Topic Models

CS 490A, Fall 2021

Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs490a_f21

Brendan O'Connor & Laure Thompson

College of Information & Computer Sciences
University of Massachusetts Amherst

Administrivia - Posters

- Posters due **today** by 7pm
- Posters **must** be submitted tonight in order to be uploaded to the gather.town platform

Initial poster session assignments available [here]

- Session A: 4:00-4:45pm
- Session B: 4:45-5:30pm
- Visit other groups when it's **not** your session

Administrivia - Posters

On poster content / design:

- Posters will cover your **current** progress, not final report
- Must be horizontal / landscape in shape!
- Keep it simple! This is a visual aid.

JEOPARDY! Category Analysis

Evan Risas & Alisa Kotliarova

Task: Analyze each question-answer pair to determine which broad category it most closely fits, then predict category frequency for future Jeopardy games.

Dataset

200,000+ Jeopardy! Question & Answer pairs

Approach

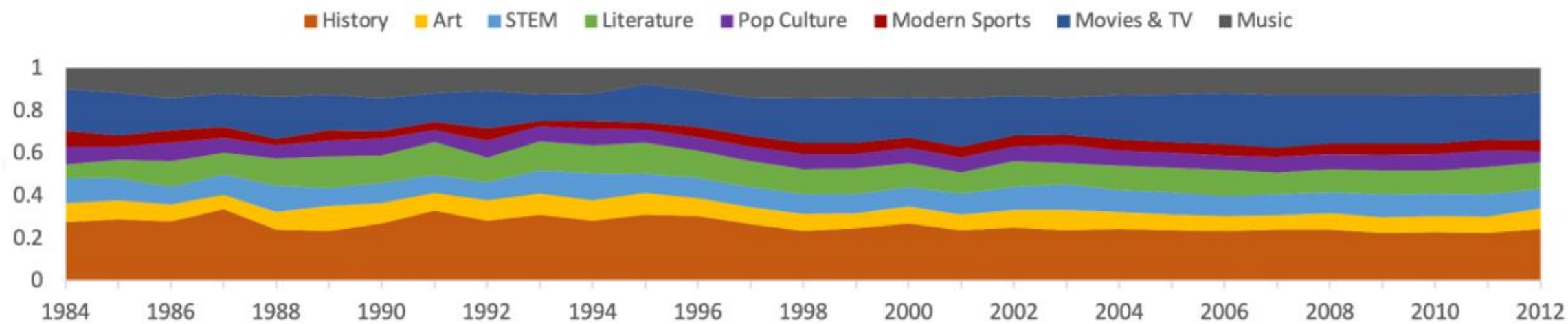
Classify into custom categories using NLP model built on Word2Vec

Observe

Examine category popularity over time

Predict

Predict future category frequency



Administrivia

- Final Project Poster Session, Tuesday, 12/7
- HW4 due next Friday, 12/10
- Final Report due Thursday, 12/16
- SRTI Course Surveys open!
  - There are two for this course, one for each instructor
  - Access surveys here: https://owl.umass.edu/partners/courseEvalSurvey/uma/

What is a topic model?

What is a topic?

music song sing singing sang play
dance played playing songs began voice

What is a topic?

music song sing singing sang play
dance played playing songs began voice

What is a topic?

music song sing singing sang play
dance played playing songs began voice

snow cold ice winter wind frozen warm
white air weather frost night

What is a topic?

music song sing singing sang play
dance played playing songs began voice

snow cold ice winter wind frozen warm
white air weather frost night

What is a topic?

music song sing singing sang play dance played playing songs began voice

snow cold ice winter wind frozen warm white air weather frost night

computer screen data program information image system monitor display code appeared console

What is a topic?

music song sing singing sang play dance played playing songs began voice

snow cold ice winter wind frozen warm white air weather frost night

computer screen data program information image system monitor display code appeared console

What is a topic model?

For a collection of documents, a **topic model** learns…

(1) A set of K **“topics”**

What is a topic model?

For a collection of documents, a **topic model** learns…

(1) A set of K **"topics"** not necessarily topical!

"topic" = collection of related words

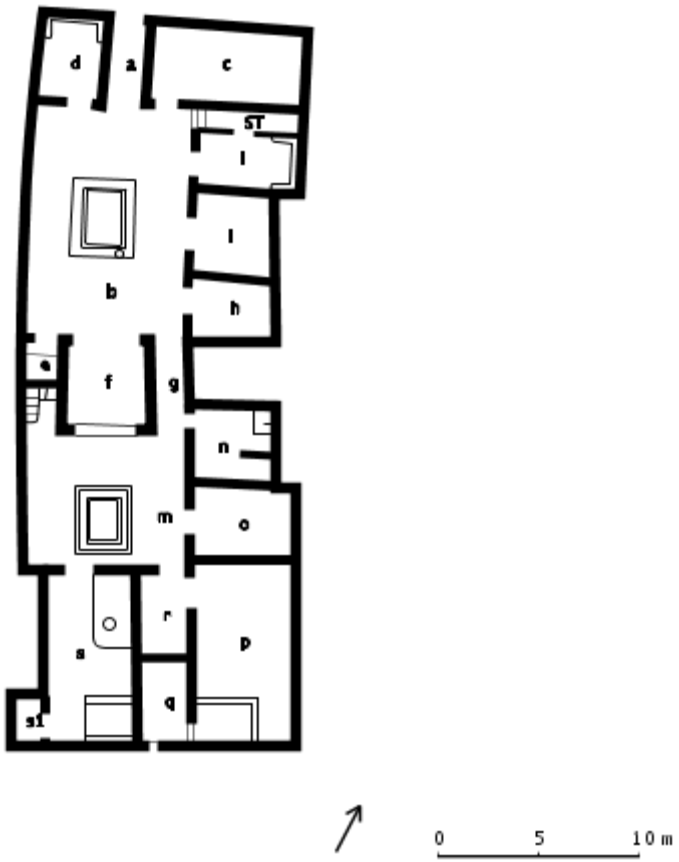
What is a topic model?

For a collection of documents, a **topic model** learns…

(1) A set of K **“topics”**

(2) A **topic distribution** for each document

Applicable to more than words



	Type 3 (*atrium*)
0.279	pottery amphora/amphoretta/hydria, building material, stairway, impluvium/compluvium, puteal/puteal fragment
0.108	chest/cupboard fitting, chest/cista, glass bottle/flask/pyxis, chest fitting, cupboard
0.083	coin, ceramic lamp, small glass bottle, architectural fitting, jewelry
0.077	pottery jug, ceramic lamp, table/table fittings/table base, terra sigillata bowl/cup, seashell/conch/snail shell
	Type 4 (*cubiculum*)
0.122	recess, built-in cupboard, stairway, niche, unidentified fixture/mound
0.116	shelving/mezzanine/suspension nails, recess, pottery amphora/amphoretta/hydria, cistern head, pottery pot
0.092	coin, ceramic lamp, small glass bottle, architectural fitting, jewelry
0.065	pottery jug, pottery amphora/amphoretta/hydria, pottery pot, pottery jar/vase, pottery plate/dish/tray

