Fracking

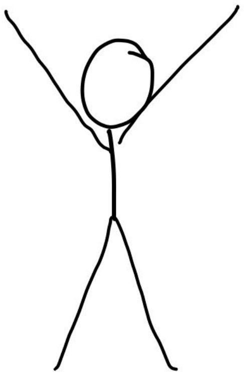
AGU100 ADVANCING EARTH AND SPACE SCIENCE

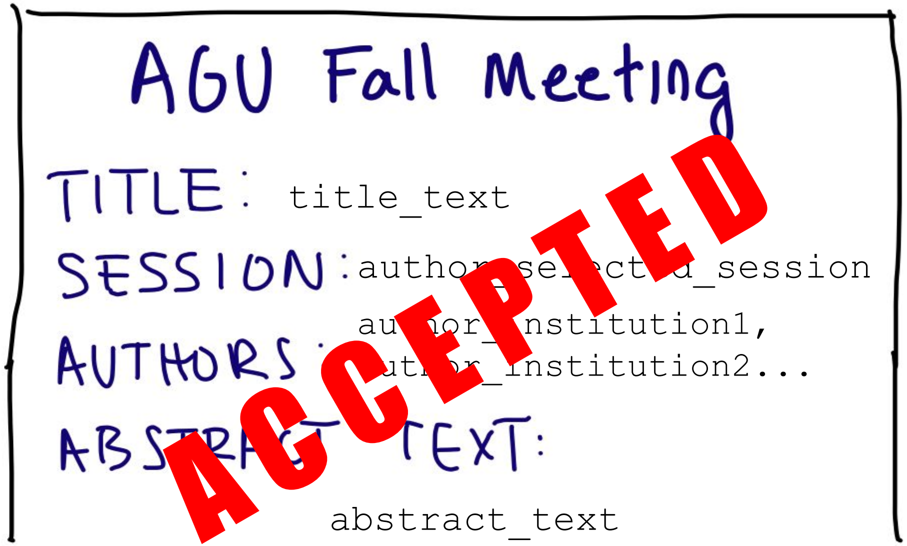
FALL MEETING

Washington, D.C. | 10–14 Dec 2018

Jordan Landers | Thinkful Final Capstone | July 11, 2019

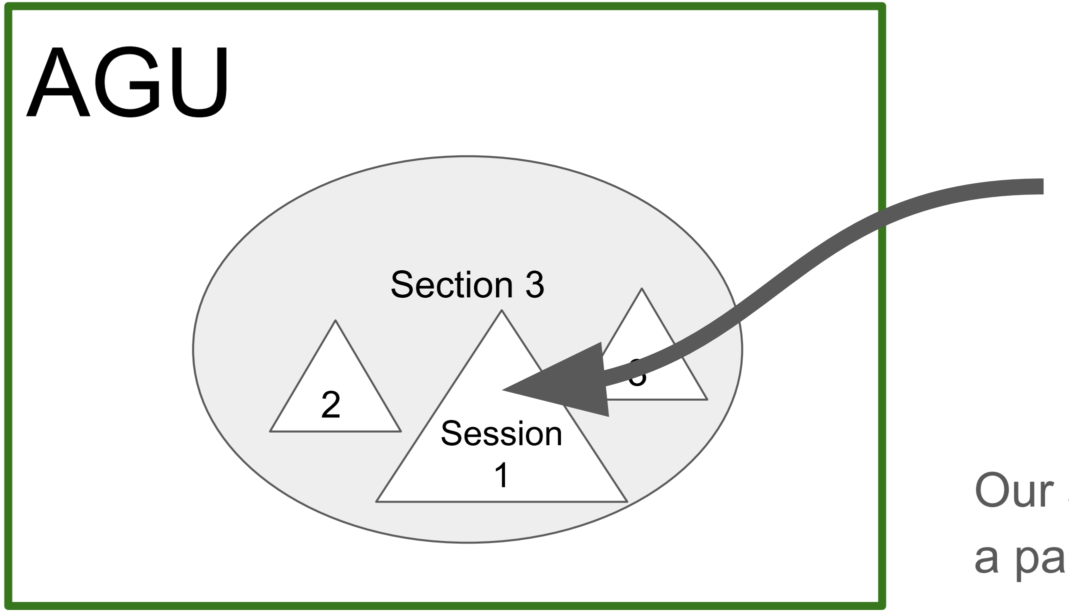
Mentor: Jason Grafft

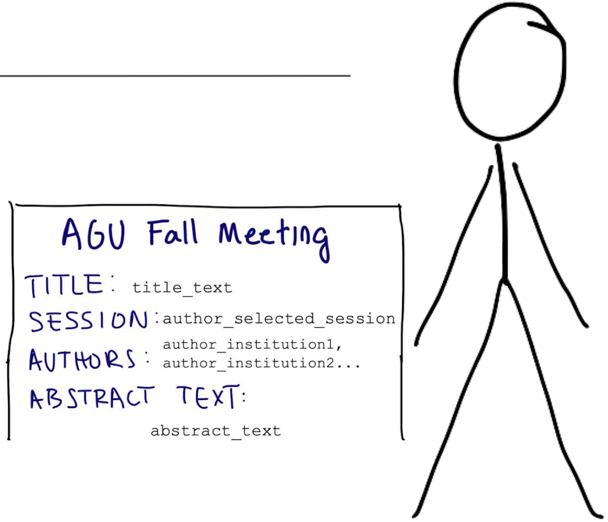




There once was an earth scientist with an accepted abstract...

Organizing AGU

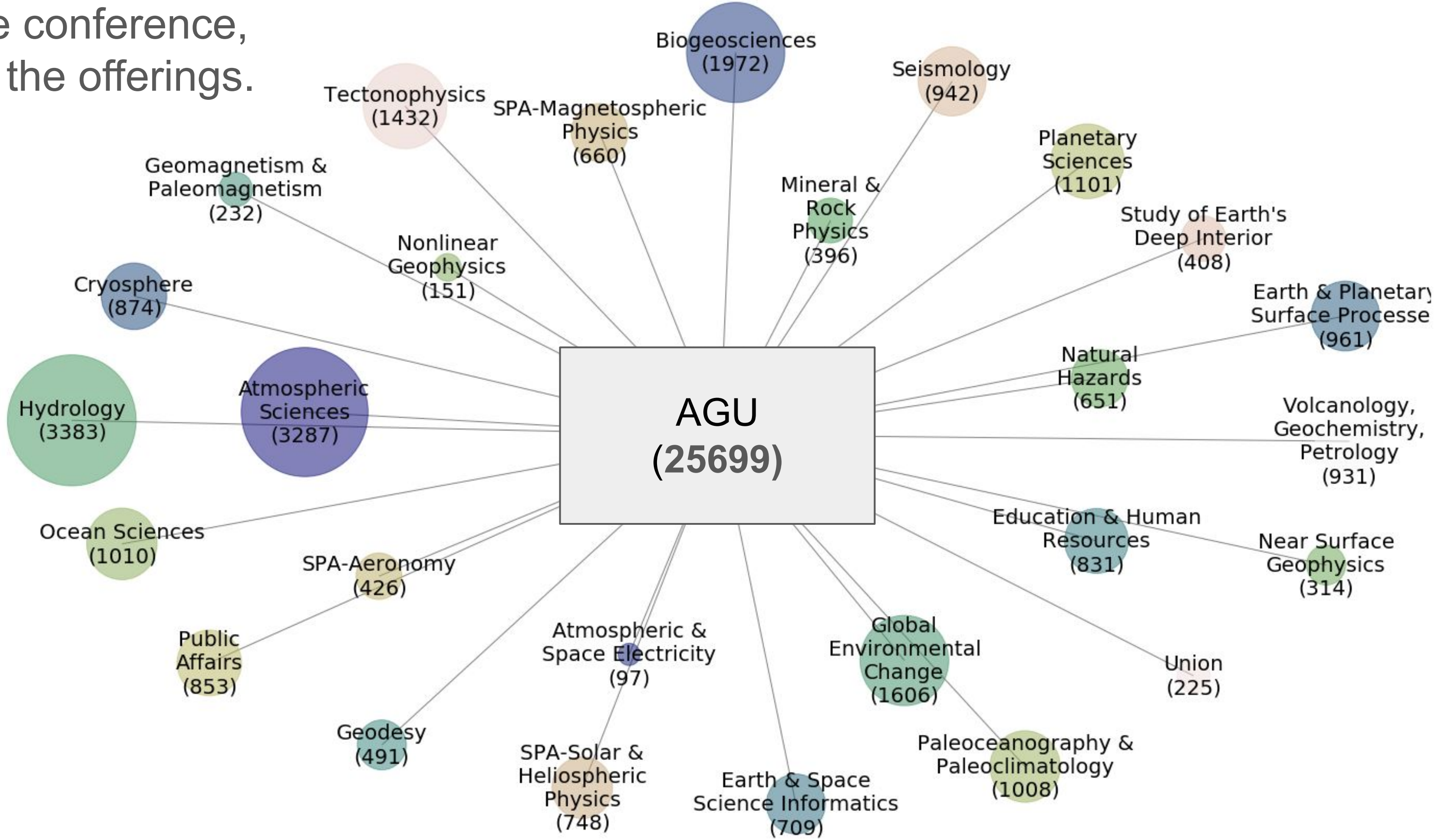




Our scientist submits an abstract to a particular *session*.

