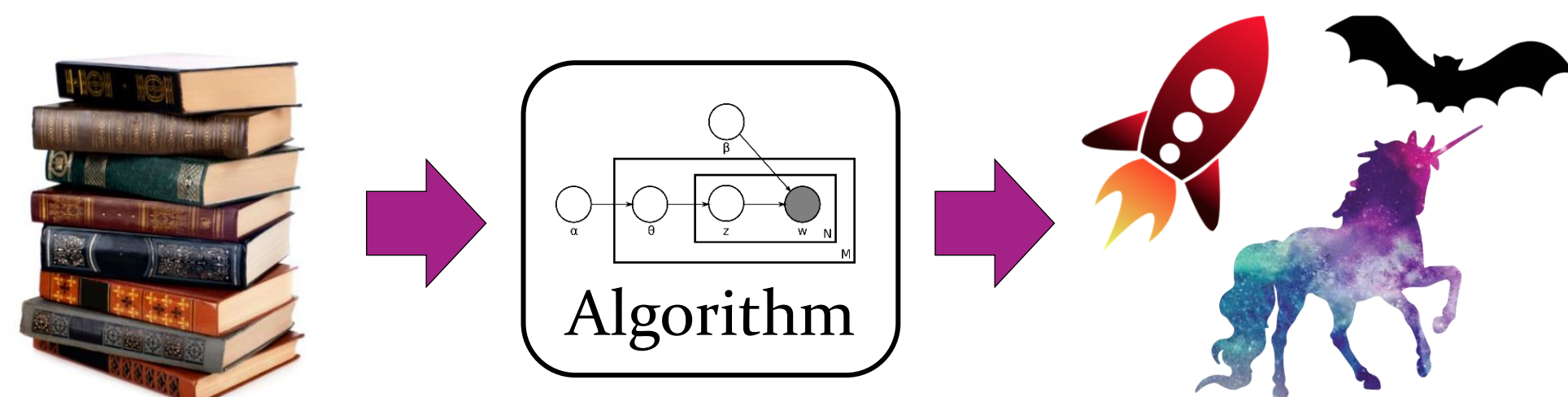# Authorless Topic Models: Biasing Models *Away* from Known Structure
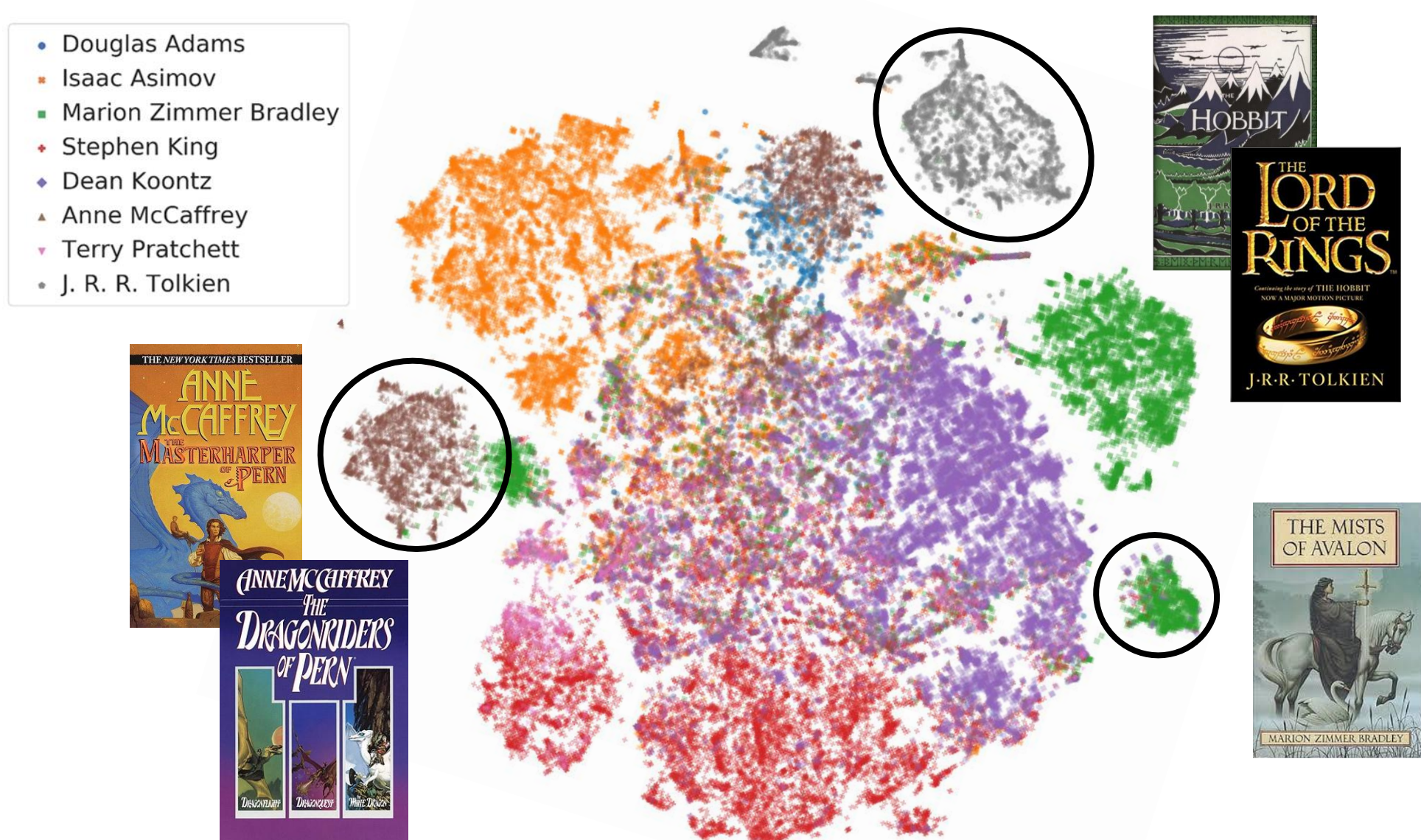
Laure Thompson
laurejt@cs.cornell.edu