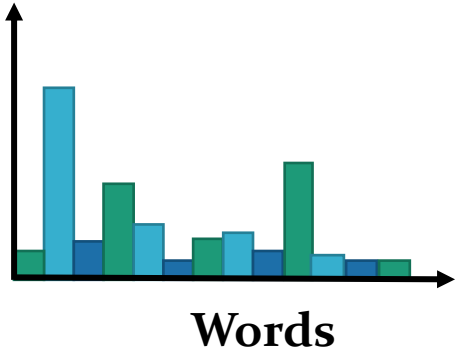
Mimno 2011

Latent Dirichlet Allocation

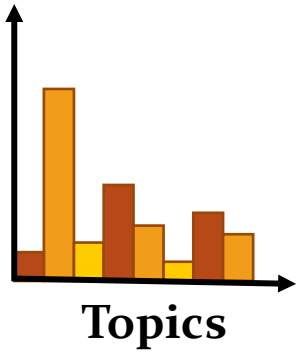
Blei et al. 2002

LDA: A Generative Model

Topics



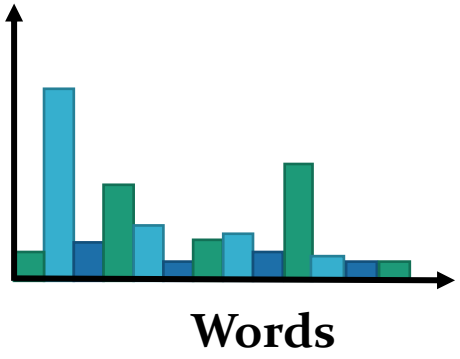
Documents



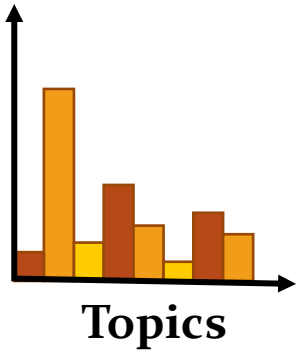
Blei et al. 2002

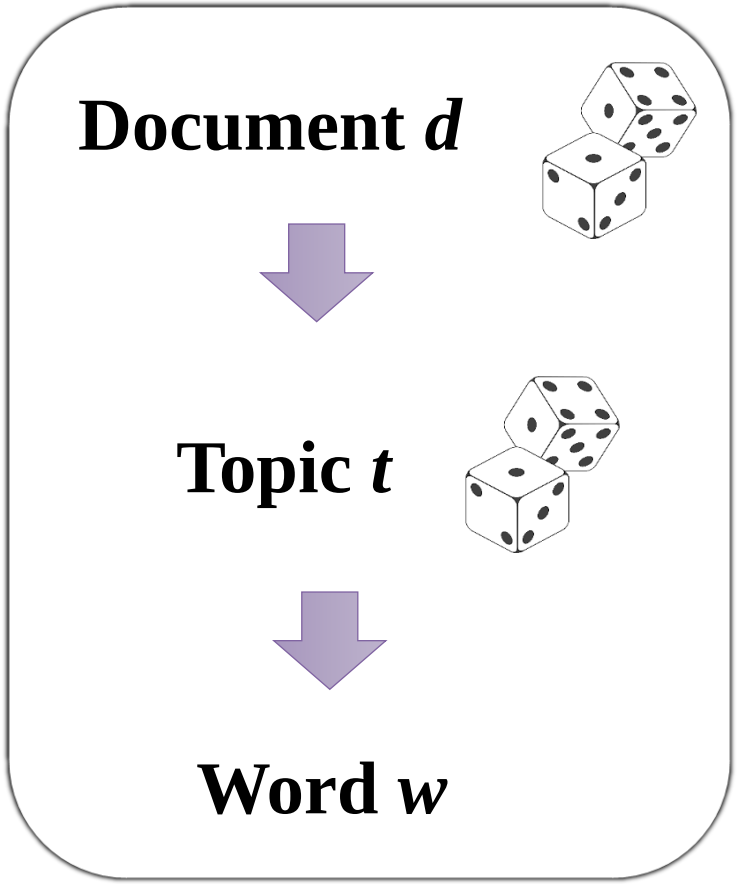
LDA: A Generative Model

Topics



Documents





Blei et al. 2002

Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

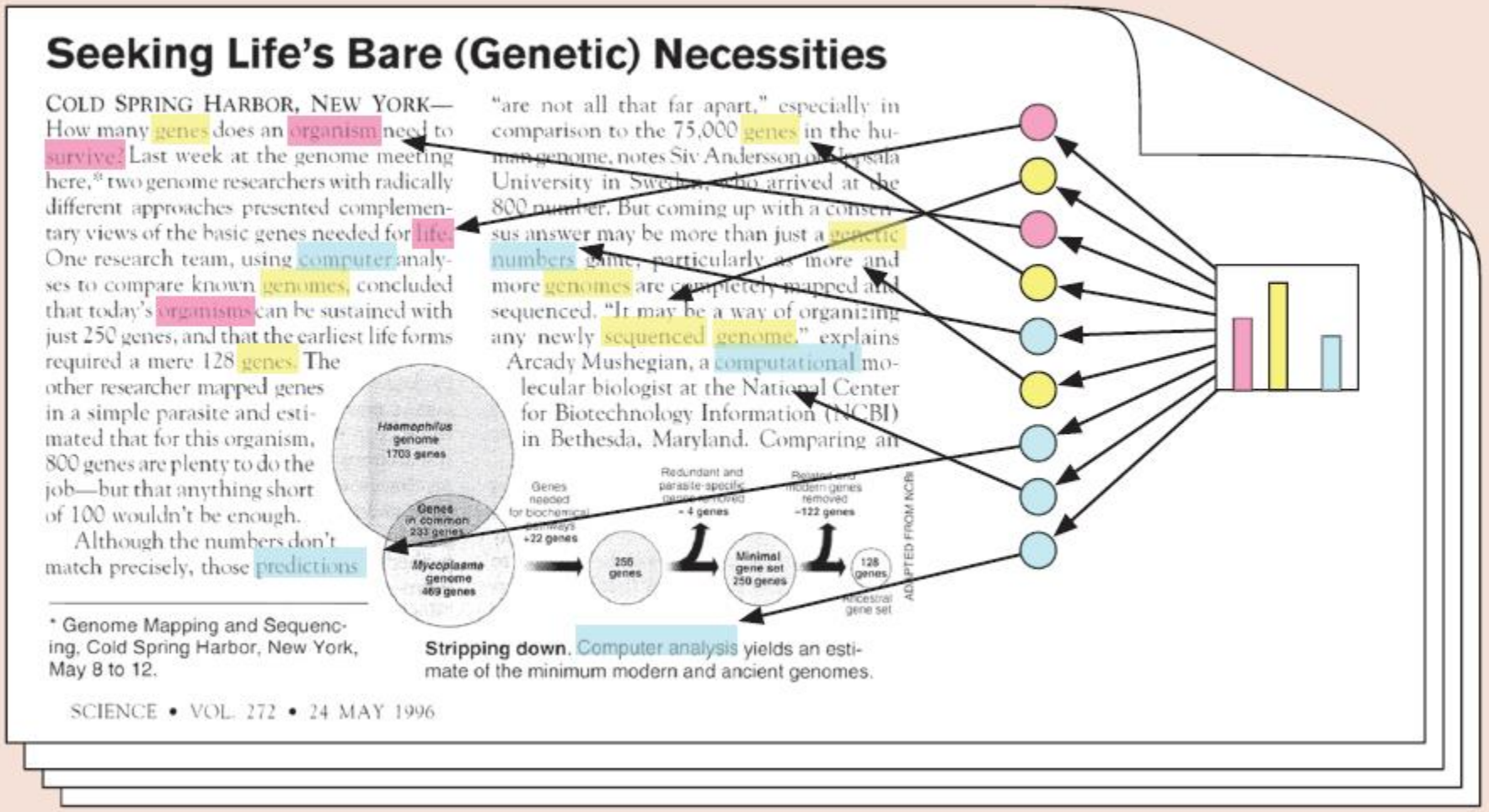
life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

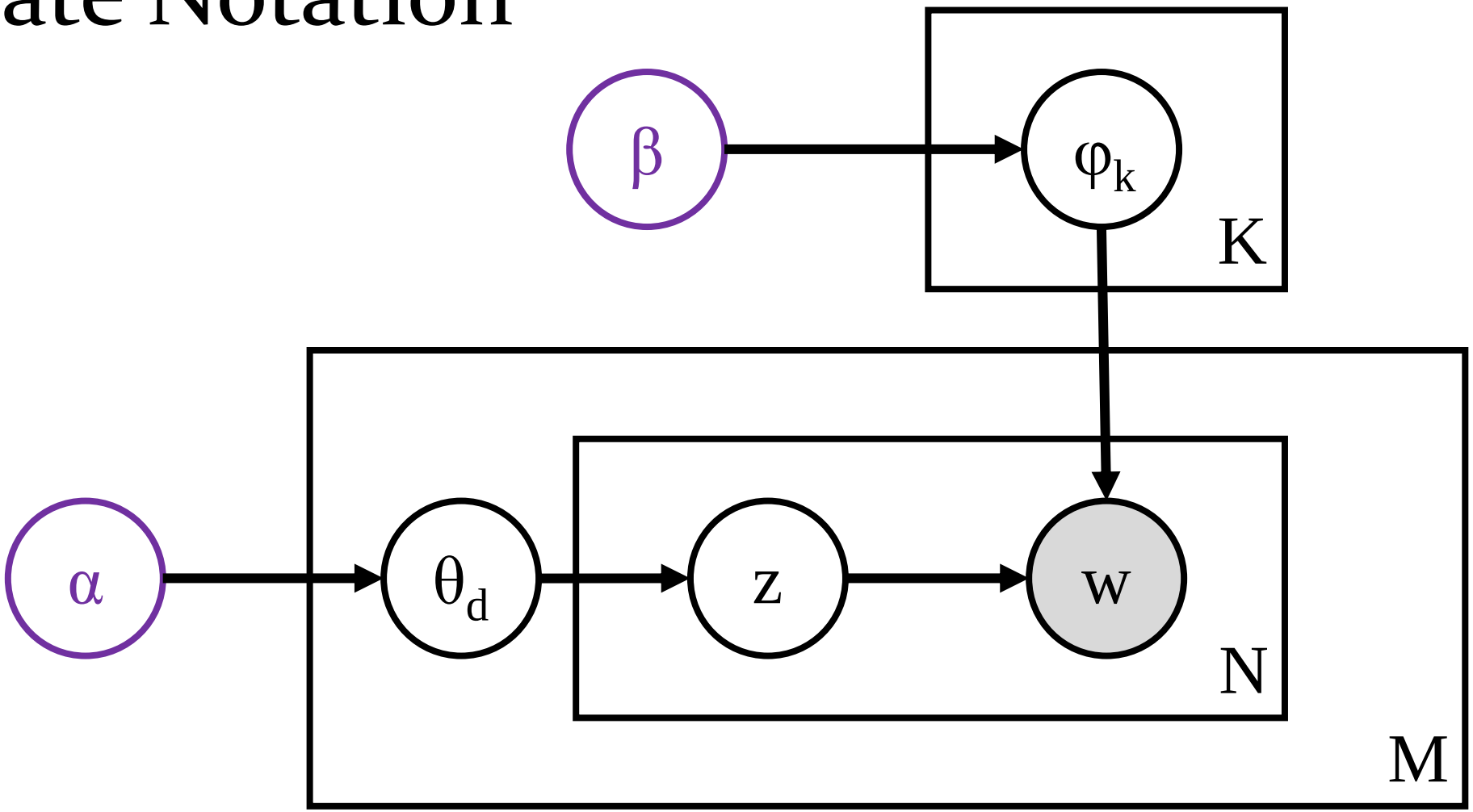
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Blei et al. 2012

Plate Notation



LDA: Contextualized, Interpretable Latent Space

Documents: K-dimensional vectors

Dimensions correspond to topics, which are both interpretable and contextualized.

Topics = Interpretable Dimensions

music song sing singing sang play
dance played playing songs began voice

snow cold ice winter wind frozen warm
white air weather frost night

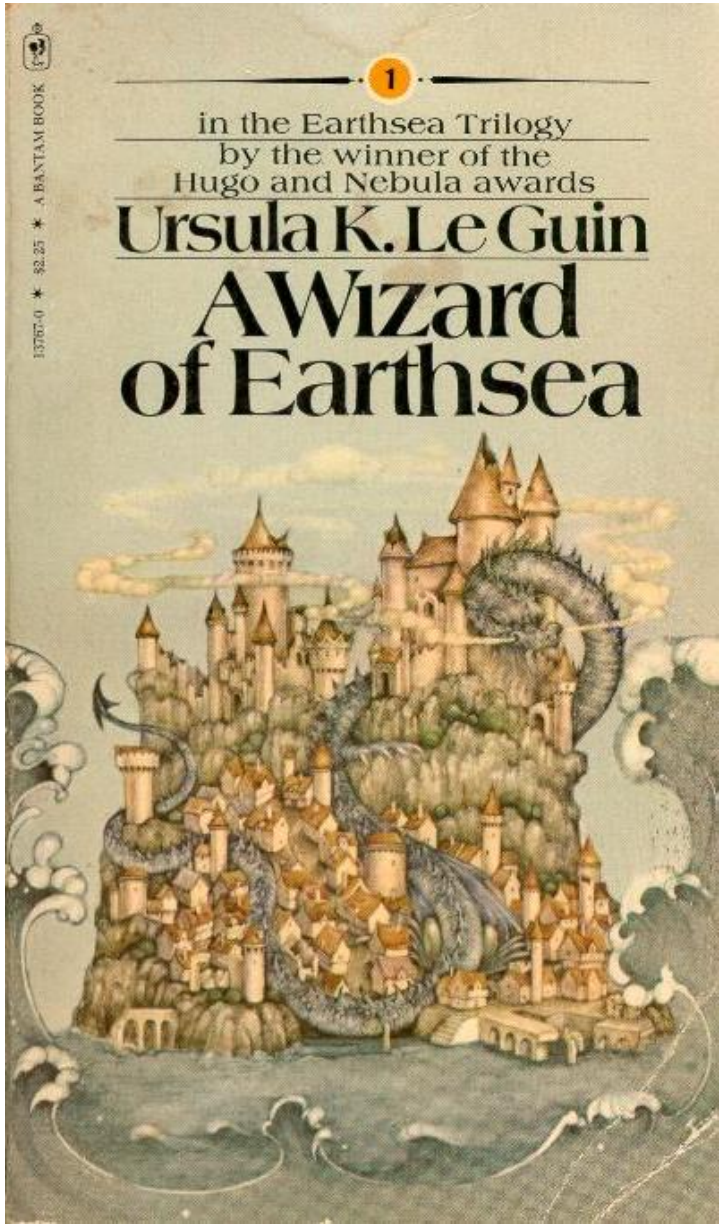
computer screen data program
information image system monitor
display code appeared console

LDA: Contextualized, Interpretable Latent Space

Documents: K-dimensional vectors

Dimensions correspond to topics, which are both interpretable and contextualized.

Idea: Each token in the working corpus has a corresponding topic assignment



1 Warriors in the Mist

The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea. Of these some say the greatest, and surely the greatest voyager, was the man called Sparrowhawk, who in his day became both dragonlord and Archmage. His life is told of in the *Deed of Ged* and in many songs, but this is a tale of the time before his fame, before the songs were made.

He was born in a lonely village called Ten Alders, high on the mountain at the head of the Northward Vale. Below the village the pastures and plowlands of the Vale slope downward level below level towards the sea, and other towns lie on the bends of the River Ar; above the village only forest rises ridge behind

ridge to the stone and snow of the heights.

The name he bore as a child, Duny, was given him by his mother, and that and his life were all she could give him, for she died before he was a year old. His father, the bronze-smith of the village, was a grim unspeaking man, and since Duny's six brothers were older than he by many years and went one by one from home to farm the land or sail the sea or work as smith in other towns of the Northward Vale, there was no one to bring the child up in tenderness. He grew wild, a thriving weed, a tall, quick boy, loud and proud and full of temper. With the few other children of the village he herded goats on the steep meadows above the river-springs; and when he was strong enough to push and pull the long bellows-sleeves, his father made him work as smith's boy, at a high cost in blows and whippings. There was not much work to be got out of Duny. He was always off and away; roaming deep in the forest, swimming in the pools of the River Ar that like all Gontish rivers runs very quick and cold, or climbing by cliff and scarp to the heights above the forest, from which he could see the sea, that broad northern ocean where, past Perregal, no islands are.

A sister of his dead mother lived in the village. She had done what was needful for him as a baby, but she had business of her own and once he could look after himself at all she paid no more heed to him. But one day when the boy was seven years old, untaught and knowing nothing of the arts and powers that are in the world, he heard his aunt crying out words to a goat which had jumped up onto the thatch of a hut and would not come down: but it came jumping when she cried a certain rhyme to it. Next day herding the longhaired goats on the meadows of High Fall, Duny shouted to them the words he had heard, not knowing their use or meaning or what kind of words they were:

The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

Sea/Ocean: sea water boat island beach ship ocean …

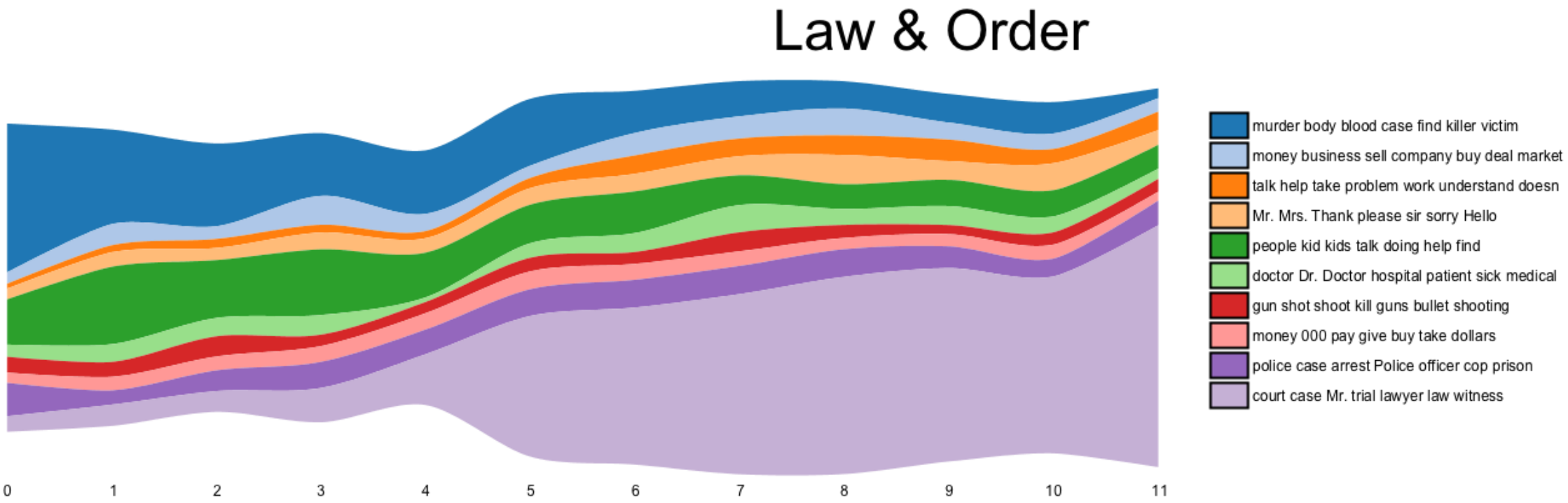
Magic: magic spell witch power demon wizard magician …