25699 abstracts distributed across **1993** sessions that belong to **27** sections

This is the conference,
these are the offerings.

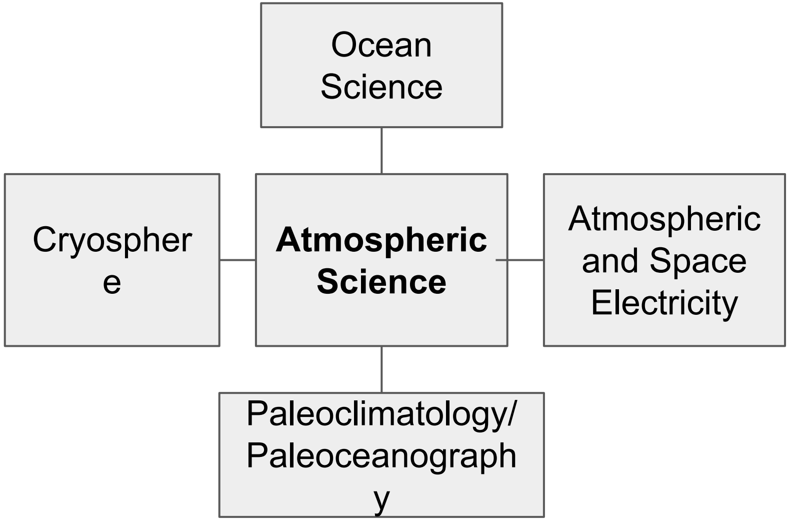


As a matter of practicality, our scientist will likely limit itinerary planning to...

(1) the section their abstract is in

Atmospheric Science
Sub-Km Grid Spacing Simulations Over the Vale do Cobro Using the WRF-ARW During Perdigao 2017
Stratification Effects on Flow through a Microscale Gap
A Comparison of Atmospheric Profilers and Environmental Soundings in Complex Terrain during the 2017 VORTEX-SE Field Campaign
Characteristics of Boundary-Layer Convection in a Deep and Wide Valley: A Large-Eddy Simulation Study
Weather Patterns Associated with the US-Bangla BS211 Aircraft Accident at TIA, Kathmandu Valley, Nepal as Revealed by WRF-ARW Simulation
Fast-response, high-resolution wind modeling over complex terrain
Effects of Topography on Residence Time and Export Fraction of Gases Emitted within Forests
On the Parameterization of Turbulence in Katabatic Flow
etc...

(2) the sections instinctively similar to the section their abstract is in



AGU Fall Meeting

TITLE: Stratification Effects on Flow through a Microscale Gap

SESSION: Boundary Layer Processes and Turbulence III

AUTHORS: ['Vassallo, D*, University of Notre Dame, Notre Dame, IN, United States'], ['Krishnamurthy, R, University of Notre Dame, Notre Dame, IN, United States'],...]

ABSTRACT TEXT: While the flow effects of mesoscale gaps and passes in mountains are well documented, studies into flow response to microscale topographic anomalies under various stability conditions are few and far between. Small gaps between localized peaks on a ridge may play an important role in the immediate surrounding environment by causing flow distortion and jetting, thus changing potential loads on wind turbines, affecting the dispersion of pollutants, and modifying the spread of forest fires.

The Perdigão Campaign, which occurred in the Spring of 2017, aimed to study flow in/over a parallel double ridge configuration, with a focus on microscale flow. One of the ridges had a densely instrumented gap that was approximately 700 m in length and 60 m in depth, allowing for an analysis of microscale gap flows. A novel triple Doppler lidar system was used to obtain data both within the gap region and on the leeward slope, while a dual Doppler lidar system collected flow information on the windward slope. Additionally, well instrumented

They will review abstracts based on:

title text and author list

OR

title text, author list AND abstract text

Section labels make the conference program less daunting, but do they help an attendee accurately build an itinerary consistent with their interests?

Do they reflect the underlying structure of the content presented at the AGU Fall Annual Meeting?

Study Design

1. Acquire data and process into `title_features` and `abstract_features`
2. Are all ways of picking abstracts to investigate created equal?

   Here we simulate the scientist's similar-abstract-search strategies using doc2vec representations of the abstracts and compare them to the average similarity across the set of abstracts returned from an analysis of the whole program.

3. All else equal, are section labels the best way to cluster abstracts?

   Here we algorithmically cluster abstracts and compare them to section label clustering

4. With training, can a classifier learn to predict section labels?

   Here we compare the performance of classifiers trained on cluster labels and section labels

Collecting and Cleaning Data

- Abstract text and metadata from the 2018 Fall Meeting was scraped using `scrapy` and `selenium`
- Raw HTML data were stripped of tags, stop words, numbers, and punctuation, set to lower case and lemmatized using `nltk`

```
"title": ["<h2>Unexpected and
significant biospheric CO<sub>2</sub>
fluxes in the Los Angeles Basin
indicated by atmospheric radiocarbon
(<sup>14</sup>CO<sub>2</sub>)</h2>"]
```

→

```
"title": unexpected
significant biospheric co2
fluxes los angeles basin
indicated by atmospheric
radiocarbon 14co2
```

- Author names and institutions were formatted to create unique identifiers

Miller, J B*, Global Monitoring Division, NOAA/ESRL, Boulder, CO, United States'

→

millerjbglobalmonitoringdivisionnoaa esrlbouldercounitedstates

Concept of Doc2Vec

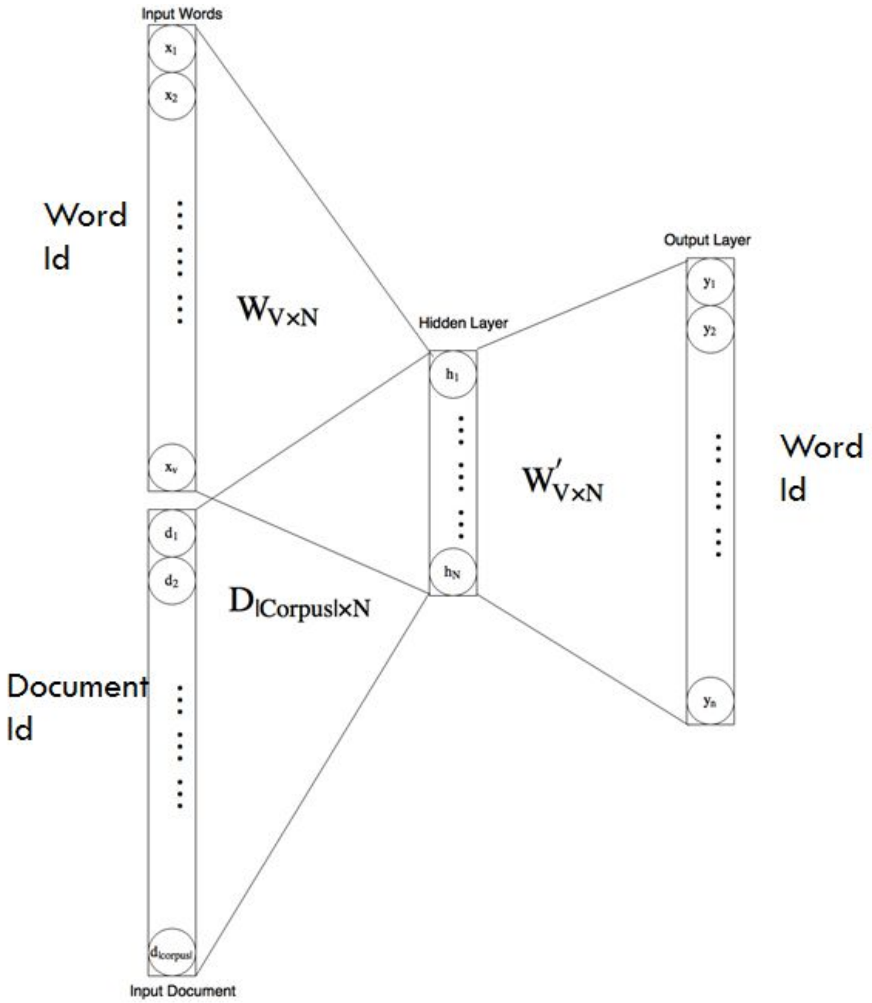
“The bug in her code caused her computer to crash.”

