

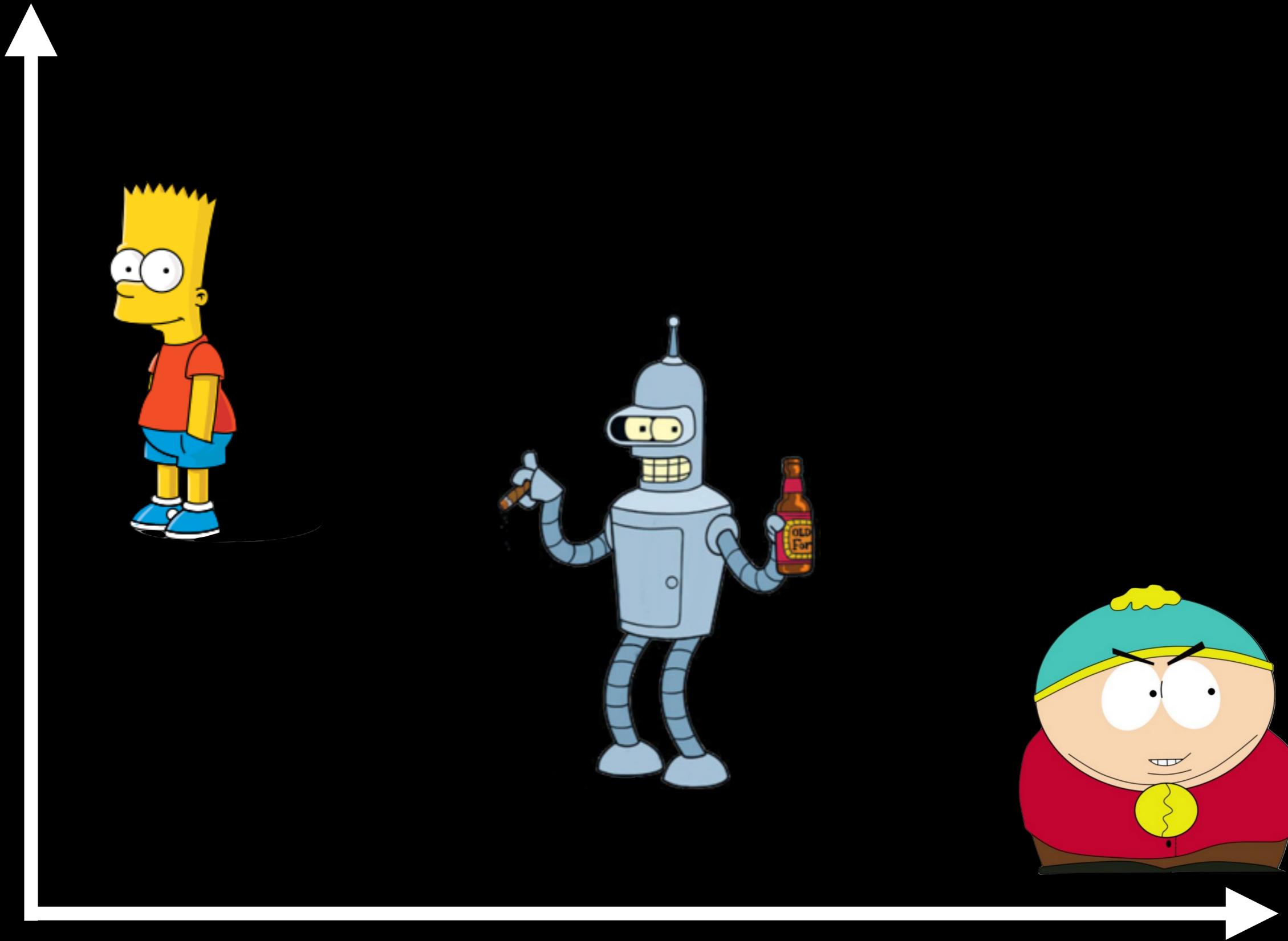
The hero with (several) thousand faces







Agreeableness



Sociopathy

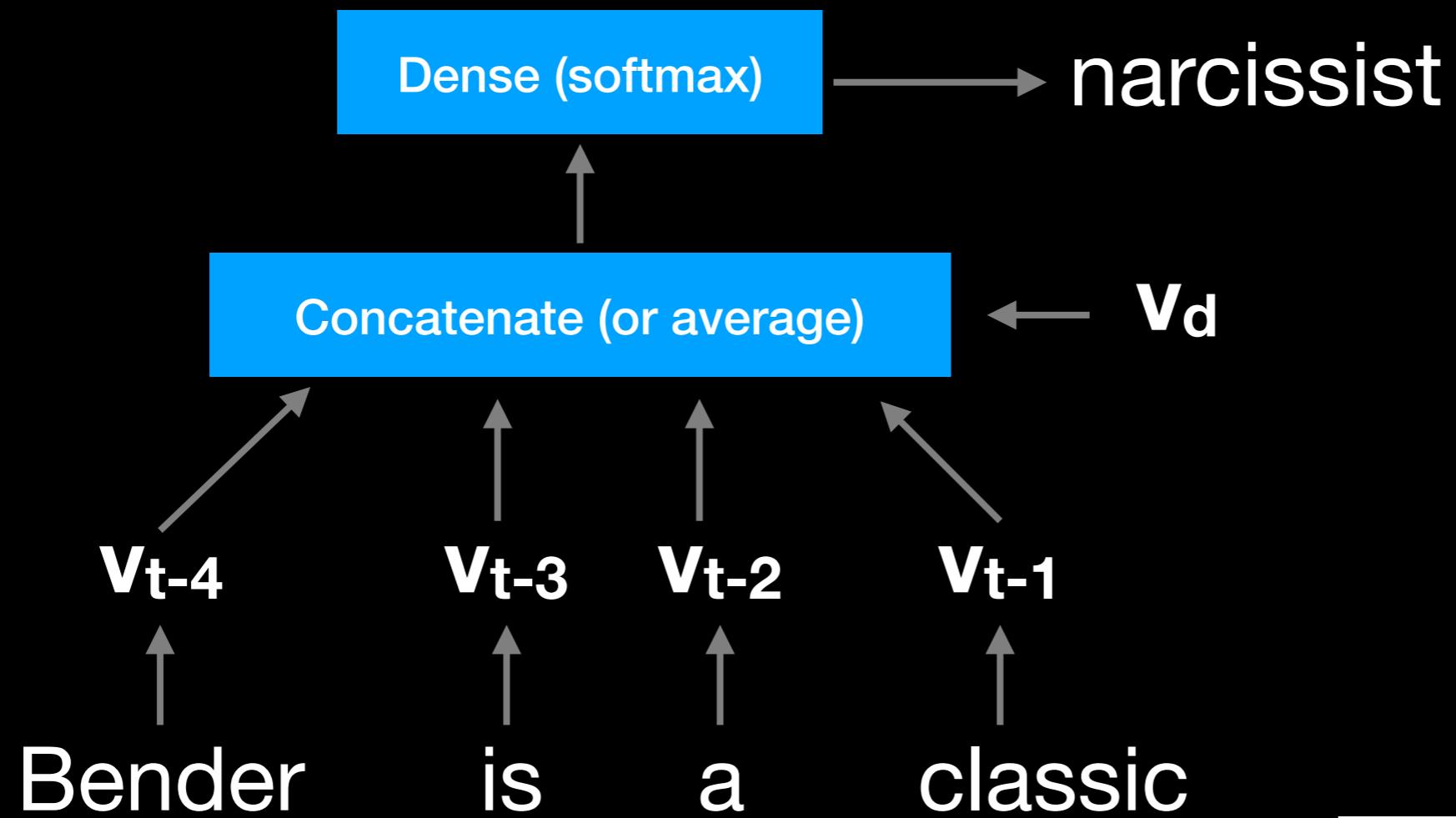