Topic Modeling for Collection Exploration

Topic Label	Top Words	Top Classics	Sample Review
The Future (Dystopias)	world, people, society, human, war, future, new, power, history, political, science, fiction, thought, today, humans	1984 The Man in the High Castle Animal Farm Brave New World Do Androids Dream of Electric Sheep? Fahrenheit 451 The Handmaid's Tale	*This was typical Phillip K. Dick far, clever philosophical* ***science fiction*** *contemplating ideas about religion,* ***society*** *and in many ways what it is to be* ***human****…I felt that it provided clever parallels with the daily grind of* ***today's*** *modern world* *— Jonathan*

Walsh & Antoniak 2021

Topic Modeling for Collection Exploration



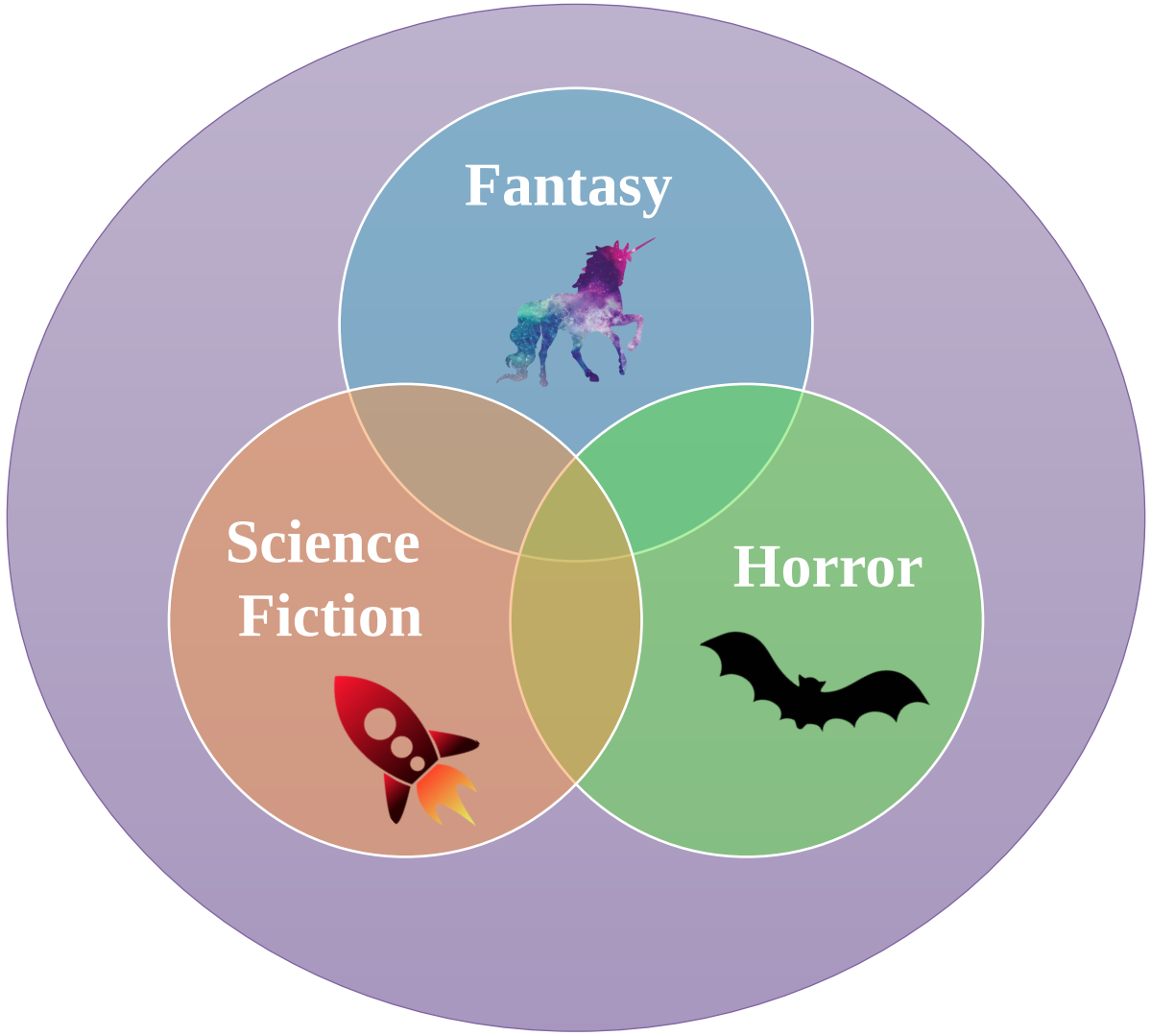
Schmidt 2014

Class Activity

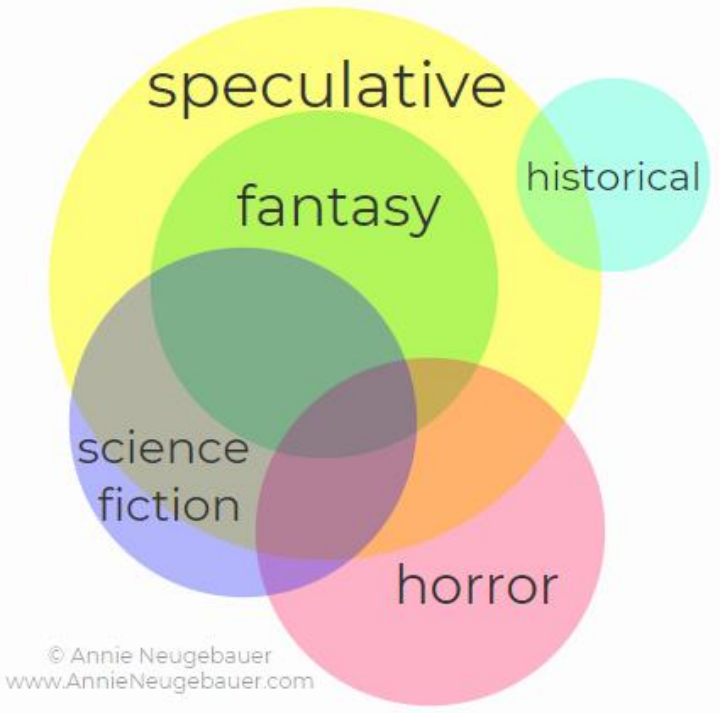
https://mimno.infosci.cornell.edu/jsLDA/

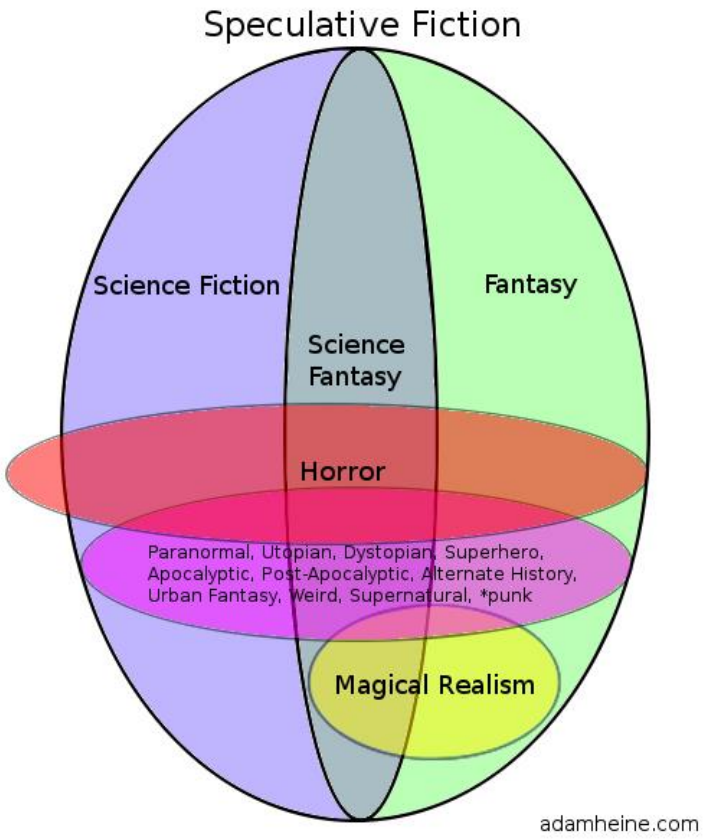
Studying Genre at Scale

What is Speculative Fiction?

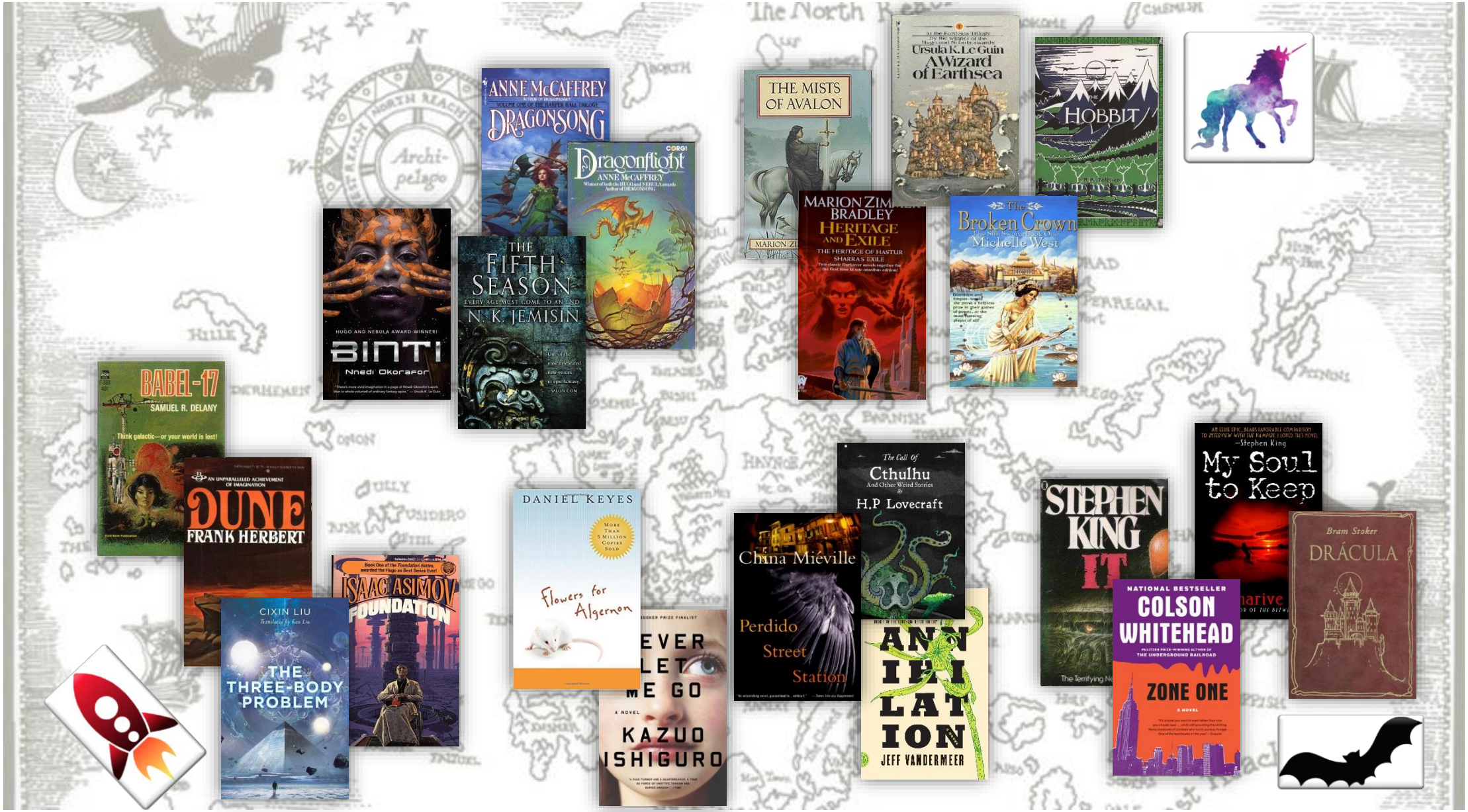


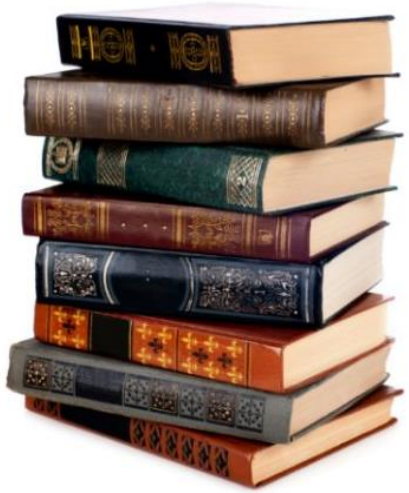
What is Speculative Fiction?