“The little kid spent the afternoon in the backyard digging in the dirt and putting bugs in her jar.”

1. Word2Vec uses these examples to build word context for “bug”
2. Doc2Vec adds information about each sentence

Ex: “The bug in her _____ caused her computer to crash.”

The model might predict “code” rather than “dirt”



Building Doc2Vec models

Train Doc2Vec models for `abstract_features` and `title_features` on each of 22 parameter sets.

Doc2Vec parameters combinations were created from:

- min_alpha = [.0001, .0003], min_count = [8, 12, 16], vector_size = [15, 35], window = [3, 6]

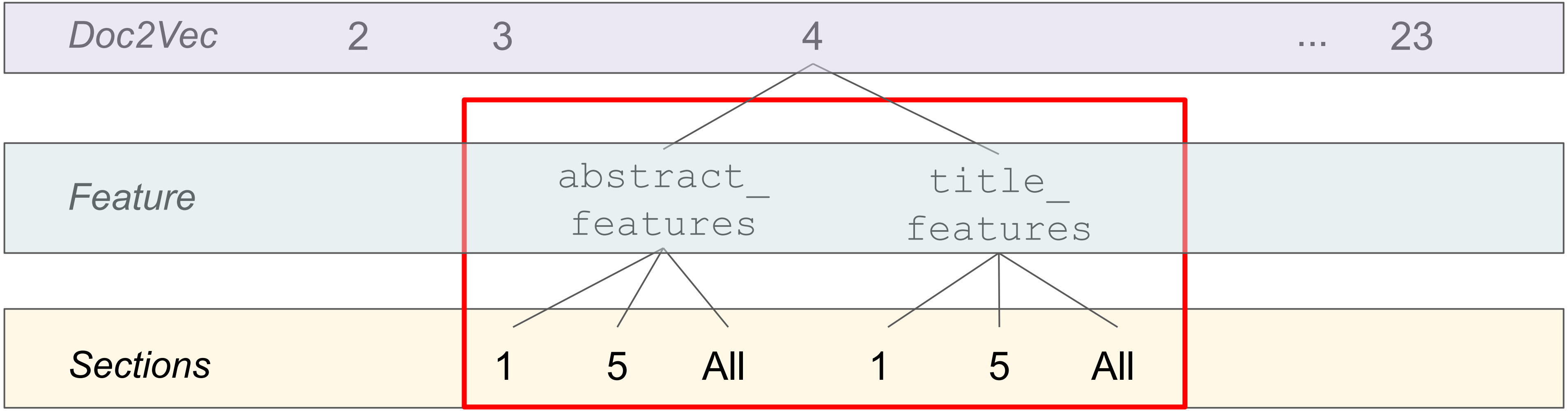
Doc2Vec	2	3	4		...	23
Feature			`abstract_features`	`title_features`		

Experiment 1:

Doc2Vec Similarity Analysis

Experiment 1: Overview

Does “*filter by my section and maybe a few similar sections*” yield the most similar set of abstracts to my seed?



Experiment 1: “Closest 4 sections”

The data structure to keep in mind:

```
{ Atmospheric Sciences: {  Ocean Sciences: [   ],
                           Natural Hazards: [   ], …
                           (all sections)  }

        (all sections) }
```

For every **section**:

For every abstract:

1. calculate the 50 most similar abstracts by doc2vec.most_similar()
2. Identify the section for each add the ***corresponding score*** to the corresponding section list

Average the lists for each section

Pick the four top scoring sections to be the “closest 4”

Experiment 1: “Closest 4 sections” (example)

The data structure to keep in mind:

{ Atmospheric Sciences: { Ocean Sciences: [   ],
Natural Hazards: [   ], …
(all sections) },...
section_label: {} (for all sections)}

Atm Sci, #5867	Corr. section	Ocean Sci.	Nat. Haz.	Paleo	Hydro	Biogeo
	Sim. Score	.9	.87	.8	.7	.6
Atm Sci, #5834	**Corr. section**	**Nat. Haz.**	**Hydro**	**Biogeo**	**Paleo**	**Ocean Sci.**
	Sim. Score	.8	.63	.6	.55	.5

{ Atm Sci.:
{ Ocean Sci.: [.9, .5],
Nat. Haz.: [.87, .8],
Paleo: [.8, .55],
Biogeo.: [.6, .6],
Hydro.: [.7, .63] …
(all sections) }

μ
.7
.835
.675
.6
.665

Closest 4 Sections for Atmospheric Sciences: [Ocean Sci, Nat. Haz., Paleo, Hydro]

Experiment 1: Calculations

[Example: Doc2Vec parameter set 5, title_features, abstract 6085]

```
df_agu.iloc[agu_inds][['title2', 'section', 'session']][:60]
```

	title2	section	session
1666	Have we seen the largest earthquakes in easter...	Geodesy	Plate Motion, Continental Deformation, and Int...
5584	Energetic onset of earthquakes	Seismology	Earthquake Source Physics Inferred from Macros...
5634	Aftershock density decay in space and time: Ob...	Seismology	Induced Seismicity in the United States and Ca...
12822	Slip Distribution of the 1960 Chile Earthquake...	Tectonophysics	Seismotectonic Processes Along Active Latin Am...
18756	Sifting Fact from Science Fiction for the Publ...	Public Affairs	The Hazards of Hazard Communication: Importanc...
25653	Hey Alexa, Open USGS Did You Feel It? Explori...	Public Affairs	Leveraging Social Media, Crowdsourcing, Citize...
18604	Shear Wave Splitting Tomography at Kilauea	Volcanology, Geochemistry, Petrology	The 2018 Eruptions of Klauea Volcano, Hawaii, ...
5430	Rapid Characterization of Large Earthquakes wi...	Seismology	Extracting Information from Geophysical and Ge...
5892	The Weak Determinism of Large Earthquakes	Seismology	Earthquake Source Physics: Unified Perspective...
13149	Quantifying seismic hazard from interseismic l...	Tectonophysics	Shallow Subduction Zone Structure and Dynamics II
6266	Earthquake Similarity through Graphical Modeli...	Seismology	Recent Progress in Nuclear Test Monitoring Cap...
14392	Uncovering the physical controls of episodic t...	Tectonophysics	Whose Fault Is It? Relating Structural and Com...
14479	Normal fault connectivity through time: an exa...	Tectonophysics	Three-Dimensional Fault Architecture and Geome...
9111	Hydroacoustic records from non-tsunamigenic ev...	Natural Hazards	Integrated Approach for Earth, Ocean, Atmosphe...
5283	The Chicxulub Impact Produced a Powerful Globa...	Paleoceanography and Paleoclimatology	The KPg Mass Extinction and the Chicxulub Impa...
5391	Increasing complexity of earthquake cycle with...	Seismology	Seismology Contributions: Earthquakes II Posters
25617	Bridging the Gap between Earthquake Hazards Re...	Public Affairs	Science to Action: Education for Community/Sci...
14898	Vent location forecasts at calderas: a physics...	Tectonophysics	Numerical and Laboratory Analogue Models of Dy...
5872	The USGS National Earthquake Information Cente...	Seismology	New Frontiers in Global Seismic Monitoring and...

Model n	
Limitation: Seed section only	
Seed	Mean of similarity between seed and...
1	5584, 5634, 5430, 5892...
… 25699	...
Model avg.	

