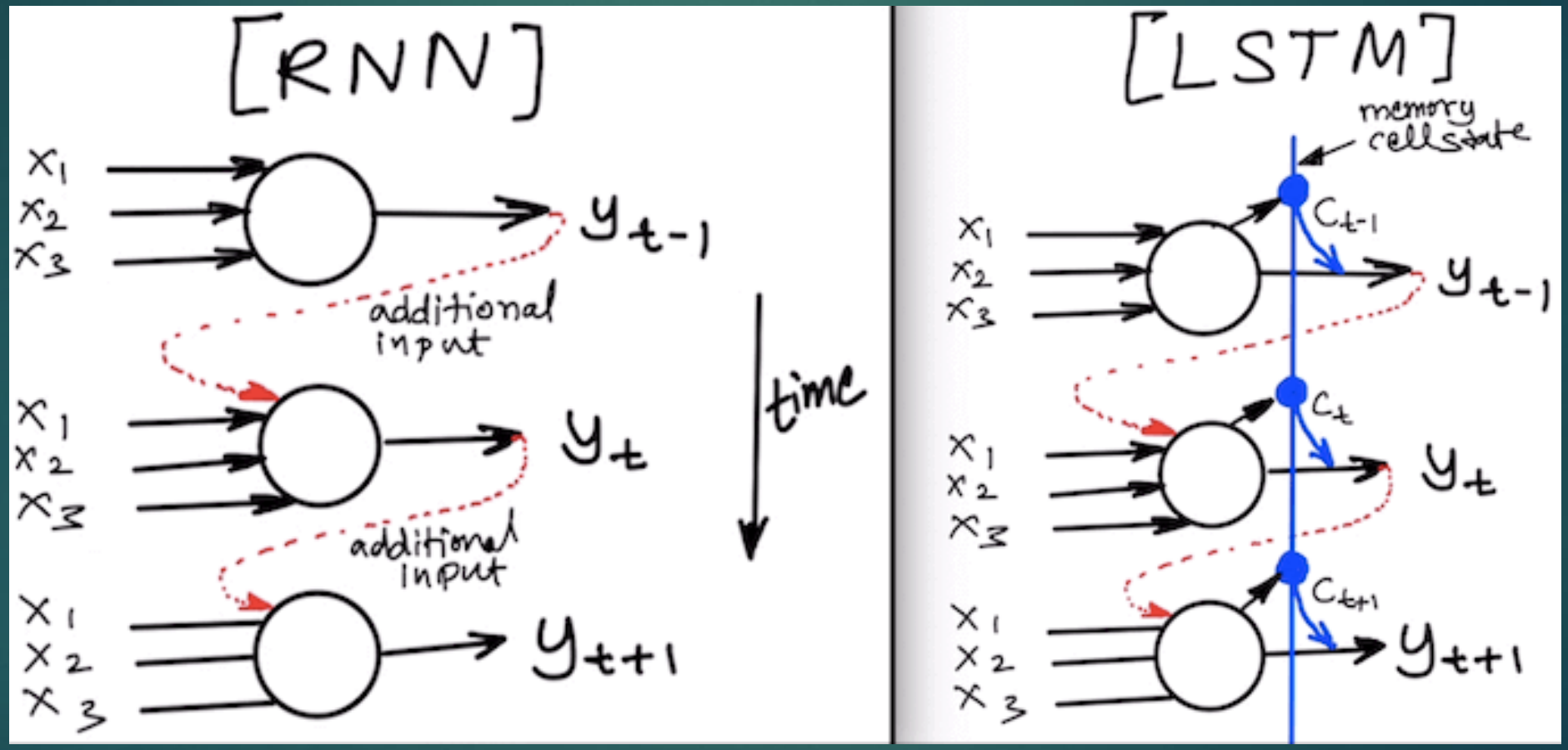
i. Standardize
ii. Tokenization
iii. Convert tokens into a numerical vector