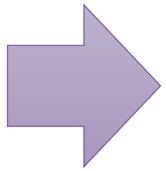


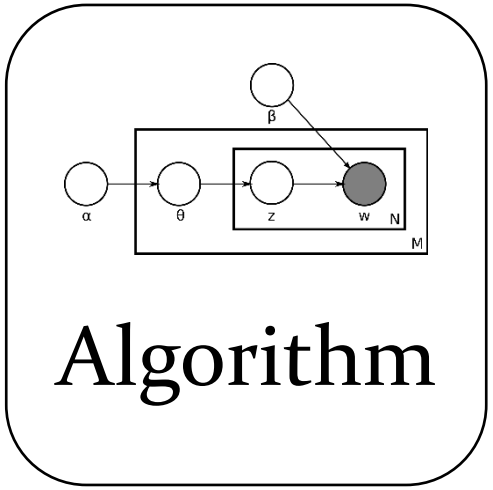


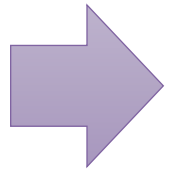
World Weaver Press

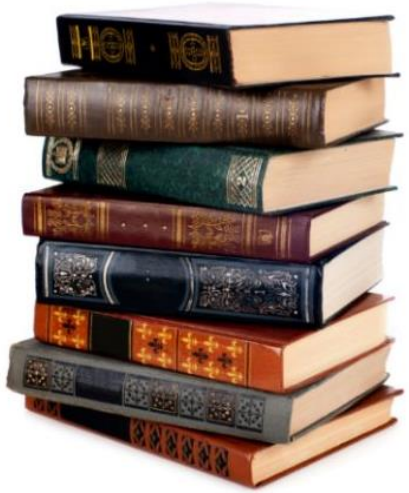

ANNE McCAFFREY
DRAGONSONG
Dragonflight
ANNE McCAFFREY
THE MISTS OF AVALON
Ursula K. Le Guin
A Wizard of Earthsea
THE HOBBIT
BINTI
Nnedi Okorafor
THE FIFTH SEASON
N. K. JEMISIN
MARION ZIMMER BRADLEY
HERITAGE AND EXILE
THE HERITAGE OF HASTUR
SHARRA'S EXILE
The Broken Crown
Michelle West
BABEL-17
SAMUEL R. DELANY
DUNE
FRANK HERBERT
ISAAC ASIMOV
FOUNDATION
CIXIN LIU
THE THREE-BODY PROBLEM
DANIEL KEYES
Flowers for Algernon
KAZUO ISHIGURO
China Miéville
Perdido Street Station
The Call Of Cthulhu
And Other Weird Stories
H.P Lovecraft
ANNIHILATION
JEFF VANDERMEER
STEPHEN KING
IT
My Soul to Keep
Stephen King
NATIONAL BESTSELLER
COLSON WHITEHEAD
ZONE ONE
Bram Stoker
DRÁCULA