Experiment 1: Calculations

[Example: Doc2Vec parameter set 5, title_features, abstract 6085]

```
df_agu.iloc[agu_inds][['title2', 'section', 'session']][:60]
```

	title2	section	session
1666	Have we seen the largest earthquakes in easter...	Geodesy	Plate Motion, Continental Deformation, and Int...
5584	Energetic onset of earthquakes	Seismology	Earthquake Source Physics Inferred from Macros...
5634	Aftershock density decay in space and time: Ob...	Seismology	Induced Seismicity in the United States and Ca...
12822	Slip Distribution of the 1960 Chile Earthquake...	Tectonophysics	Seismotectonic Processes Along Active Latin Am...
18756	Sifting Fact from Science Fiction for the Publ...	Public Affairs	The Hazards of Hazard Communication: Importanc...
25653	Hey Alexa, Open USGS Did You Feel It? Explori...	Public Affairs	Leveraging Social Media, Crowdsourcing, Citize...
18604	Shear Wave Splitting Tomography at Kilauea	Volcanology, Geochemistry, Petrology	The 2018 Eruptions of Klauea Volcano, Hawaii, ...
5430	Rapid Characterization of Large Earthquakes wi...	Seismology	Extracting Information from Geophysical and Ge...
5892	The Weak Determinism of Large Earthquakes	Seismology	Earthquake Source Physics: Unified Perspective...
13149	Quantifying seismic hazard from interseismic l...	Tectonophysics	Shallow Subduction Zone Structure and Dynamics II
6266	Earthquake Similarity through Graphical Modeli...	Seismology	Recent Progress in Nuclear Test Monitoring Cap...
14392	Uncovering the physical controls of episodic t...	Tectonophysics	Whose Fault Is It? Relating Structural and Com...
14479	Normal fault connectivity through time: an exa...	Tectonophysics	Three-Dimensional Fault Architecture and Geome...
9111	Hydroacoustic records from non-tsunamigenic ev...	Natural Hazards	Integrated Approach for Earth, Ocean, Atmosphe...
5283	The Chicxulub Impact Produced a Powerful Globa...	Paleoceanography and Paleoclimatology	The KPg Mass Extinction and the Chicxulub Impa...
5391	Increasing complexity of earthquake cycle with...	Seismology	Seismology Contributions: Earthquakes II Posters
25617	Bridging the Gap between Earthquake Hazards Re...	Public Affairs	Science to Action: Education for Community/Sci...
14898	Vent location forecasts at calderas: a physics...	Tectonophysics	Numerical and Laboratory Analogue Models of Dy...
5872	The USGS National Earthquake Information Cente...	Seismology	New Frontiers in Global Seismic Monitoring and...

Model n	
Limitation: Seed section + 4 similar	
Seed	Mean of similarity between seed and...
1	5584, 5634, 12822, ...
... 25699	...
Model avg.	

Experiment 1: Calculations

[Example: Doc2Vec parameter set 5, title_features, abstract 6085]

```
df_agu.iloc[agu_inds][['title2', 'section', 'session']][:60]
```

	title2	section	session
1666	Have we seen the largest earthquakes in easter...	Geodesy	Plate Motion, Continental Deformation, and Int...
5584	Energetic onset of earthquakes	Seismology	Earthquake Source Physics Inferred from Macros...
5634	Aftershock density decay in space and time: Ob...	Seismology	Induced Seismicity in the United States and Ca...
12822	Slip Distribution of the 1960 Chile Earthquake...	Tectonophysics	Seismotectonic Processes Along Active Latin Am...
18756	Sifting Fact from Science Fiction for the Publ...	Public Affairs	The Hazards of Hazard Communication: Importanc...
25653	Hey Alexa, Open USGS Did You Feel It? Explori...	Public Affairs	Leveraging Social Media, Crowdsourcing, Citize...
18604	Shear Wave Splitting Tomography at Kilauea	Volcanology, Geochemistry, Petrology	The 2018 Eruptions of Klauea Volcano, Hawaii, ...
5430	Rapid Characterization of Large Earthquakes wi...	Seismology	Extracting Information from Geophysical and Ge...
5892	The Weak Determinism of Large Earthquakes	Seismology	Earthquake Source Physics: Unified Perspective...
13149	Quantifying seismic hazard from interseismic l...	Tectonophysics	Shallow Subduction Zone Structure and Dynamics II
6266	Earthquake Similarity through Graphical Modeli...	Seismology	Recent Progress in Nuclear Test Monitoring Cap...
14392	Uncovering the physical controls of episodic t...	Tectonophysics	Whose Fault Is It? Relating Structural and Com...
14479	Normal fault connectivity through time: an exa...	Tectonophysics	Three-Dimensional Fault Architecture and Geome...
9111	Hydroacoustic records from non-tsunamigenic ev...	Natural Hazards	Integrated Approach for Earth, Ocean, Atmosphe...
5283	The Chicxulub Impact Produced a Powerful Globa...	Paleoceanography and Paleoclimatology	The KPg Mass Extinction and the Chicxulub Impa...
5391	Increasing complexity of earthquake cycle with...	Seismology	Seismology Contributions: Earthquakes II Posters
25617	Bridging the Gap between Earthquake Hazards Re...	Public Affairs	Science to Action: Education for Community/Sci...
14898	Vent location forecasts at calderas: a physics...	Tectonophysics	Numerical and Laboratory Analogue Models of Dy...
5872	The USGS National Earthquake Information Cente...	Seismology	New Frontiers in Global Seismic Monitoring and...

Model n	
Limitation: None (all sections)	
Seed	Mean of similarity between seed and...
1	1666, 5584, 5634, 12822, 18756...
… 25699	...
Model avg.	

Experiment 1a: How many sections to search?

1_title_feat_S1		
index	mean	std
1		
2		
etc...		
25699		

1_title_feat_S5		
index	mean	std
1		
2		
etc...		
25699		

1_title_feat_SA		
index	mean	std
1		
2		
etc...		
25699		

1_abstract_feat_S1		
index	mean	std
1		
2		
etc...		
25699		

1_abstract_feat_S5		
index	mean	std
1		
2		
etc...		
25699		

1_abstract_feat_SA		
index	mean	std
1		
2		
etc...		
25699		

Run t-test between sets, record significant set(s) w/ higher mean score(s)

	2	3	...	23		Row summary
abstract						
title						
Col summary						

T-tests for all abstract features winners

Best performing models

T-tests between param set # winners

Experiment 1a: Results

			2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		Row sum
AF																										9, 21
TF																										9, 13, 17, 5, 21
Col sum																										9, 13, 17, 5, 21

S1	S5	SAll	Title	Abstract

- "All Sections" average model scores were best for every parameterization of both `abstract_features` & `title_features`
- `title_features` were more predictive of similar titles than abstracts were of similar abstracts except for models 19 and 23, where the two were not significantly different.

Experiment 1b: Search based on title or abstract text?

1_title_feat_S1		
index	mean	std
1		
2		
etc...		
25699		

1_abstract_feat_S1		
index	mean	std
1		
2		
etc...		
25699		

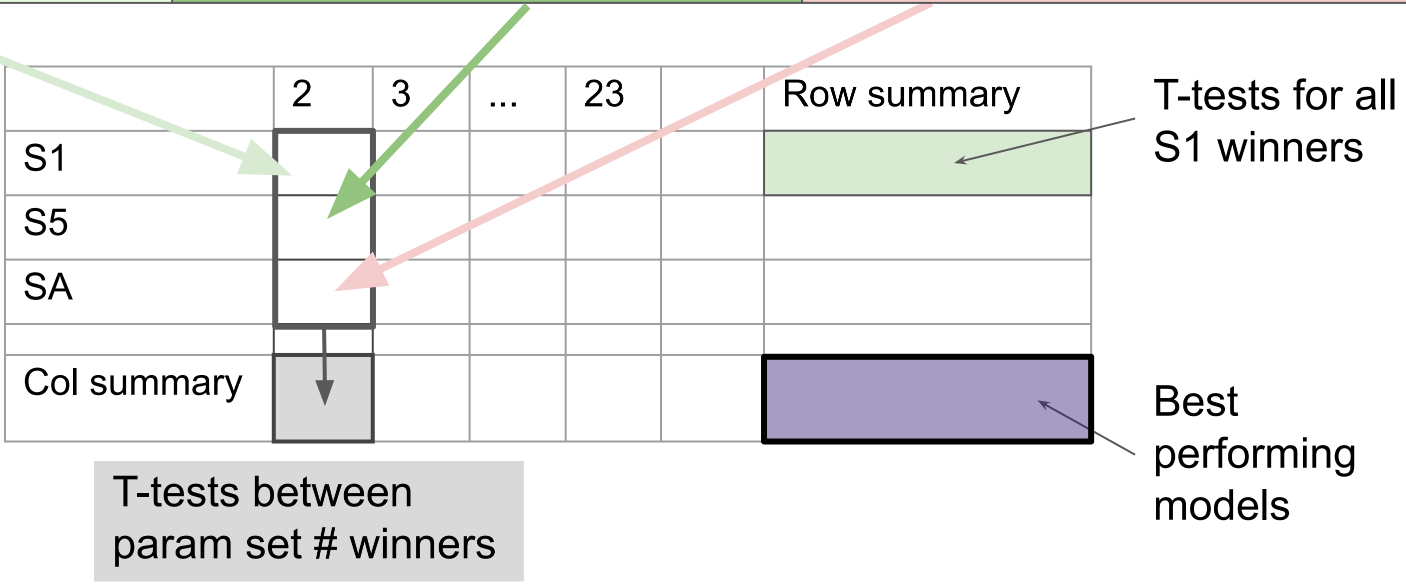
1_title_feat_S5		
index	mean	std
1		
2		
etc...		
25699		