David Mimno
mimno@cornell.edu
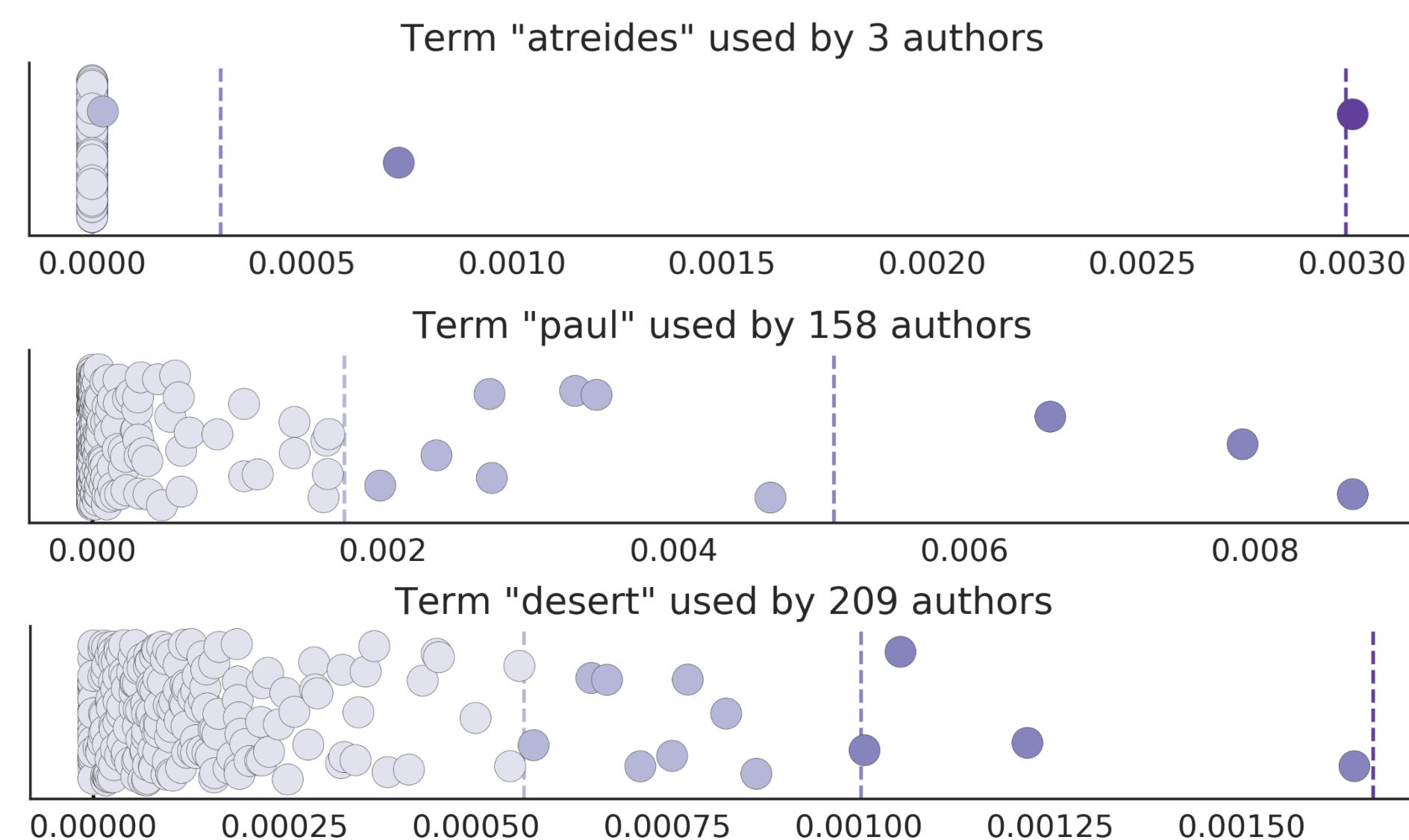