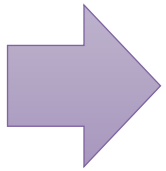


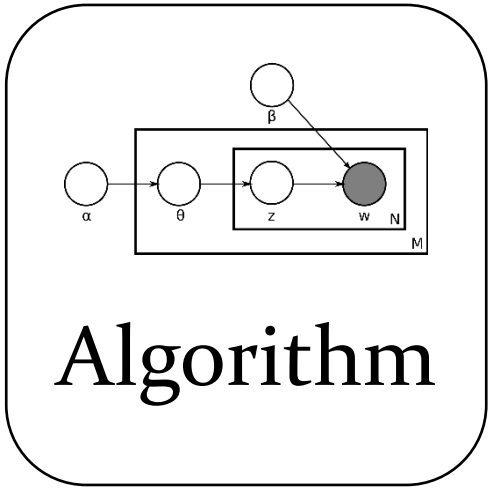



β
α
θ
z
w
N
M
Algorithm





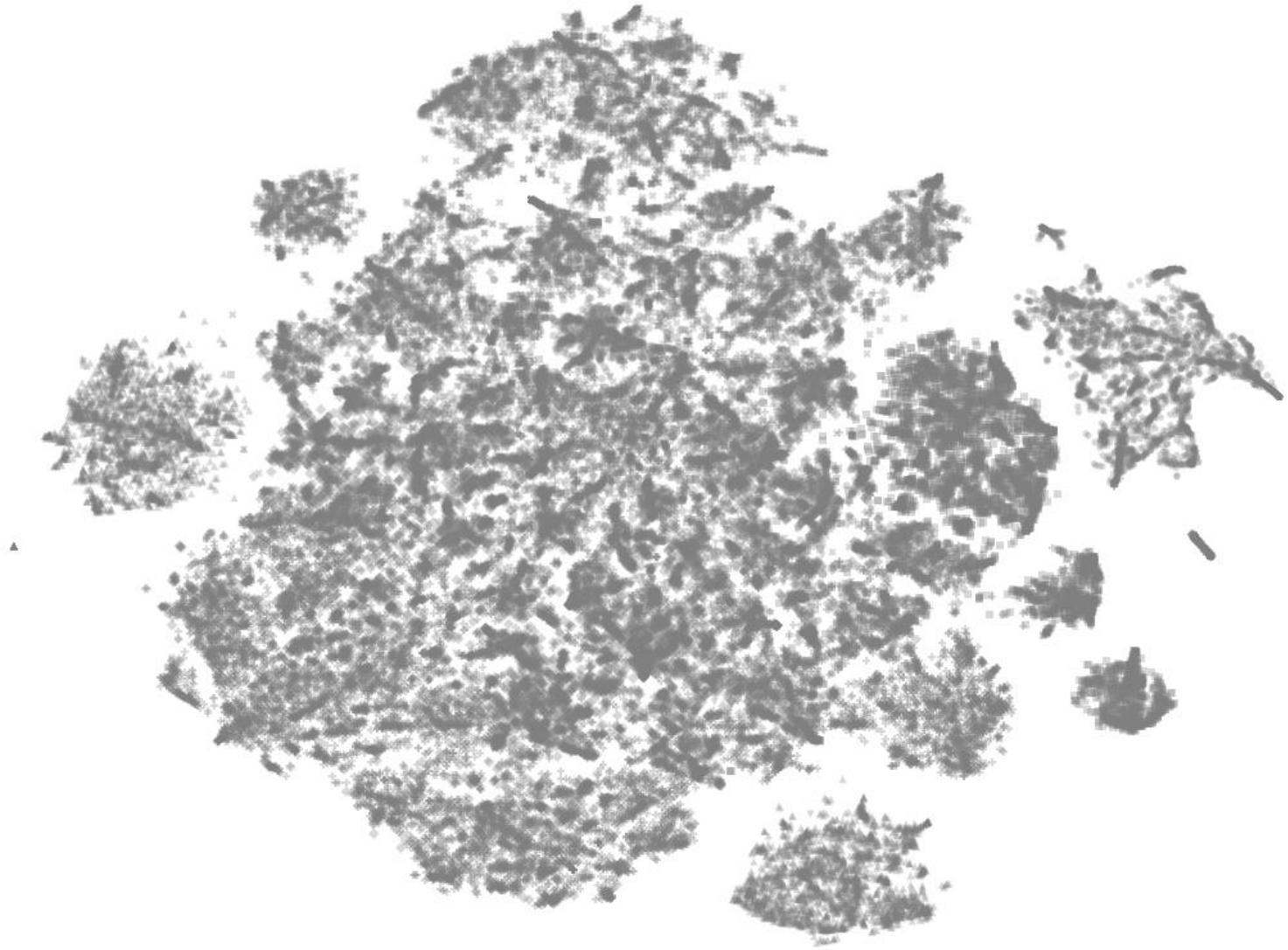


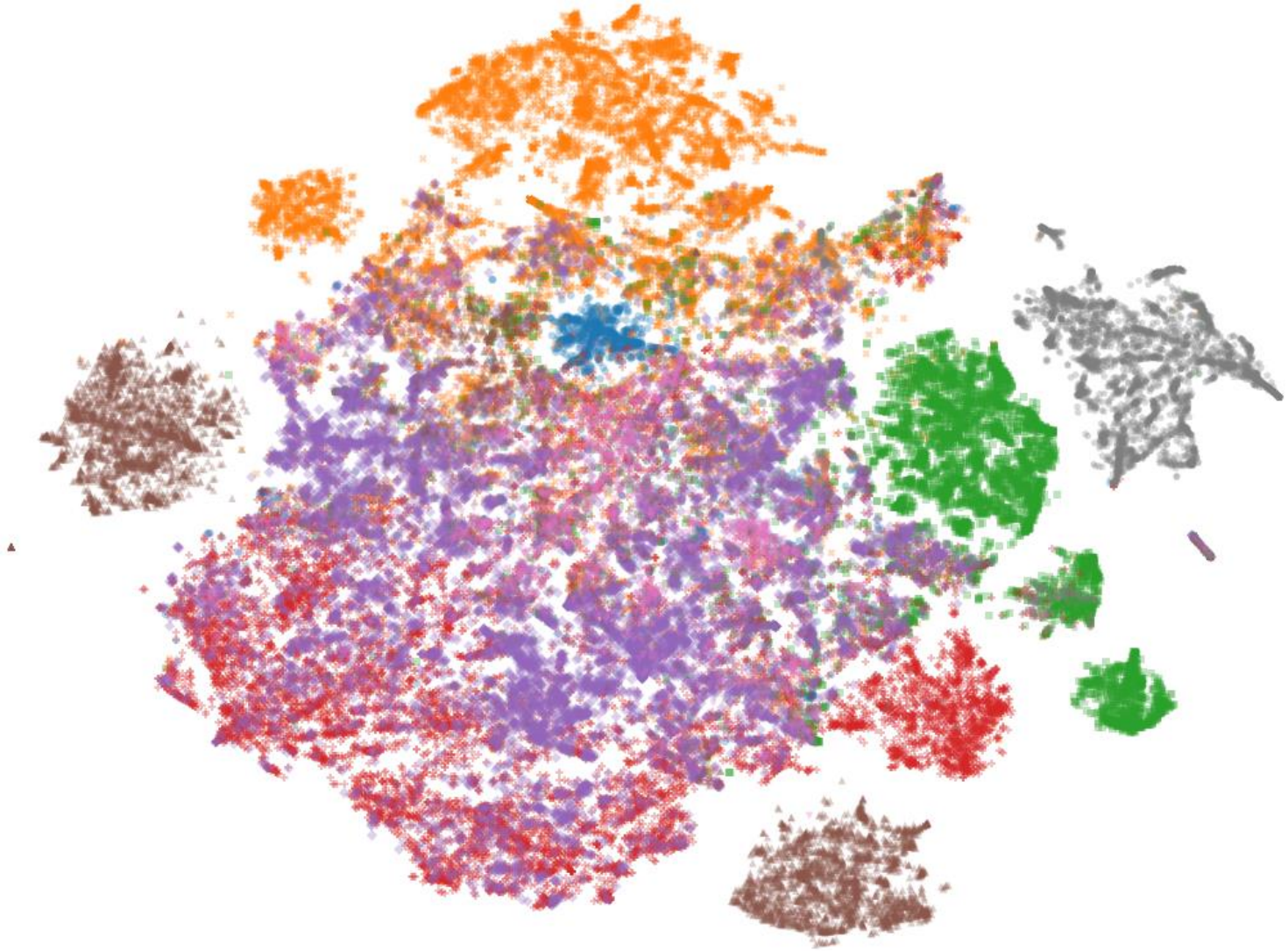

β
α
θ
z
w
N
M
Algorithm

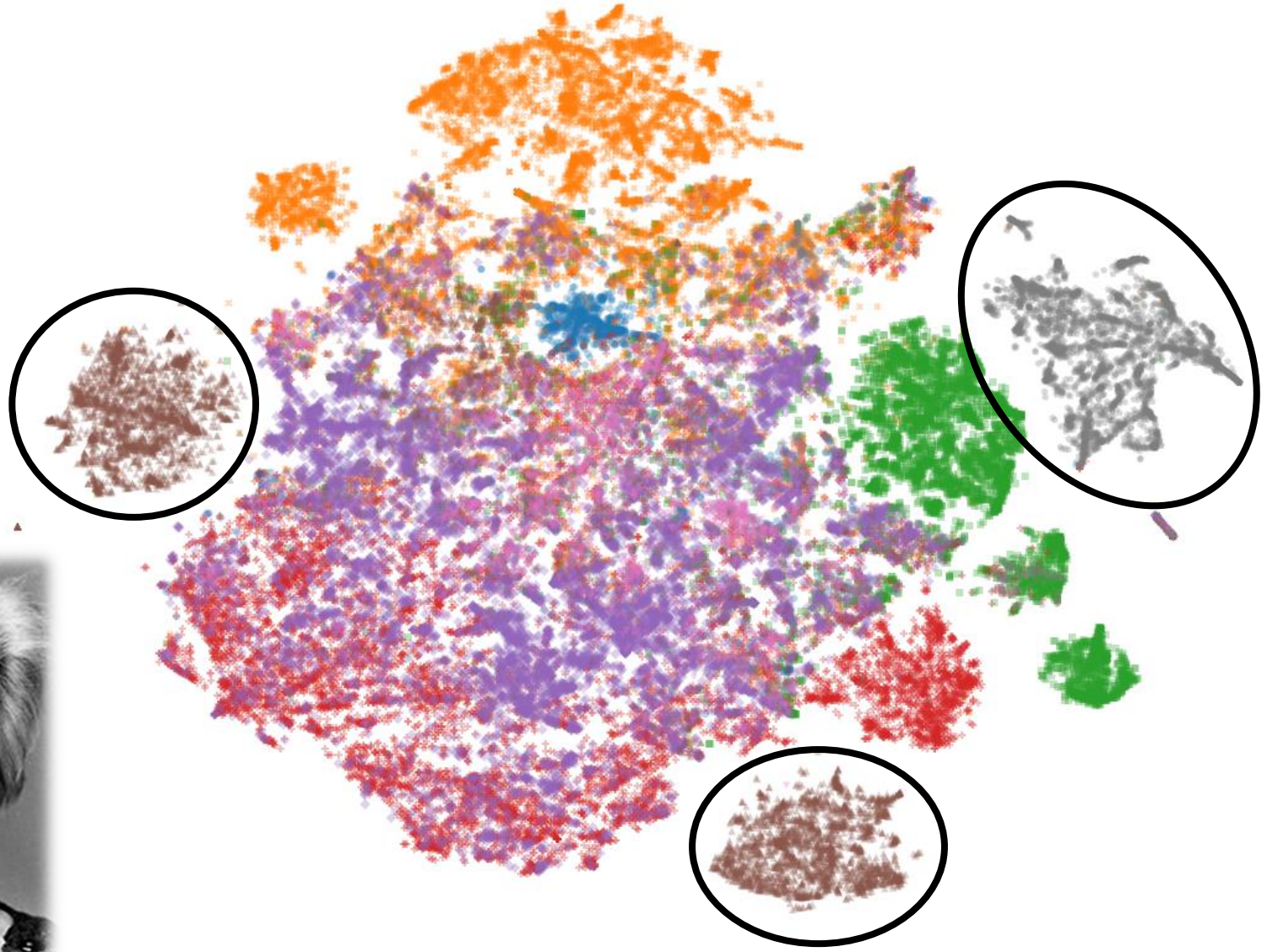
Authorless Topic Models: Biasing Models *Away* from Known Structure

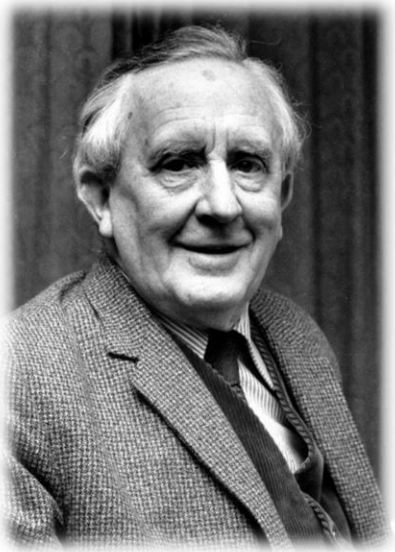
Laure Thompson and David Mimno.

COLING 2018: Best NLP engineering experiment.









What do author-correlated topics look like?

I’ll know it when I see it…

school professor work university years research science students student college study class year history scientific theory young…

I'll know it when I see it...

school professor work university years research science students student college study class year history scientific theory young...

f'lar lessa weyr robinton hold dragon f'nor lord dragons benden rider bronze harper thread mnementh brekke ramoth

…but it’s not always so easy

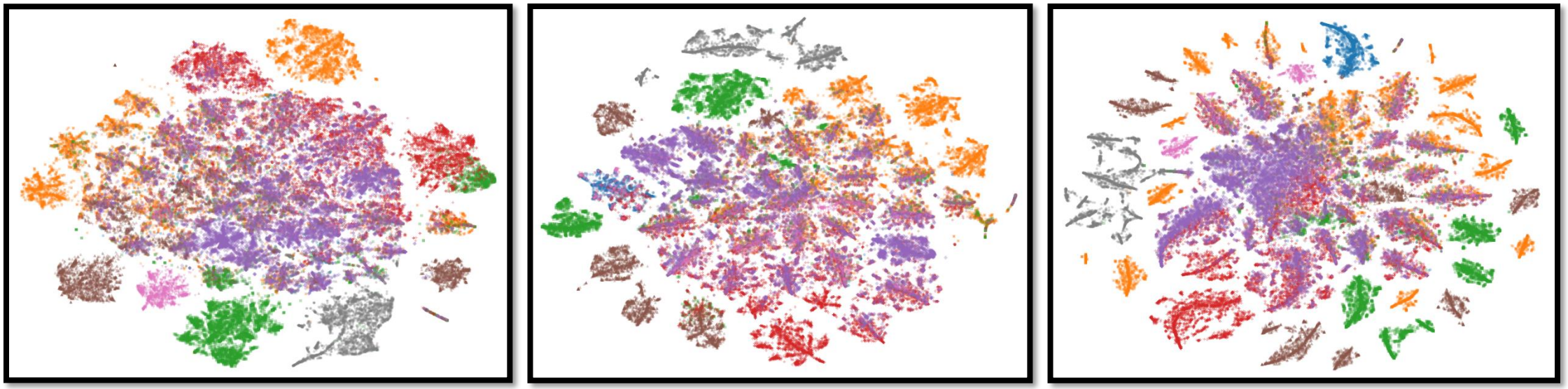
robot robots andrew human cully susan calvin brain being powell donovan law moldaug sir drake positronic bogert

…but it's not always so easy

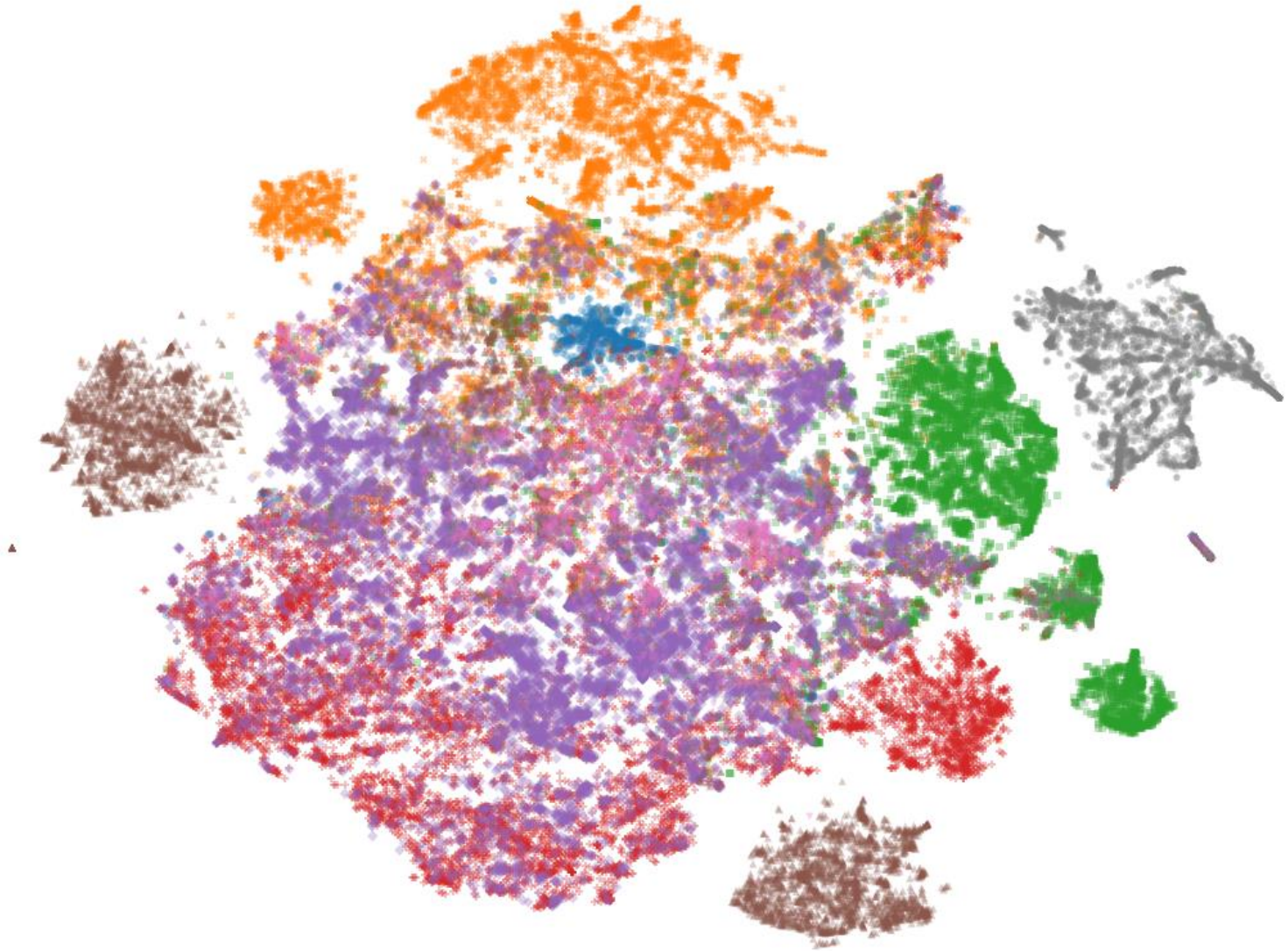
robot robots andrew human cully susan calvin brain being powell donovan law moldaug sir drake positronic bogert

sand pirx mars desert roger dust rock bass dunes crater martian jeffries kirov dune sweeney eileen rocks canyon lava camp

Adding more topics doesn't help!



250 500 1000



Preprocessing

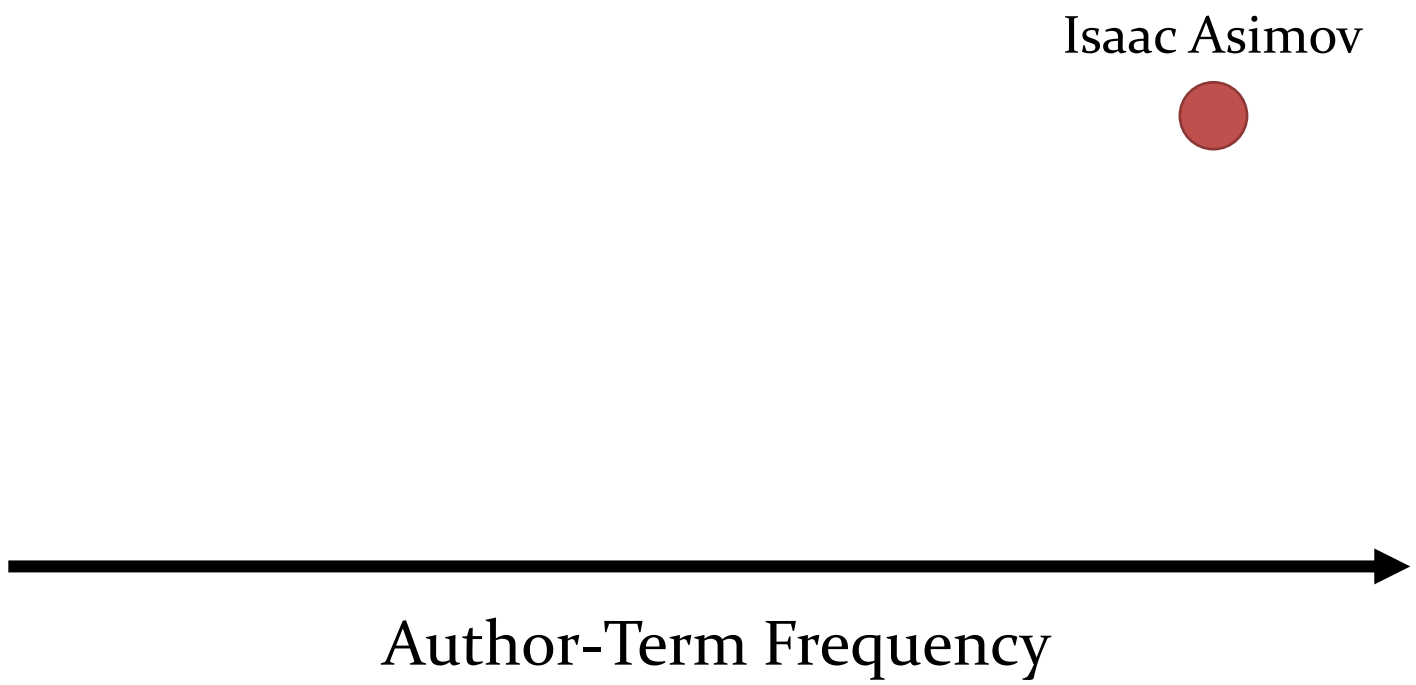
The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