1_abstract_feat_S5		
index	mean	std
1		
2		
etc...		
25699		

1_title_feat_SA		
index	mean	std
1		
2		
etc...		
25699		

1_abstract_feat_SA		
index	mean	std
1		
2		
etc...		
25699		

Run t-test between sets, record significant set(s) w/ higher mean score(s)



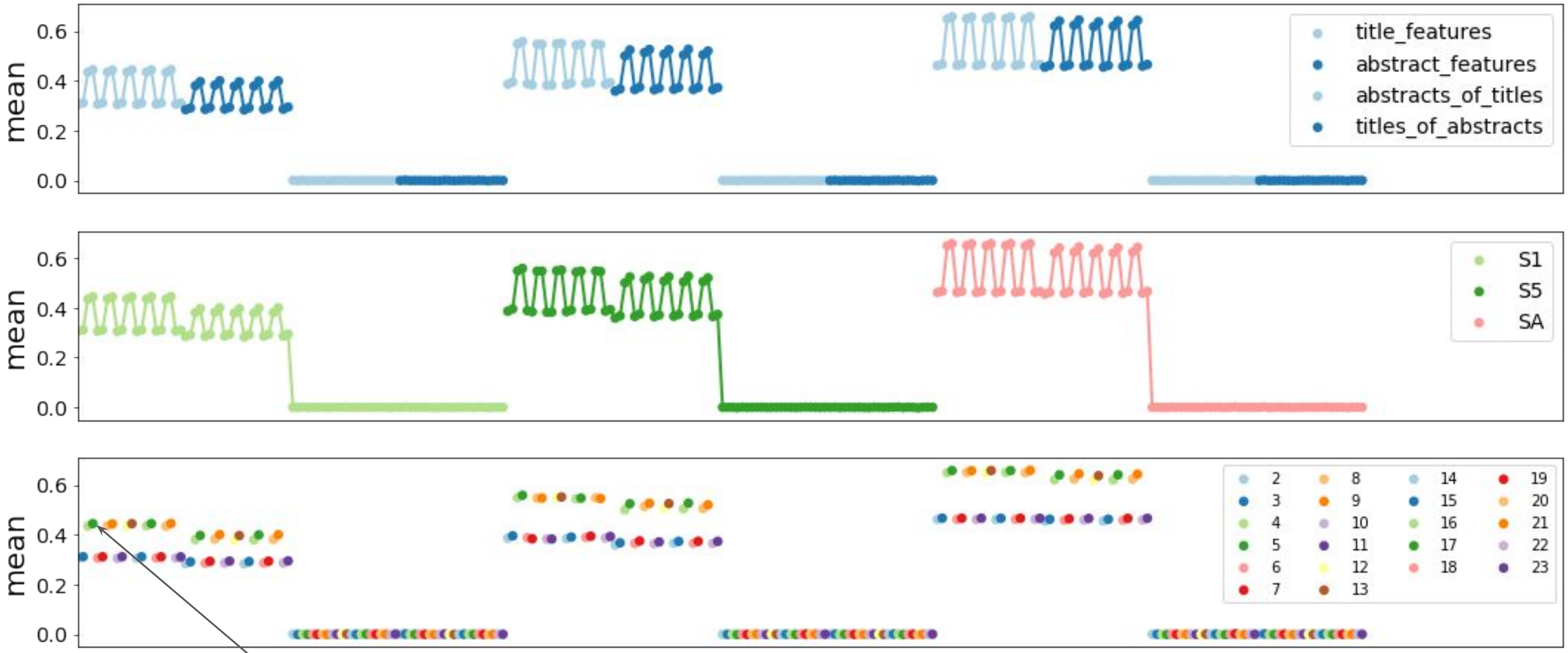
Experiment 1b: Results

			2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		Row sum
S1																										17, 21, 5, 13, 9
S5																										5
SA																										9, 13, 17, 5, 21
Col sum																										5, 17, 9, 21, 13

S1	S5	SAll	Title	Abstract

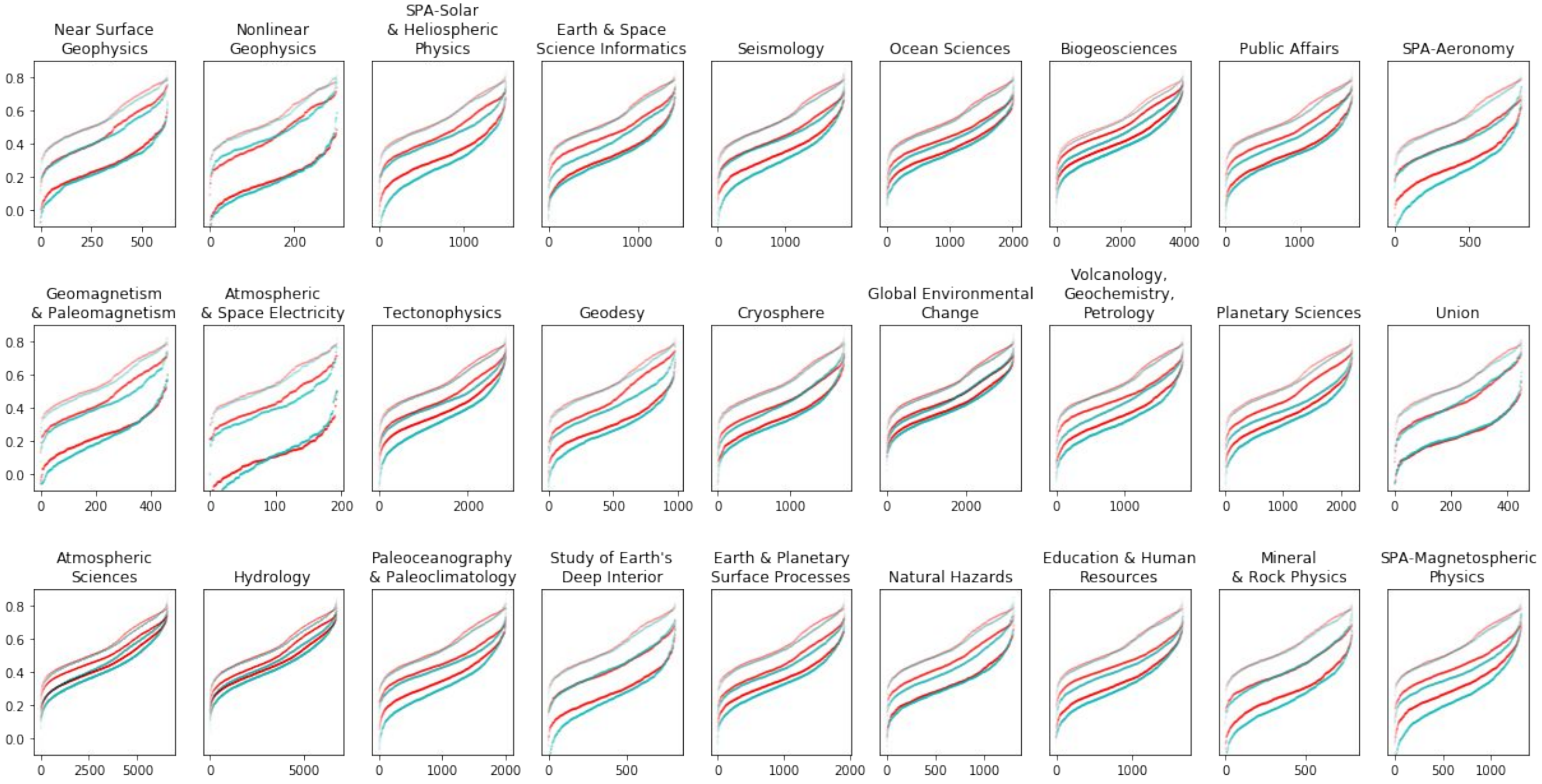
- Same column summary as experiment 1a

Experiment 1: Summary



(*5_title_feat_S1.csv*, mean
of abstract similarity scores)