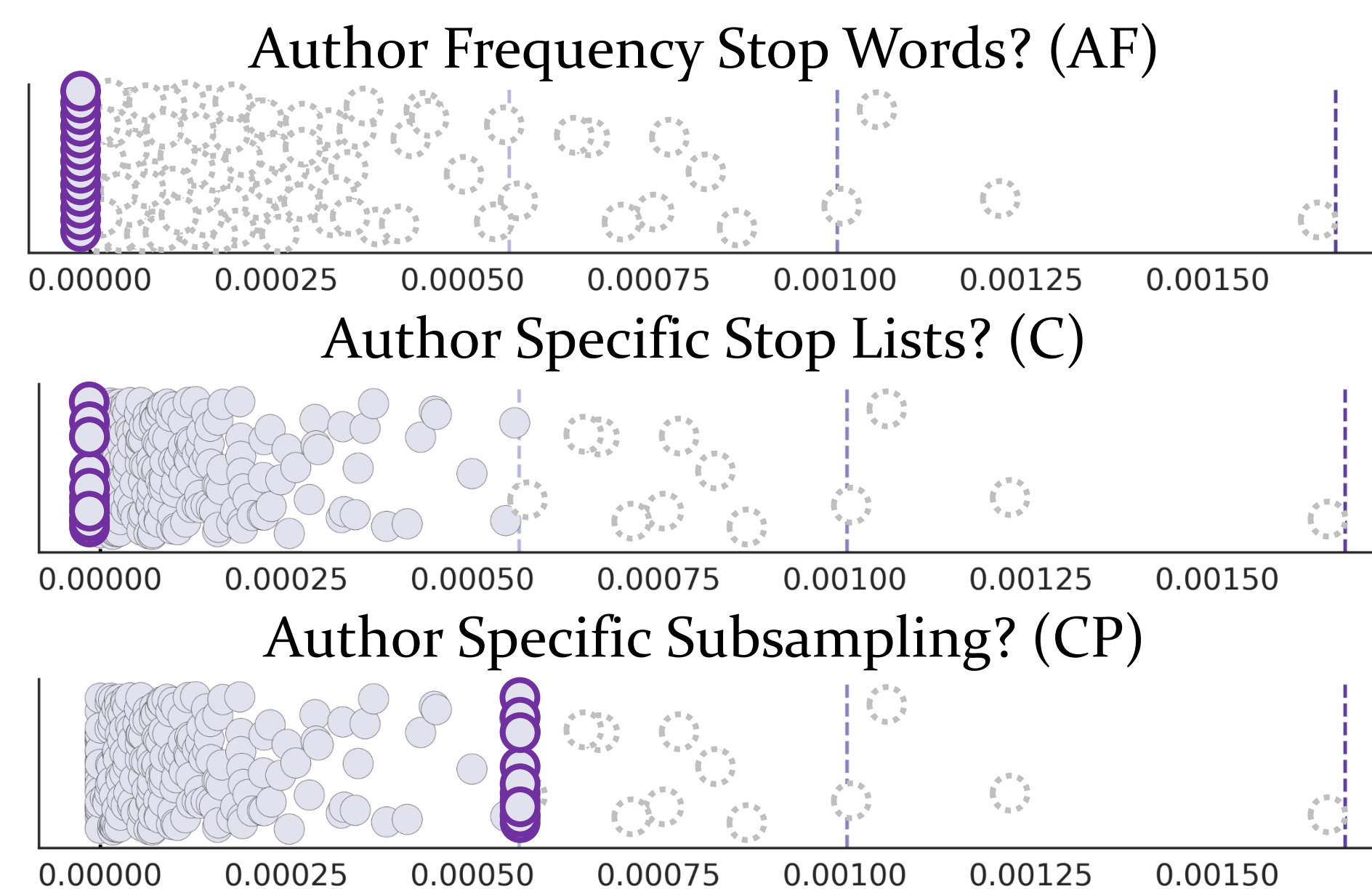## Setting: Mapping