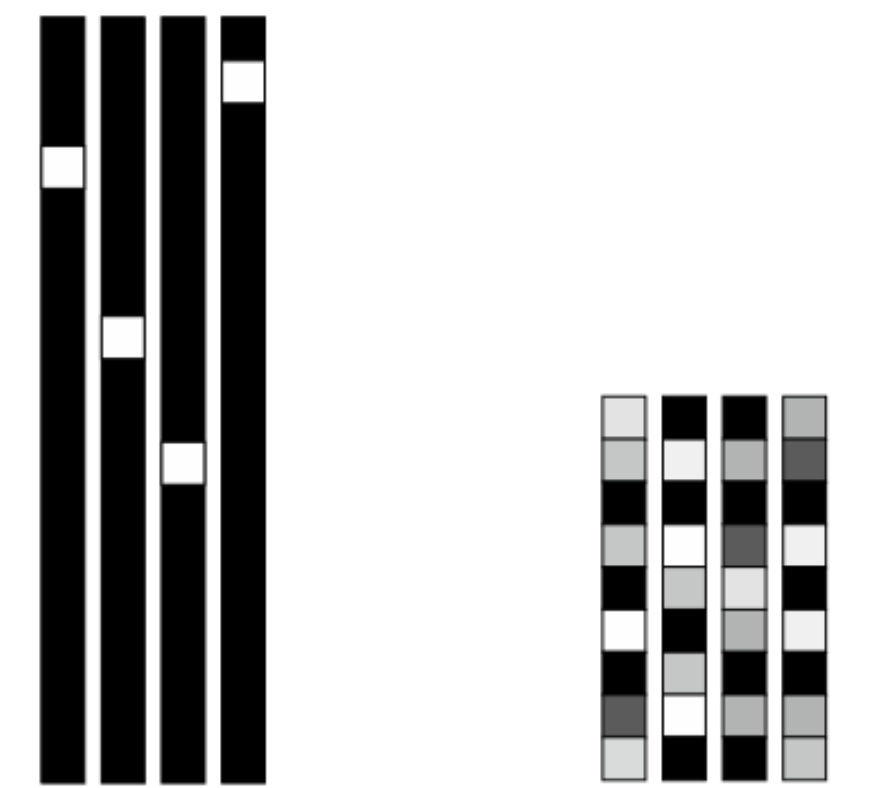
Sets vs Sequences



RNN vs LSTM


[RNN]
x_1
x_2
x_3
y_{t-1}
additional input
y_t
additional input
y_{t+1}
time
[LSTM]
memory cellstate
C_{t-1}
C_t
C_{t+1}