Results Breakdown by Section



T	A	
		1
		5
		A

Checking the results

Seed: Fully Physics-Based PSHA: Coupling RSQSim with Deterministic Ground Motion Simulations

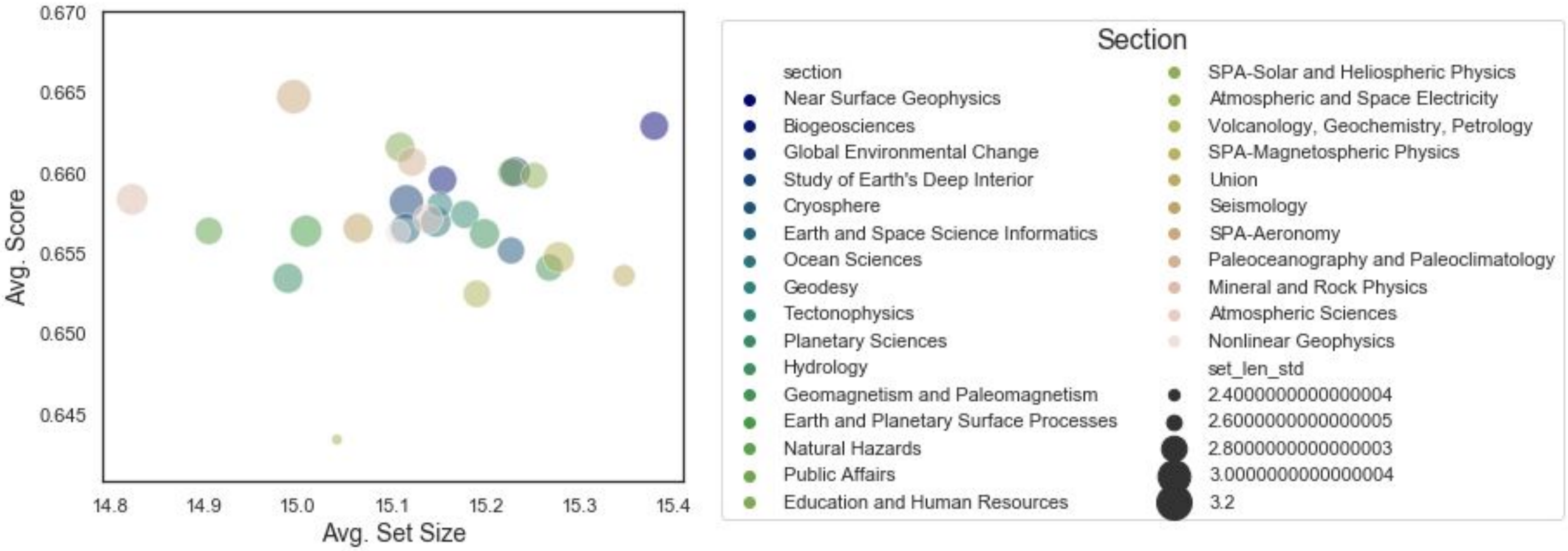
Section: Seismology, **Session**: 'Beyond the Earthquake Cycle: Field and Modeling Constraints of Earthquake Rupture Along Complex-Geometry Fault Systems and Implications for Seismic Hazard Assessment II'

```
df_agu.iloc[agu_inds][['title2', 'section', 'session']][:60]
```

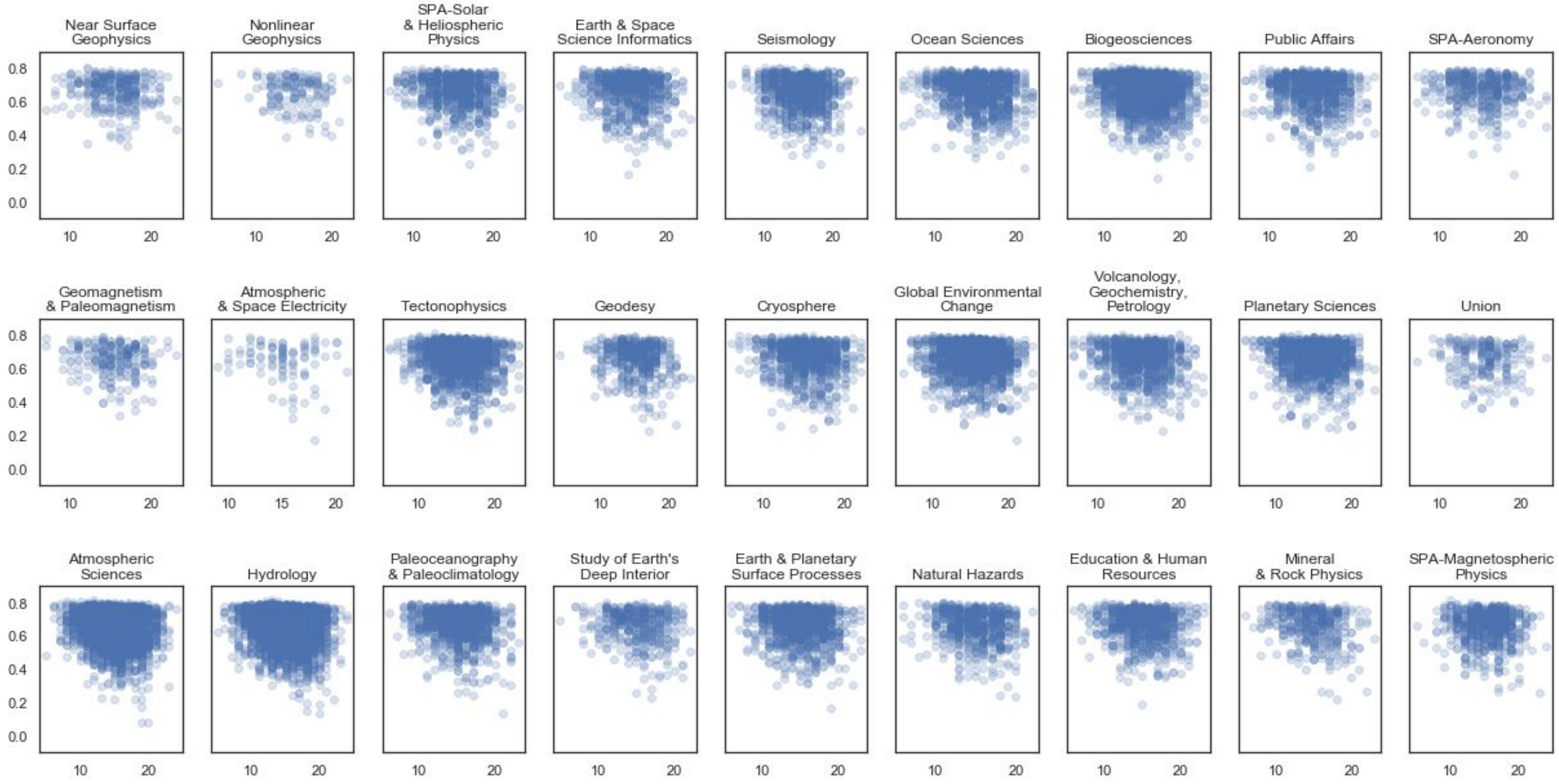
10 sections featured in top 60!

	title2	section	session
1666	Have we seen the largest earthquakes in easter...	Geodesy	Plate Motion, Continental Deformation, and Int...
5584	Energetic onset of earthquakes	Seismology	Earthquake Source Physics Inferred from Macros...
5634	Aftershock density decay in space and time: Ob...	Seismology	Induced Seismicity in the United States and Ca...
12822	Slip Distribution of the 1960 Chile Earthquake...	Tectonophysics	Seismotectonic Processes Along Active Latin Am...
18756	Sifting Fact from Science Fiction for the Publ...	Public Affairs	The Hazards of Hazard Communication: Importanc...
25653	Hey Alexa, Open USGS Did You Feel It? Explori...	Public Affairs	Leveraging Social Media, Crowdsourcing, Citize...
18604	Shear Wave Splitting Tomography at Kilauea	Volcanology, Geochemistry, Petrology	The 2018 Eruptions of Klauea Volcano, Hawaii, ...
5430	Rapid Characterization of Large Earthquakes wi...	Seismology	Extracting Information from Geophysical and Ge...
5892	The Weak Determinism of Large Earthquakes	Seismology	Earthquake Source Physics: Unified Perspective...
13149	Quantifying seismic hazard from interseismic l...	Tectonophysics	Shallow Subduction Zone Structure and Dynamics II
6266	Earthquake Similarity through Graphical Modeli...	Seismology	Recent Progress in Nuclear Test Monitoring Cap...
14392	Uncovering the physical controls of episodic t...	Tectonophysics	Whose Fault Is It? Relating Structural and Com...