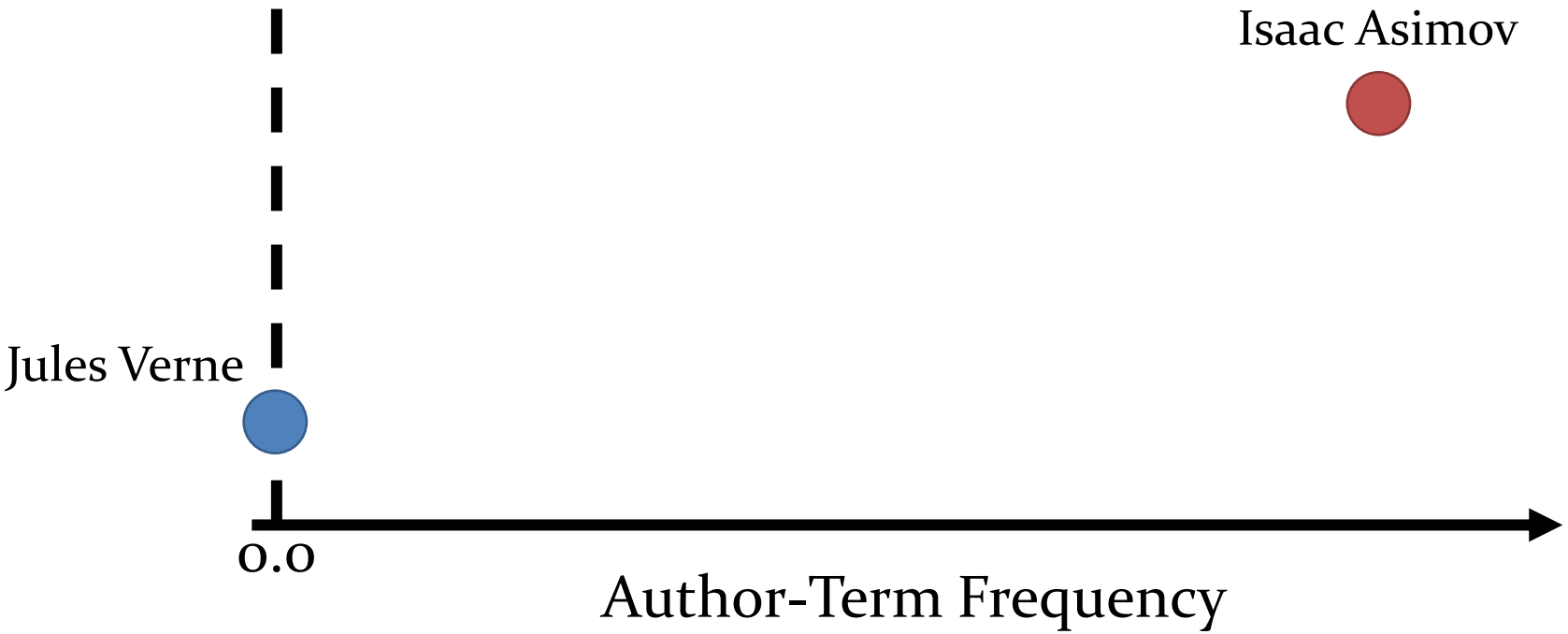
~~Preprocessing~~
Purposeful Data Modification

Author-Word Correlation: “robot”

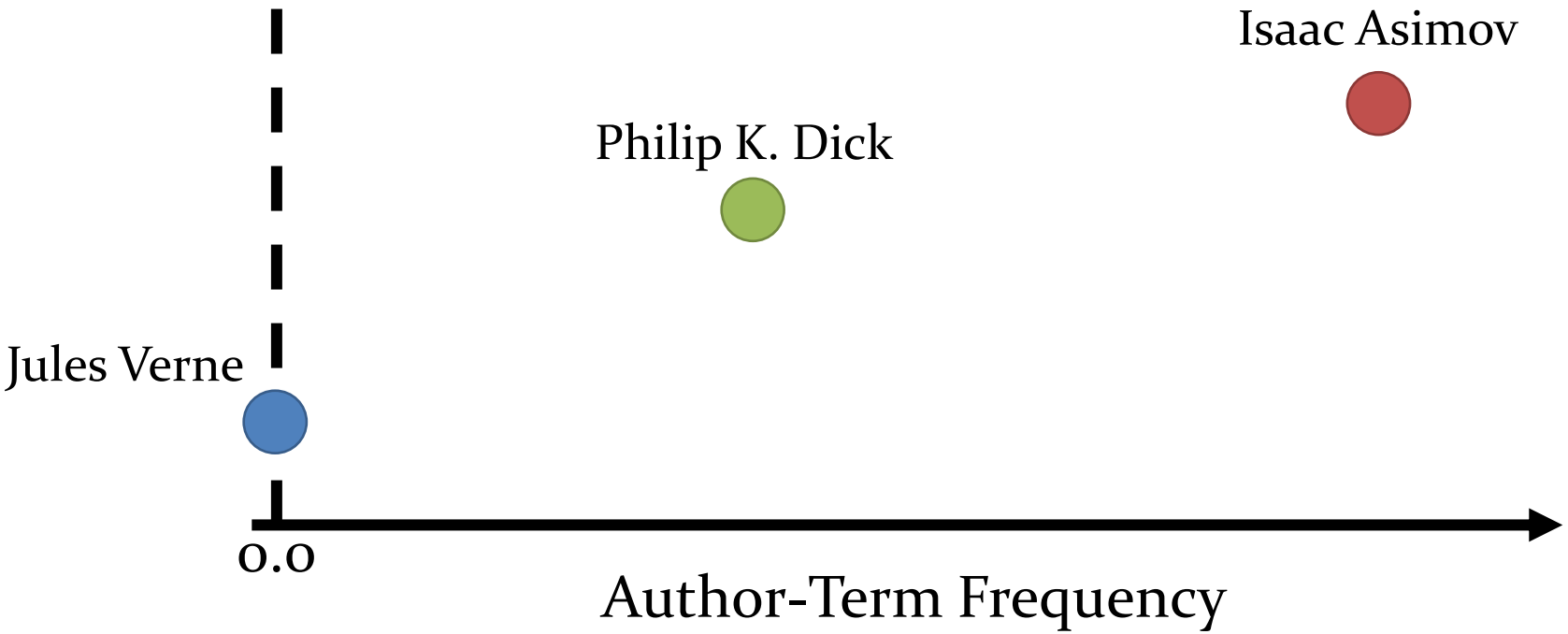
Author-Word Correlation: “robot”



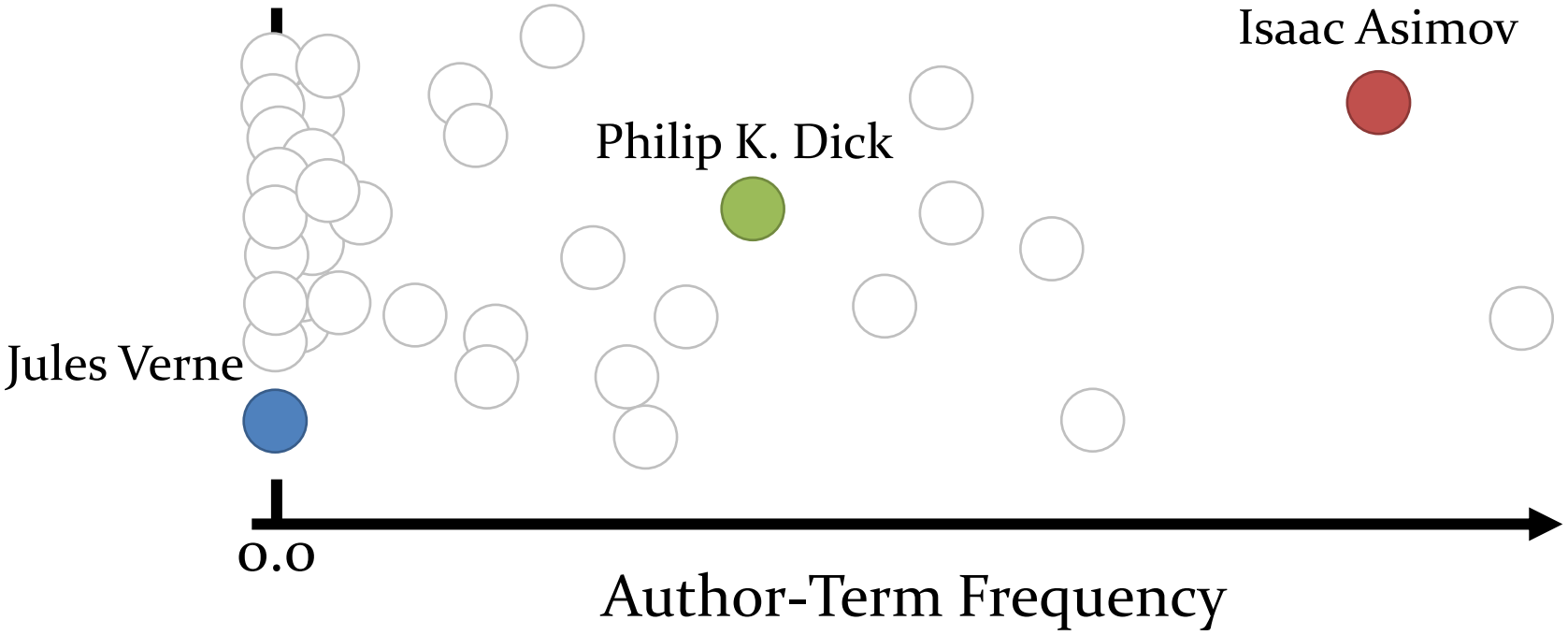
Author-Word Correlation: “robot”



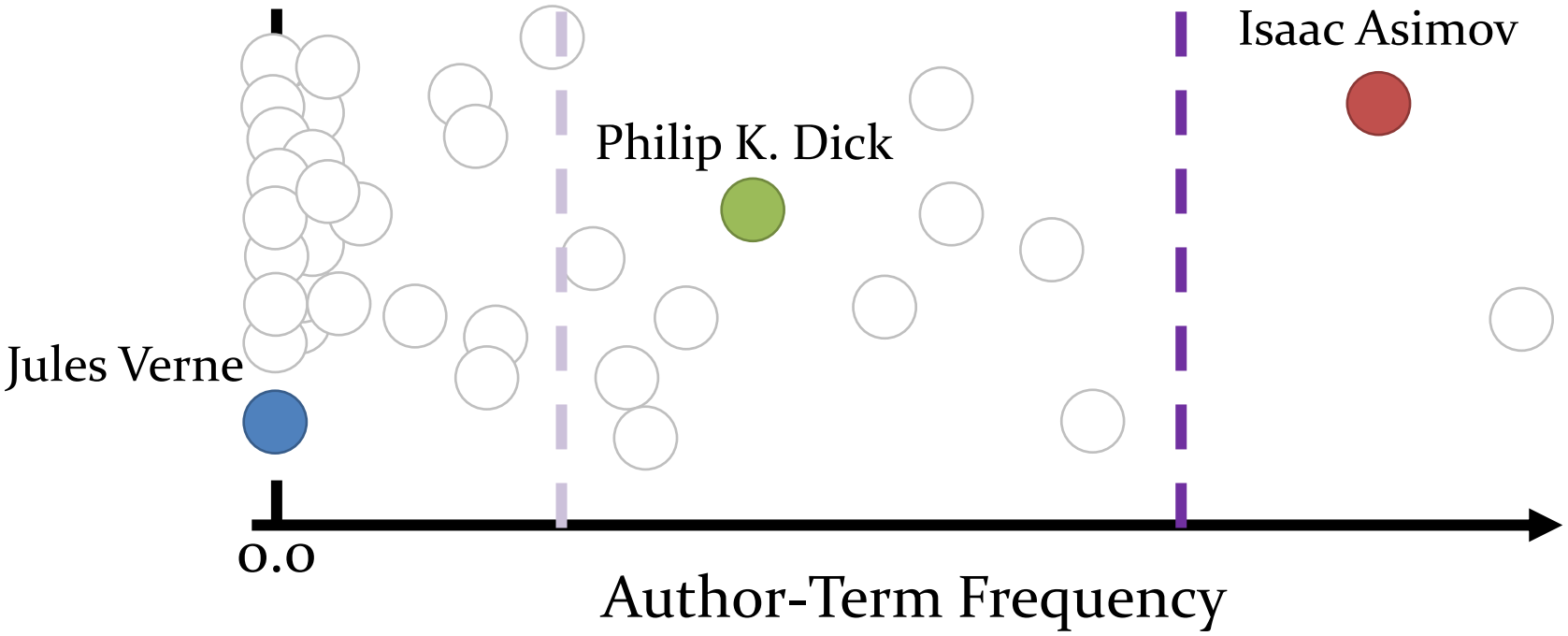
Author-Word Correlation: “robot”