Speculative Fiction
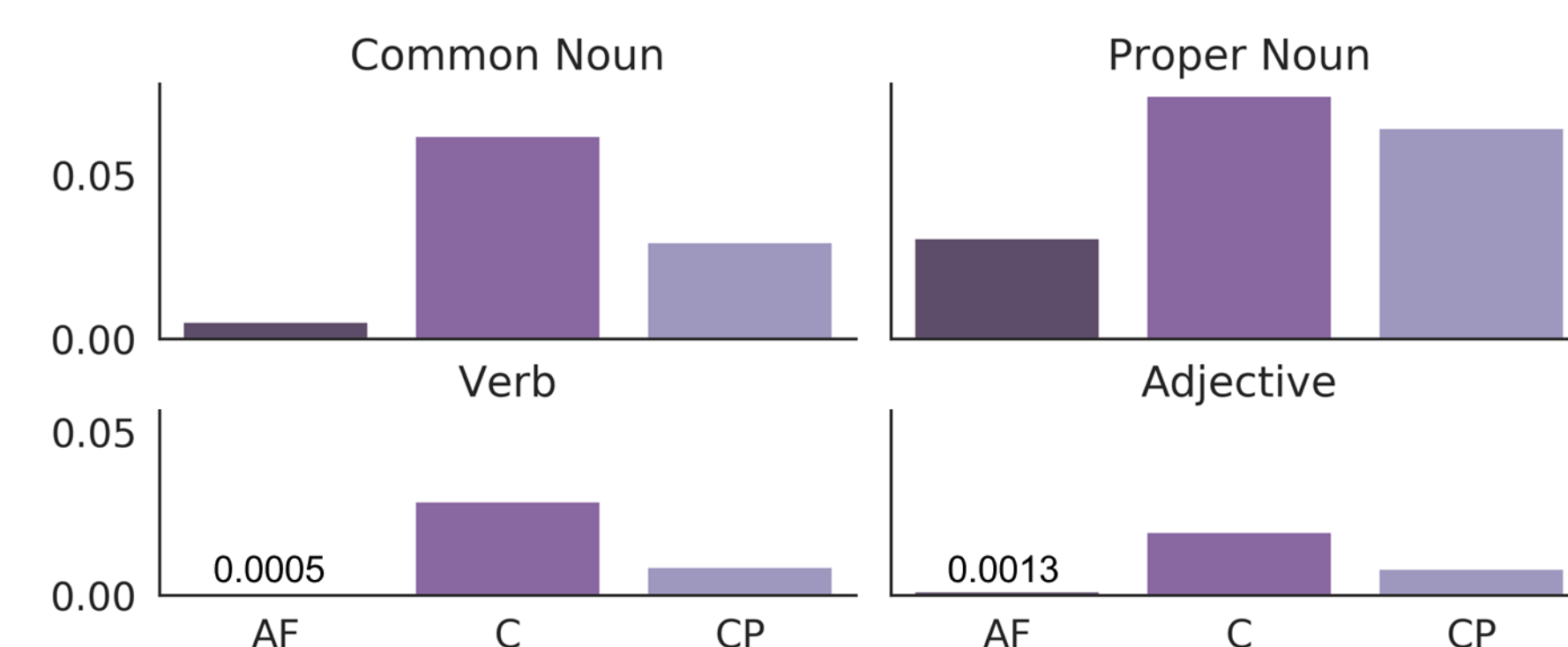