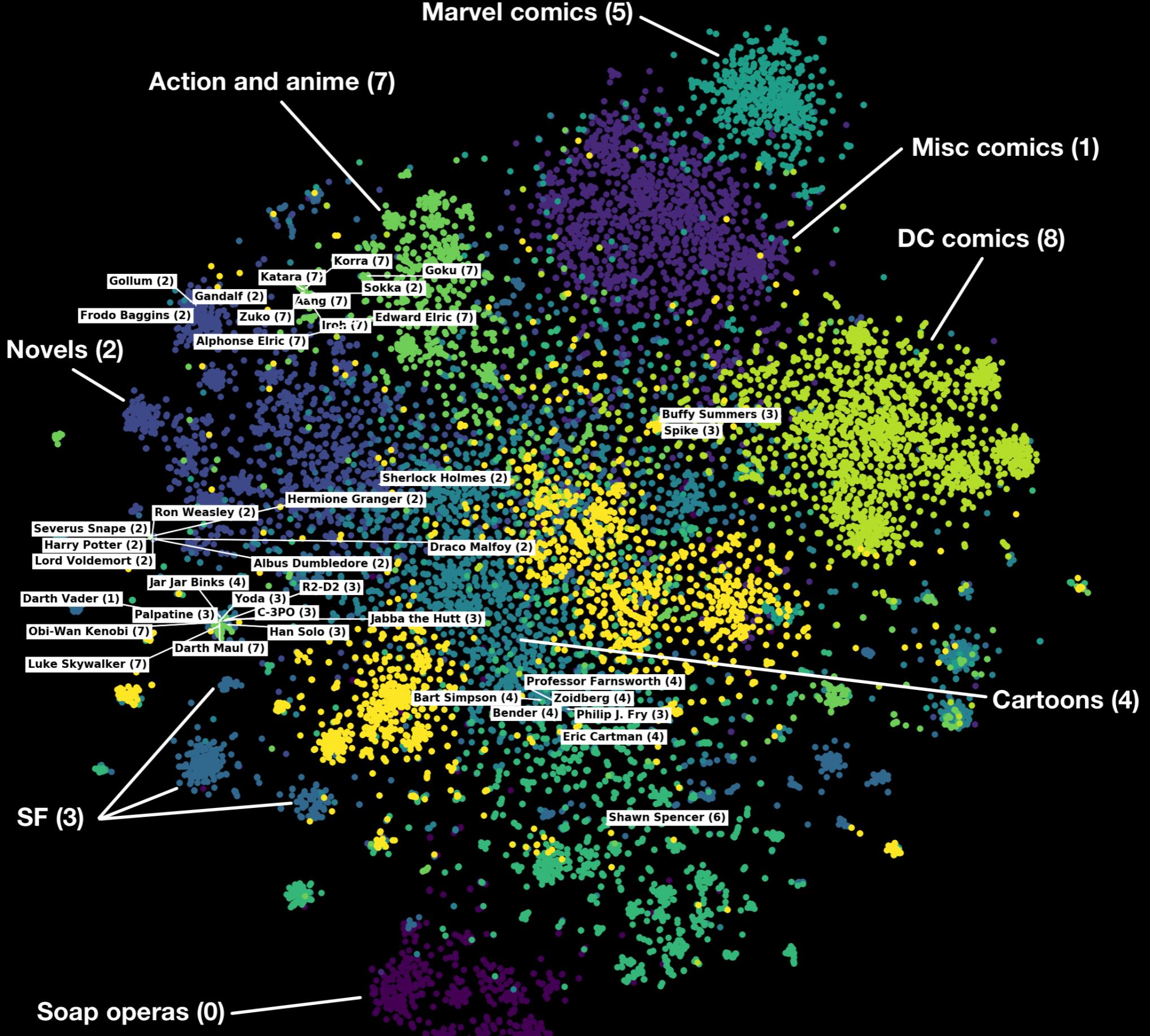
# Data set

- ~14,000 character articles from Wikipedia
- 23,616,309 raw words
- 1 to 5 million relevant ones



# Doc2Vec







Gollum (2)  
Frodo Baggins (2)  
Gandalf (2)  
Zuko (7)  
Alphonse Elric (7)  
Korra (7)  
Katara (7)  
Aang (7)  
Iroh (7)  
Sokka (2)  
Edward Elric (7)