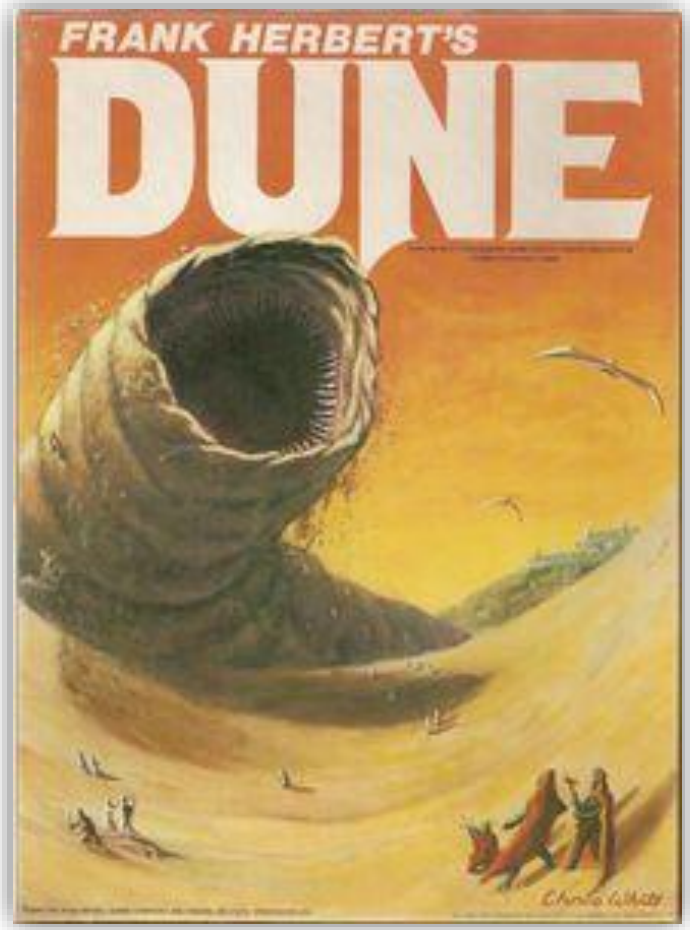


Author-Word Correlation: “robot”



Author-Word Correlation: “robot”

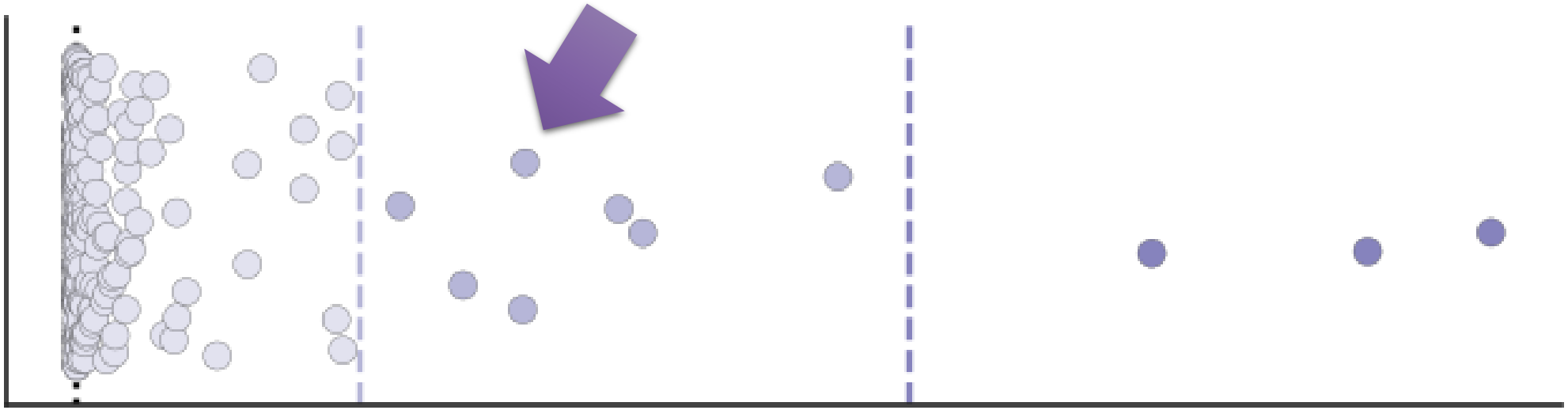



FRANK HERBERT'S
DUNE

Frank Herbert: “atreides”

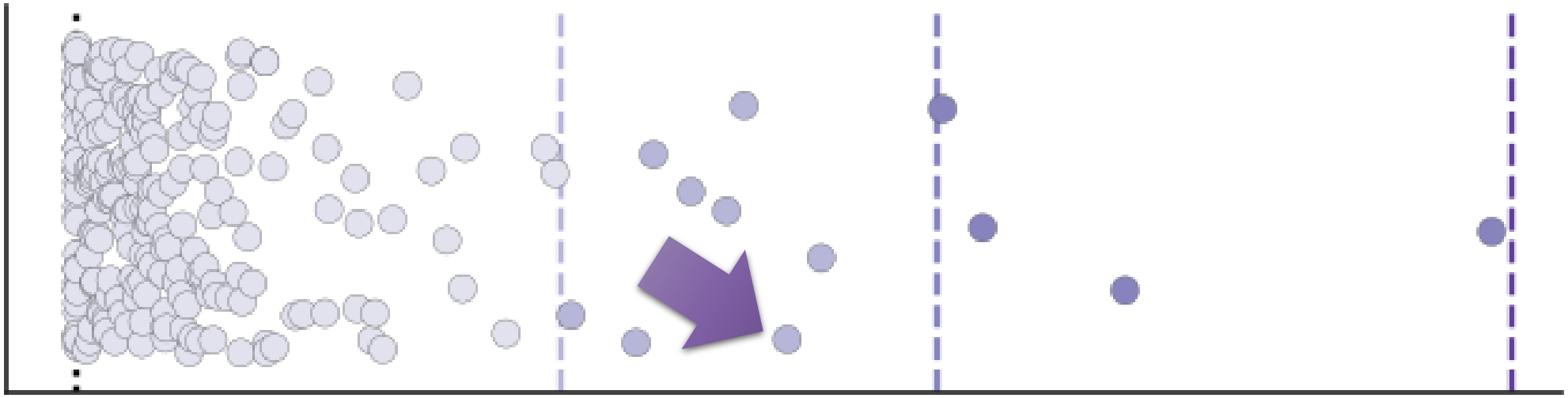
Author-Term Frequency

Frank Herbert: “paul”



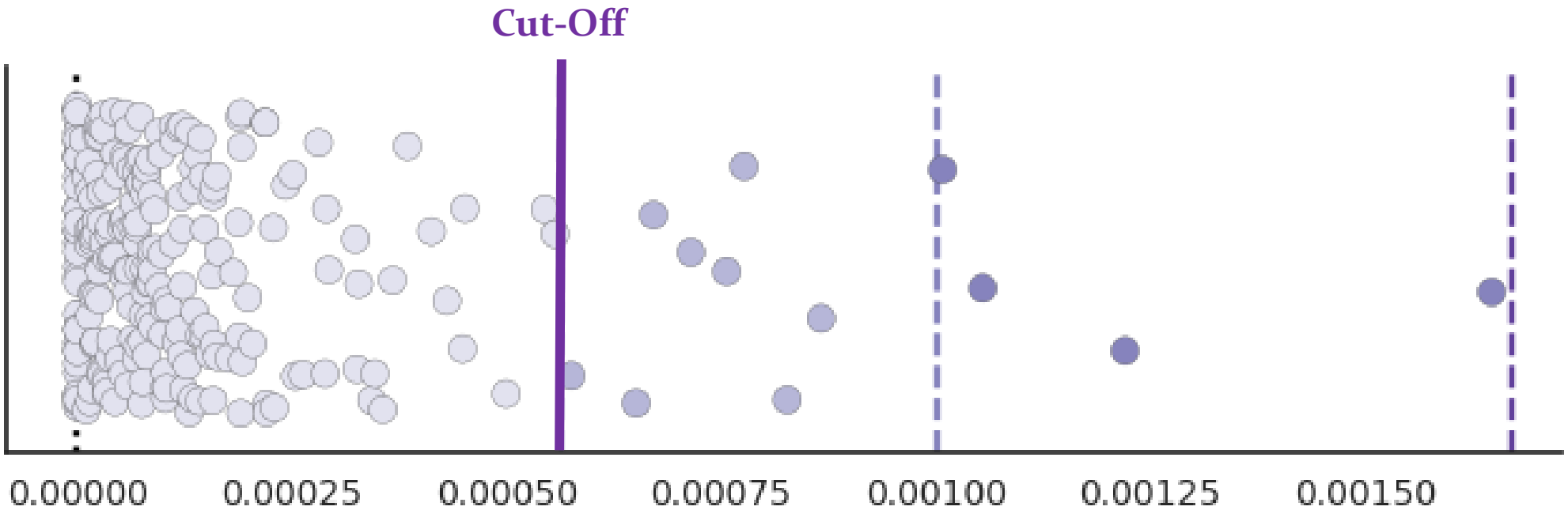
Author-Term Frequency

Frank Herbert: “desert”