Sections/Abstract for each Section in SA calcs



Sections/Abstract for each Section in SA calcs

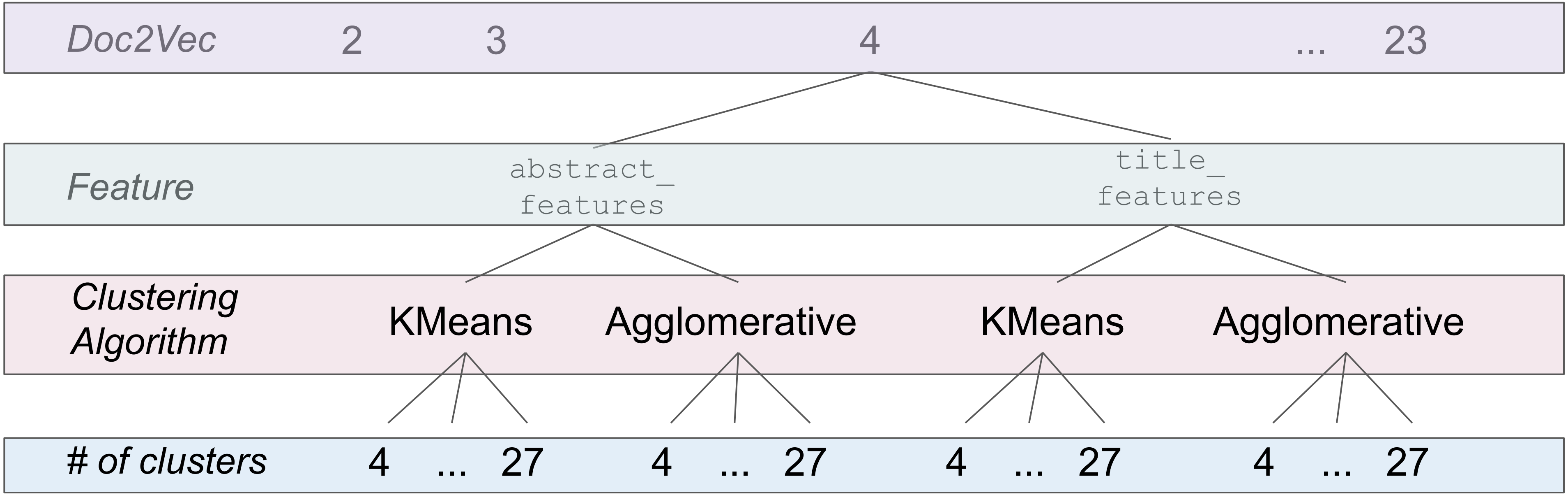


Experiment 2:

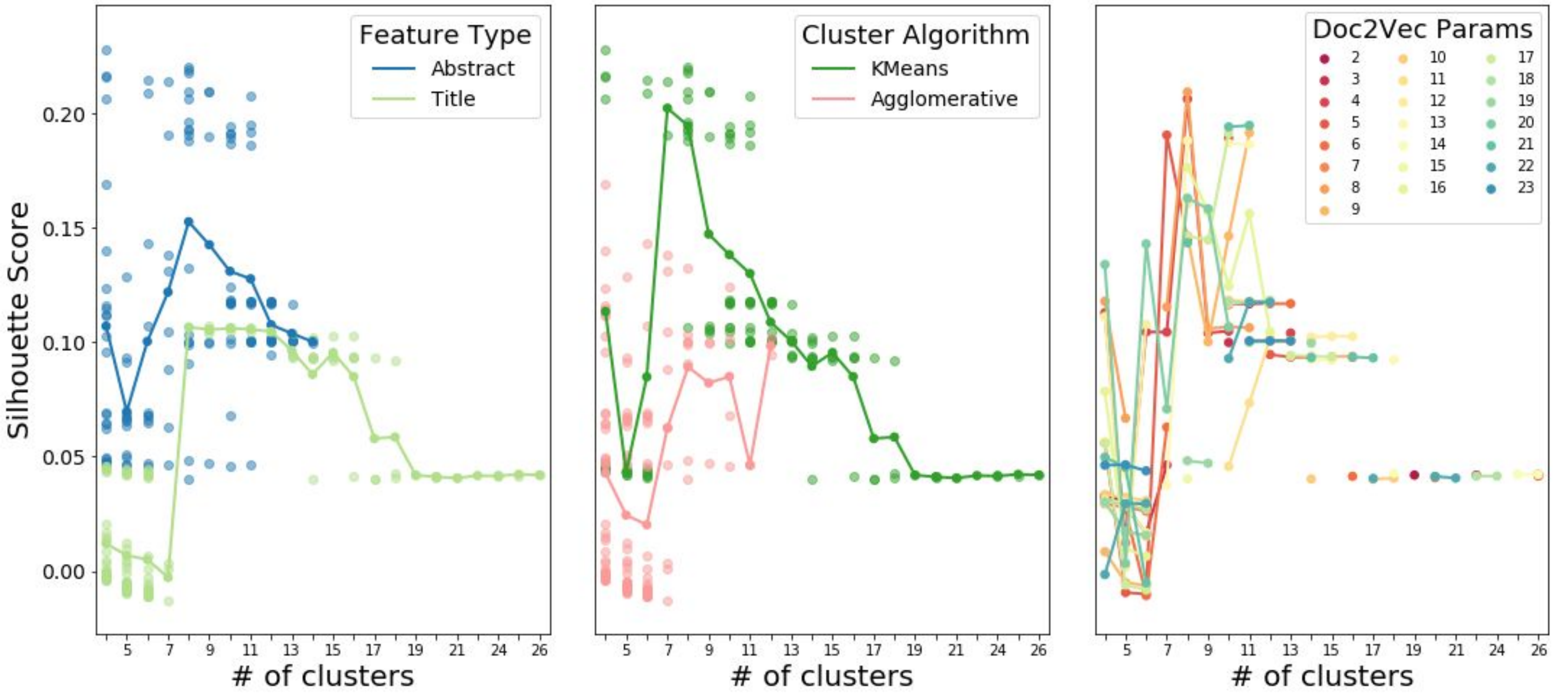
Clustering

Experiment 2: Overview

Are Section Labels the best description of the underlying group structure of the AGU offerings?