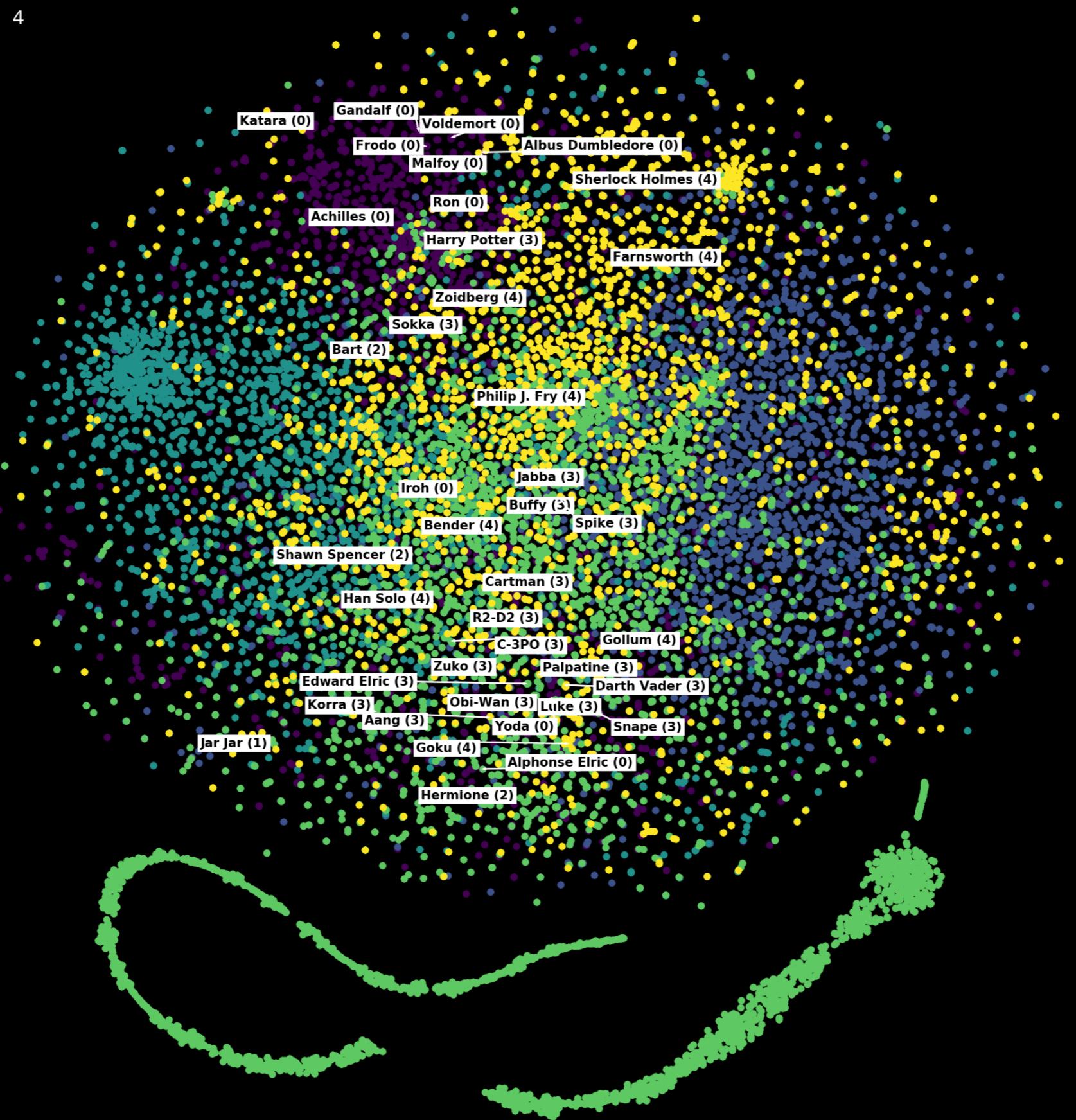
Buffy Summers (3)  
Spike (3)  
Sherlock Holmes (2)  
Ron Weasley (2)  
Hermione Granger (2)  
Severus Snape (2)  
Harry Potter (2)  
Lord Voldemort (2)  
Darth Vader (1)  
Palpatine (3)  
Obi-Wan Kenobi (7)  
Luke Skywalker (7)  
Jar Jar Binks (4)  
Albus Dumbledore (2)  
Yoda (3)  
C-3PO (3)  
Han Solo (3)  
Darth Maul (7)  
Draco Malfoy (2)  
Jabba the Hutt (3)

Professor Farnsworth (4)  
Bart Simpson (4)  
Zoidberg (4)  
Bender (4)  
Philip J. Fry (3)  
Eric Cartman (4)

Shawn Spencer (6)

**"<SUBJECT> is the mad scientist proprietor of  
the <ENTITY> delivery service , for whom the  
main characters work ."**

- 0
- 1
- 2
- 3
- 4



# Raw text similarity

```
1 model.docvecs.most_similar('Yoda')  
  
[ ('Obi-Wan Kenobi', 0.8464793562889099),  
 ('Count Dooku', 0.8301168084144592),  
 ('R2-D2', 0.8243268132209778),  
 ('Qui-Gon Jinn', 0.82326340675354),  
 ('C-3PO', 0.8140591382980347),  
 ('Luke Skywalker', 0.8125448822975159),  
 ('Palpatine', 0.7923175096511841),  
 ('Admiral Ackbar', 0.7624970078468323),  
 ('Darth Maul', 0.7532809376716614),  
 ('Jar Jar Binks', 0.7466059923171997)]
```

# Filtered text similarity

```
1 d2v_model.docvecs.most_similar('Yoda')  
  
[ ('Qui-Gon Jinn', 0.6633567810058594),  
 ('Luke Skywalker', 0.6167474985122681),  
 ('Optimus Prime', 0.6107190251350403),  
 ('Scourge (Transformers)', 0.5825557708740234),  
 ('Kanan Jarrus', 0.5685822367668152),  
 ('Zuko', 0.5668112635612488),  
 ('Gimli (Middle-earth)', 0.565727710723877),  
 ('Mokujin', 0.5602509379386902),  
 ('Albus Dumbledore', 0.5537698268890381),  
 ('Beorn', 0.5510784387588501)]
```

# Future directions

- I still have mistags and other noise.
- Implement Doc2Vec in Keras/Tensorflow.
- Improve sentence extractor and apply to plot summaries.
- Understand at the axes of the space (axis of evil, et cetera).