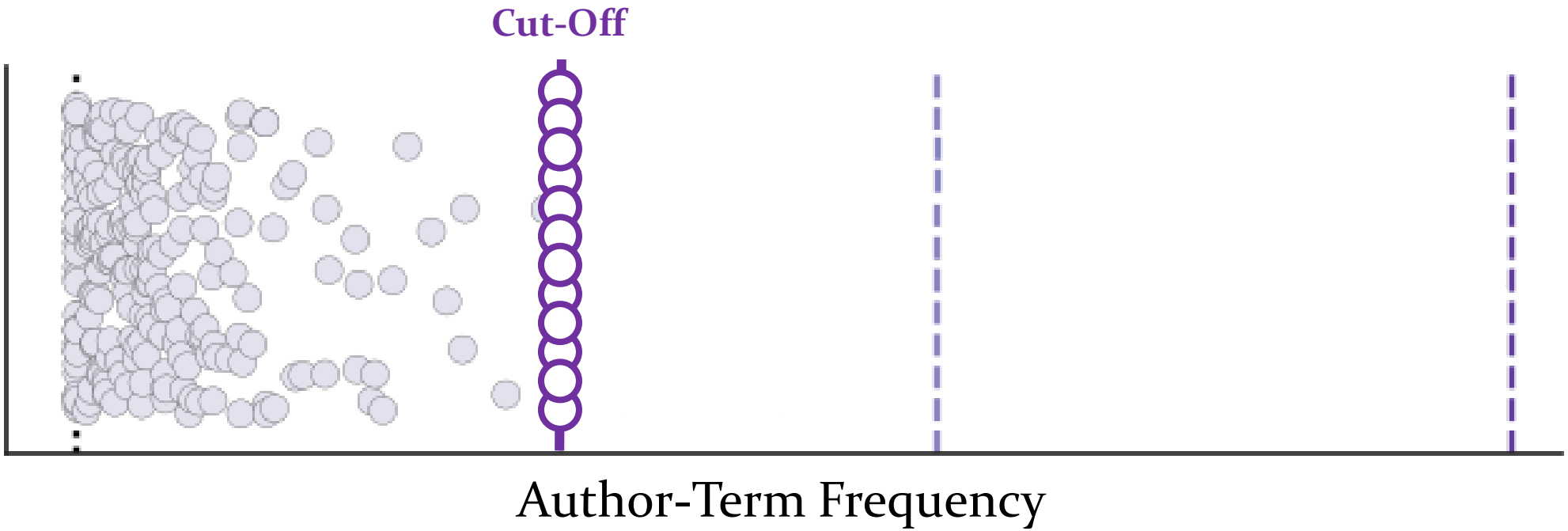


Author-Term Frequency

Author-Specific Subsampling



Author-Specific Subsampling

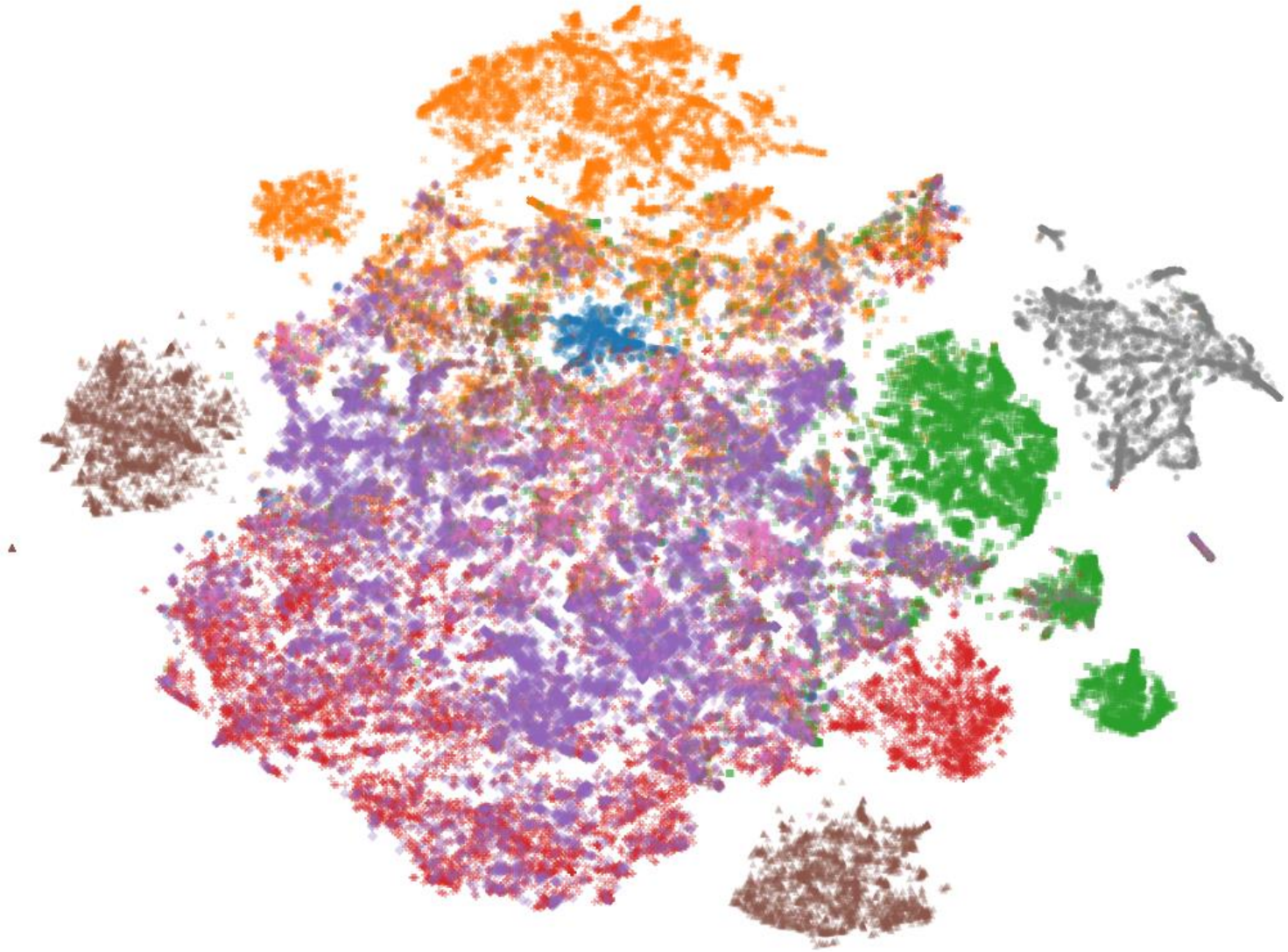


The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

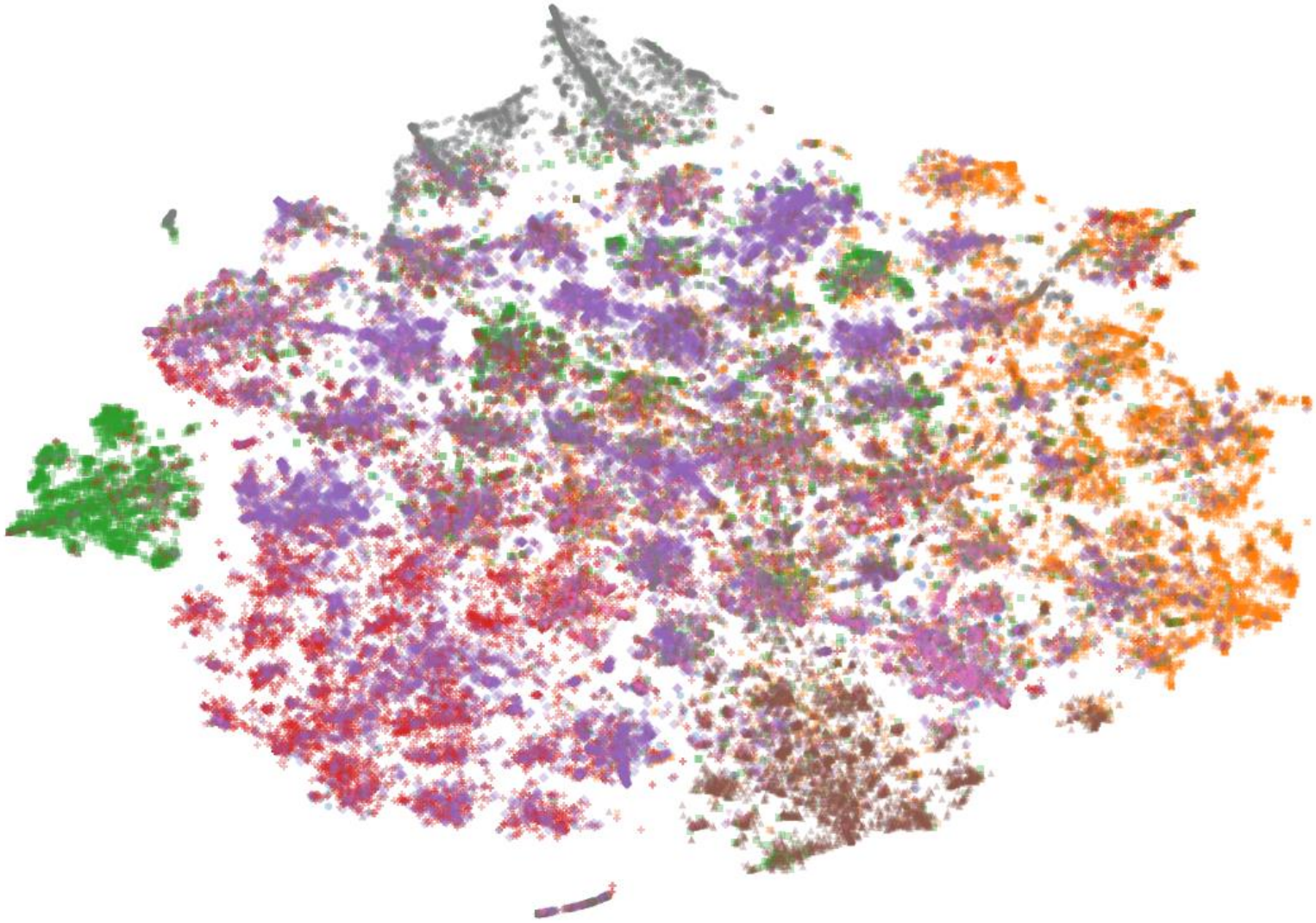
The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

The island of Gont, a single mountain that lifts its peak a mile above the storm-racked Northeast Sea, is a land famous for wizards. From the towns in its high valleys and the ports on its dark narrow bays many a Gontishman has gone forth to serve the Lords of the Archipelago in their cities as wizard or mage, or, looking for adventure, to wander working magic from isle to isle of all Earthsea.

Before

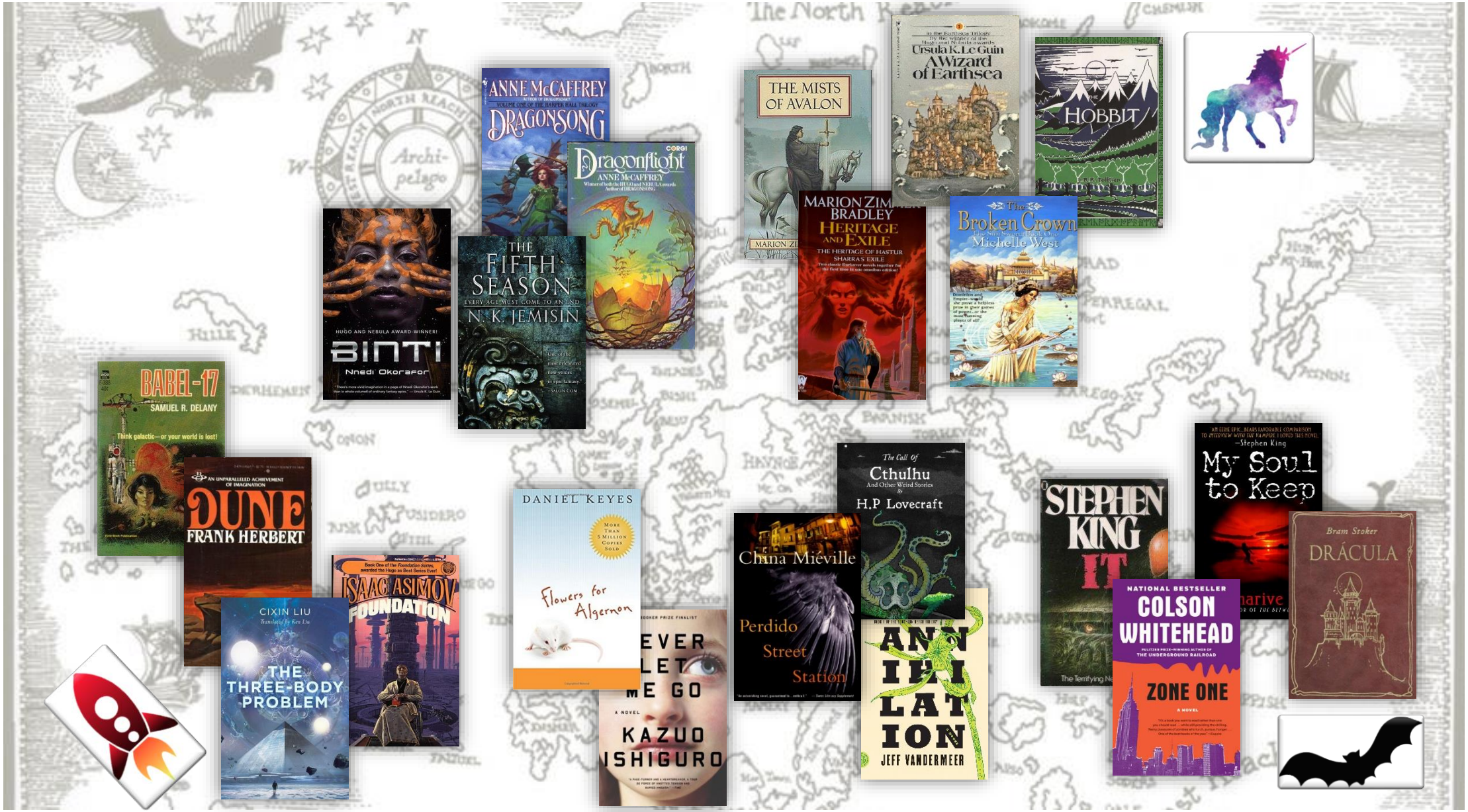


After



Topics look more cohesive & meaningful!

school professor work university years research science students	professor university college student students research school science
robot robots andrew human cully susan calvin brain being powell	machine robot machines robots human mechanical metal brain
sand pirx mars desert roger dust rock bass dunes crater martian	sand desert rock mountains mountain dust land surface plain
f'lar lessa weyr robinton hold dragon f'nor lord dragons benden rider	lord hold between master queen star enough turns high good


ANNE McCAFFREY
DRAGONSONG
Dragonflight
ANNE McCAFFREY
THE MISTS OF AVALON
Ursula K. Le Guin
A Wizard of Earthsea
THE HOBBIT
BINTI
Nnedi Okorafor
THE FIFTH SEASON
N. K. JEMISIN
MARION ZIMMER BRADLEY
HERITAGE AND EXILE
THE HERITAGE OF HASTUR
SHARRA'S EXILE
The Broken Crown
Michelle West
BABEL-17
SAMUEL R. DELANY
DUNE
FRANK HERBERT
ISAAC ASIMOV
FOUNDATION
CIXIN LIU
THE THREE-BODY PROBLEM
DANIEL KEYES
Flowers for Algernon
KAZUO ISHIGURO
China Miéville
Perdido Street Station
The Call Of Cthulhu
And Other Weird Stories
H.P Lovecraft
ANNIHILATION
JEFF VANDERMEER
STEPHEN KING
IT
My Soul to Keep
Stephen King
NATIONAL BESTSELLER
COLSON WHITEHEAD
ZONE ONE
Bram Stoker
DRÁCULA