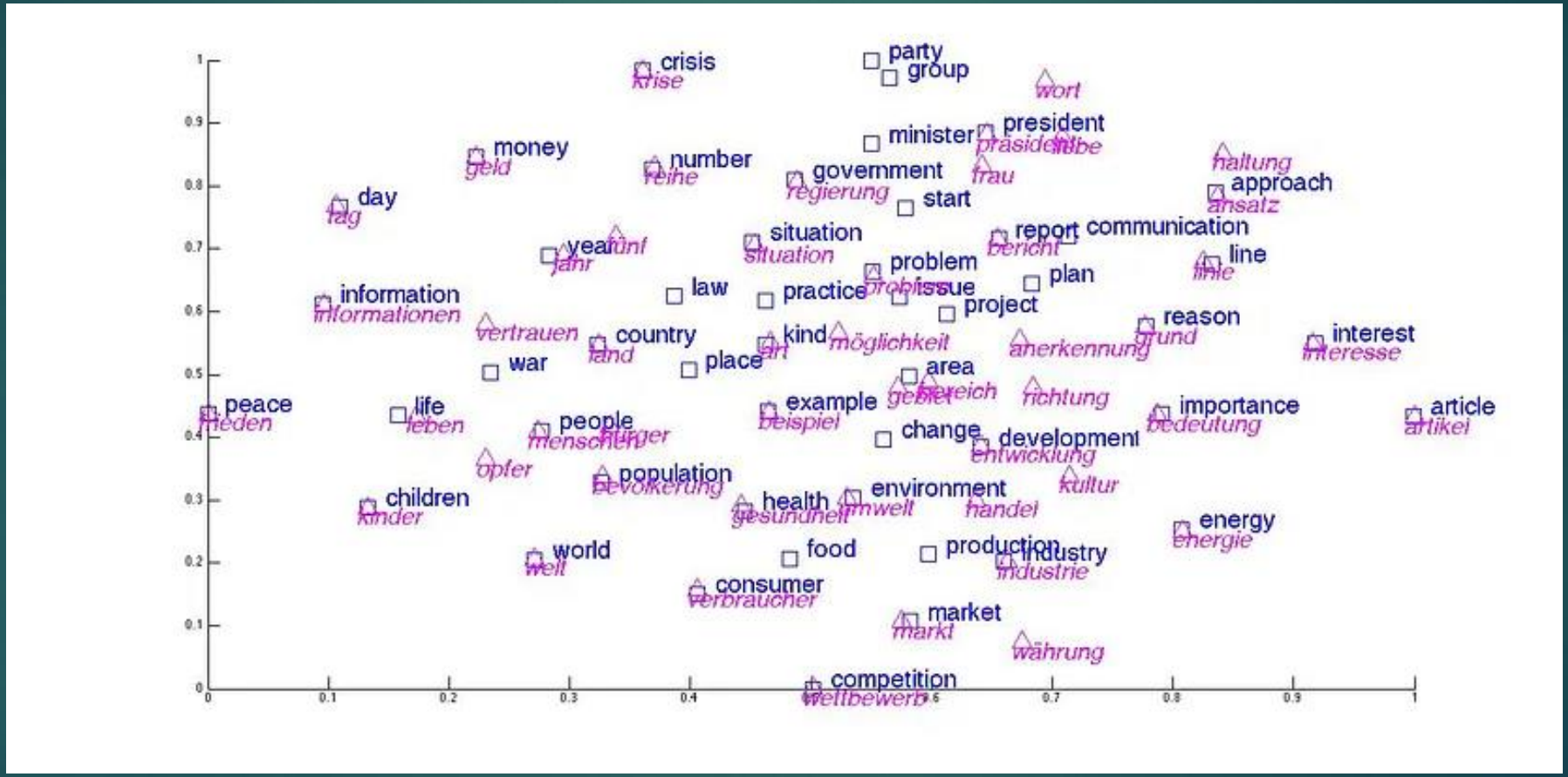
One-hot vs Embedding



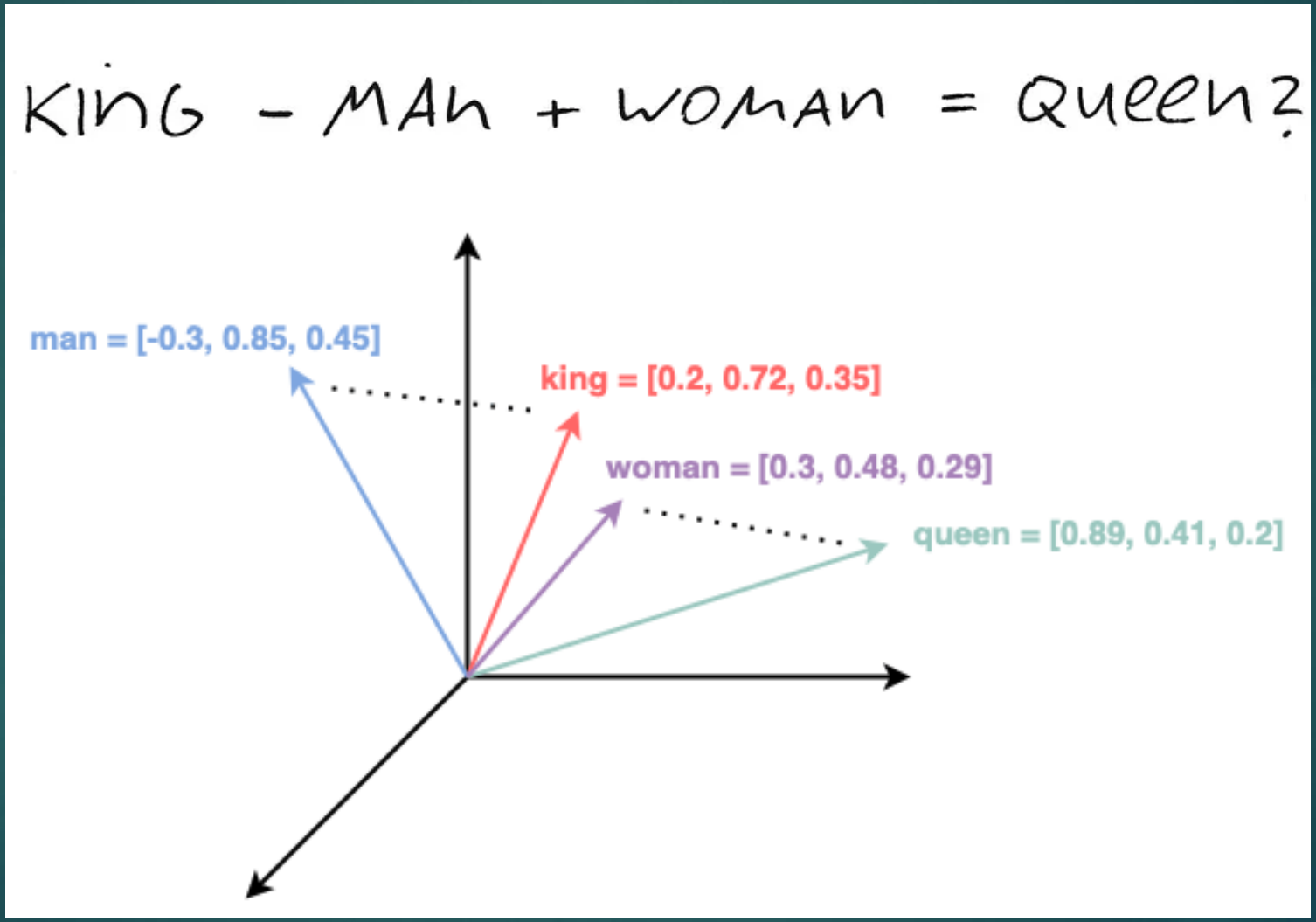
One-hot word vectors:
- Sparse
- High-dimensional
- Hardcoded

Word embeddings:
- Dense
- Lower-dimensional
- Learned from data

Embedding


crisis
krise
party
group
wort
president
minister
präsident
money
geld
number
reihe
government
regierung
frau
start
haltung
approach
ansatz
day
tag
situation
report
bericht
communication
line
linie
problem
plan
information
informationen
law
practice
issue
project
reason
grund
vertrauen
country
land
kind
möglichkeit
anerkennung
interest
interesse
war
place
area
peace
frieden
life
leben
example
beispiel
gebiet
richtung
importance
bedeutung
article
artikel
people
menschen
change
development
entwicklung
opfer
population
bevölkerung
kultur
children
kinder
health
gesundheit
environment
umwelt
handel
energy
energie
world
welt
food
production
industry
industrie
consumer
verbraucher
market
markt
währung
competition
wettbewerb

Embedding


KING - MAN + WOMAN = QUEEN?
man = [-0.3, 0.85, 0.45]
king = [0.2, 0.72, 0.35]
woman = [0.3, 0.48, 0.29]
queen = [0.89, 0.41, 0.2]

The End