...but topics correlate with author/series!

Author correlations can be hard to see!

A	B	Top Words
✓	✓	school professor work university years research science students student college study class
✓	✗	sand **pirx** mars desert roger dust rock **bass** dunes crater martian **jeffries kirov** dune
✗	✓	old night yes cried town last men **rocket** god years hands house upon stood wind boy shut
✓	✗	jack emma **malenfant** trip janet michael ing **wireman leonard nemoto** sally **jeannine**
✗	✗	**f'lar lessa weyr robinton hold dragon f'nor lord dragons benden rider bronze harper**

A. Does the topic have high author diversity?
B. Do its authors use similar words?

What about word-level correlations?

Preprocessing
Purposeful Data Modification!

Author Frequency Stop Words? (AF)

Author Specific Stop Lists? (C)

Author Specific Subsampling? (CP)

Removing more than names!

Author Diversity vs. Topic Stability
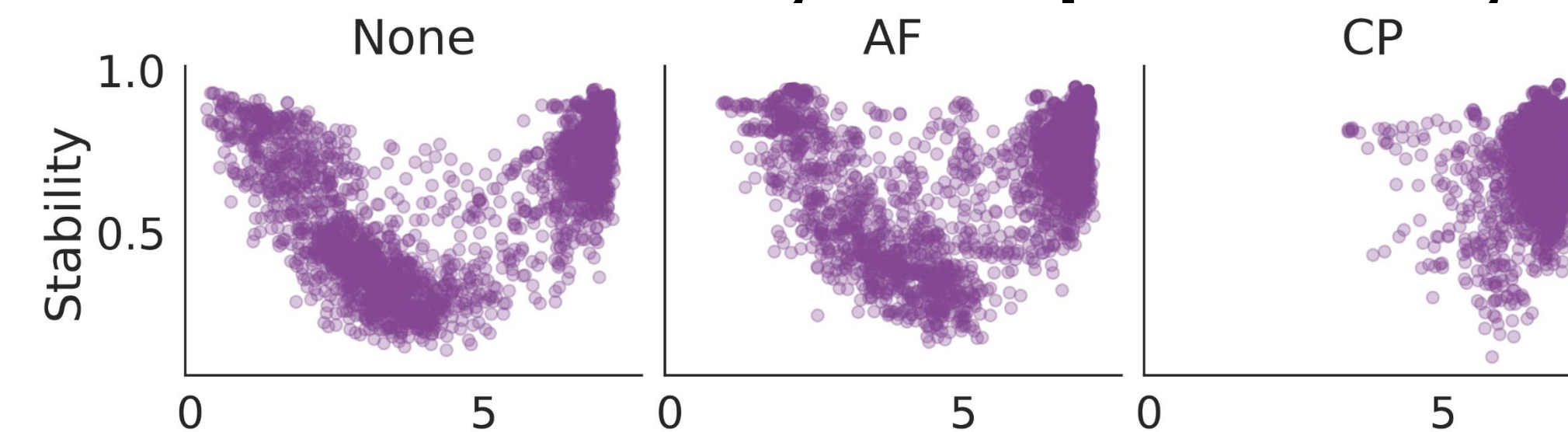
Subsampling yields cross-cutting topics

Topics look more cohesive, meaningful

A	B	Top Words
✓	✓	professor university college student students research school science work years year
✓	✓	sand desert rock mountains mountain dust land surface plain water across ground earth
✓	✓	night wind dark cold sky moon light stars air gone sat old day sun last stood white under
		No Match!
✗	✓	lord hold between master queen star enough turns high good such young ruth hall red wings