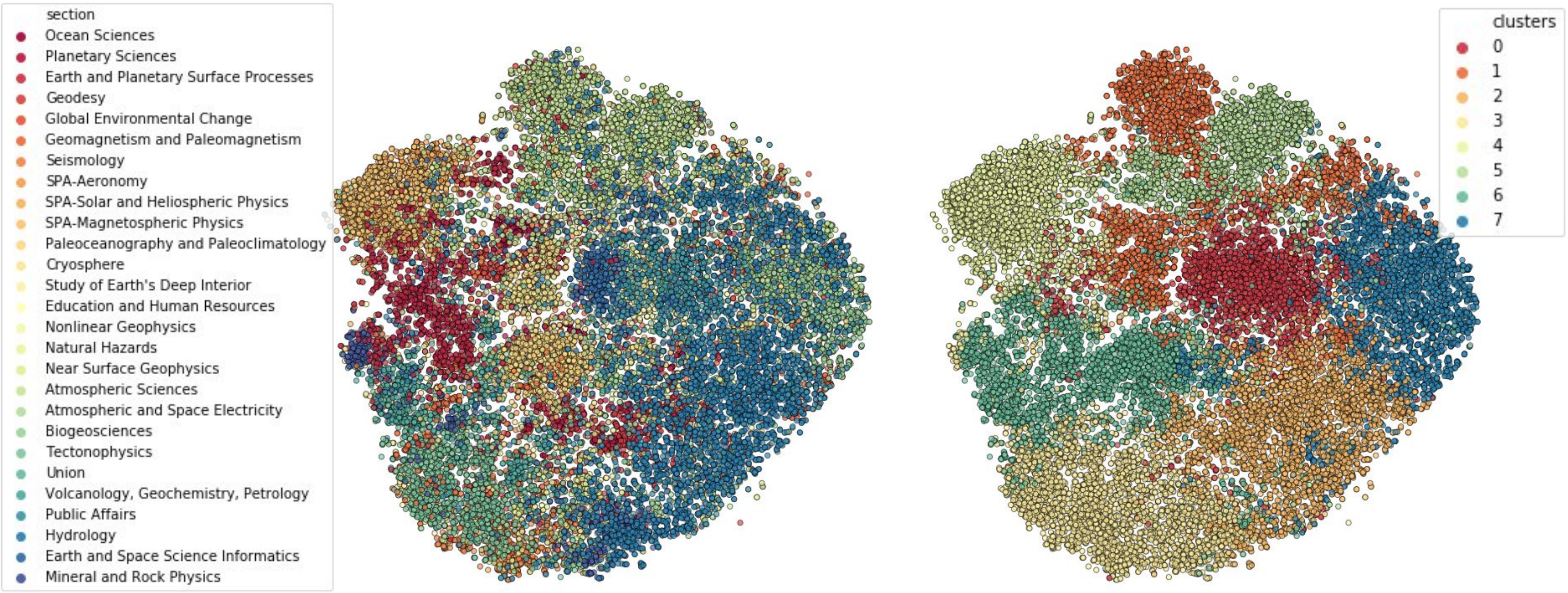


Results Summary

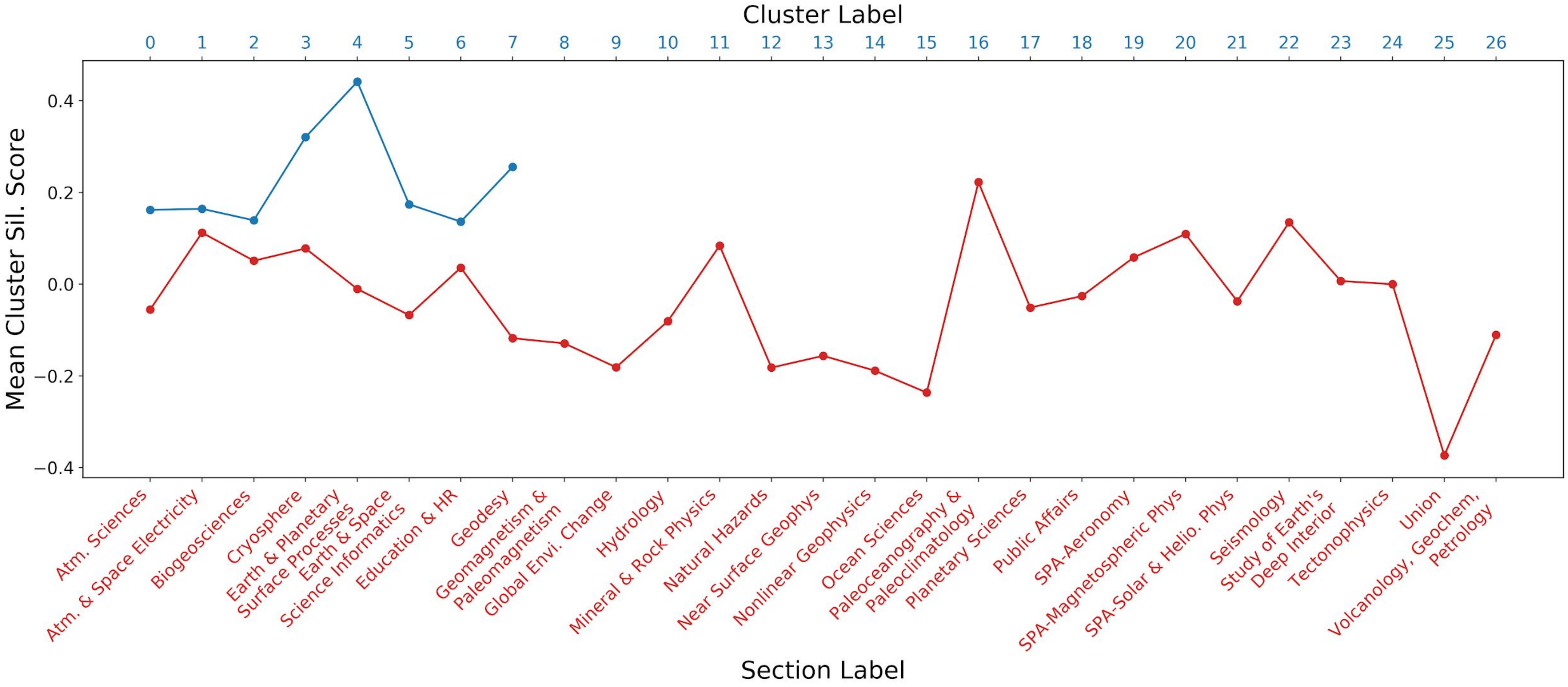


# of clusters for the three highest scoring models for each feature_type-cluster_alg-doc2vec configurations

Results Summary



Results Summary

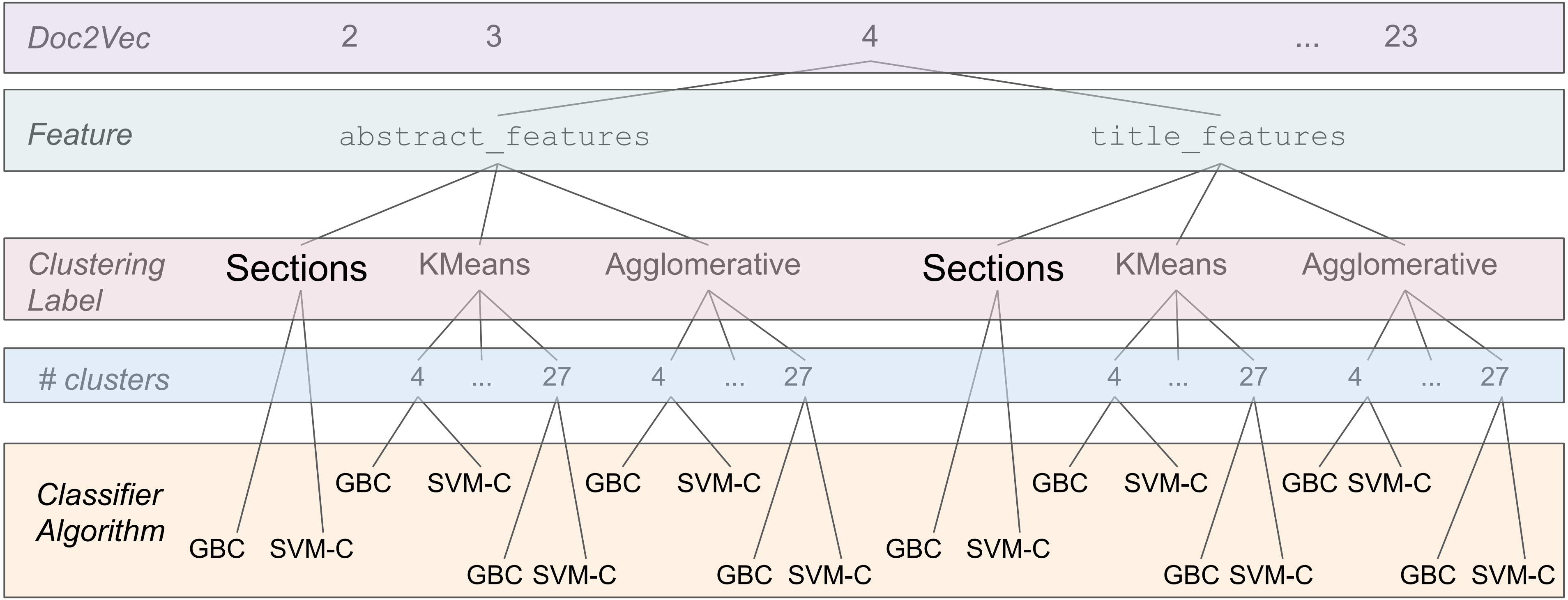


Experiment 3:

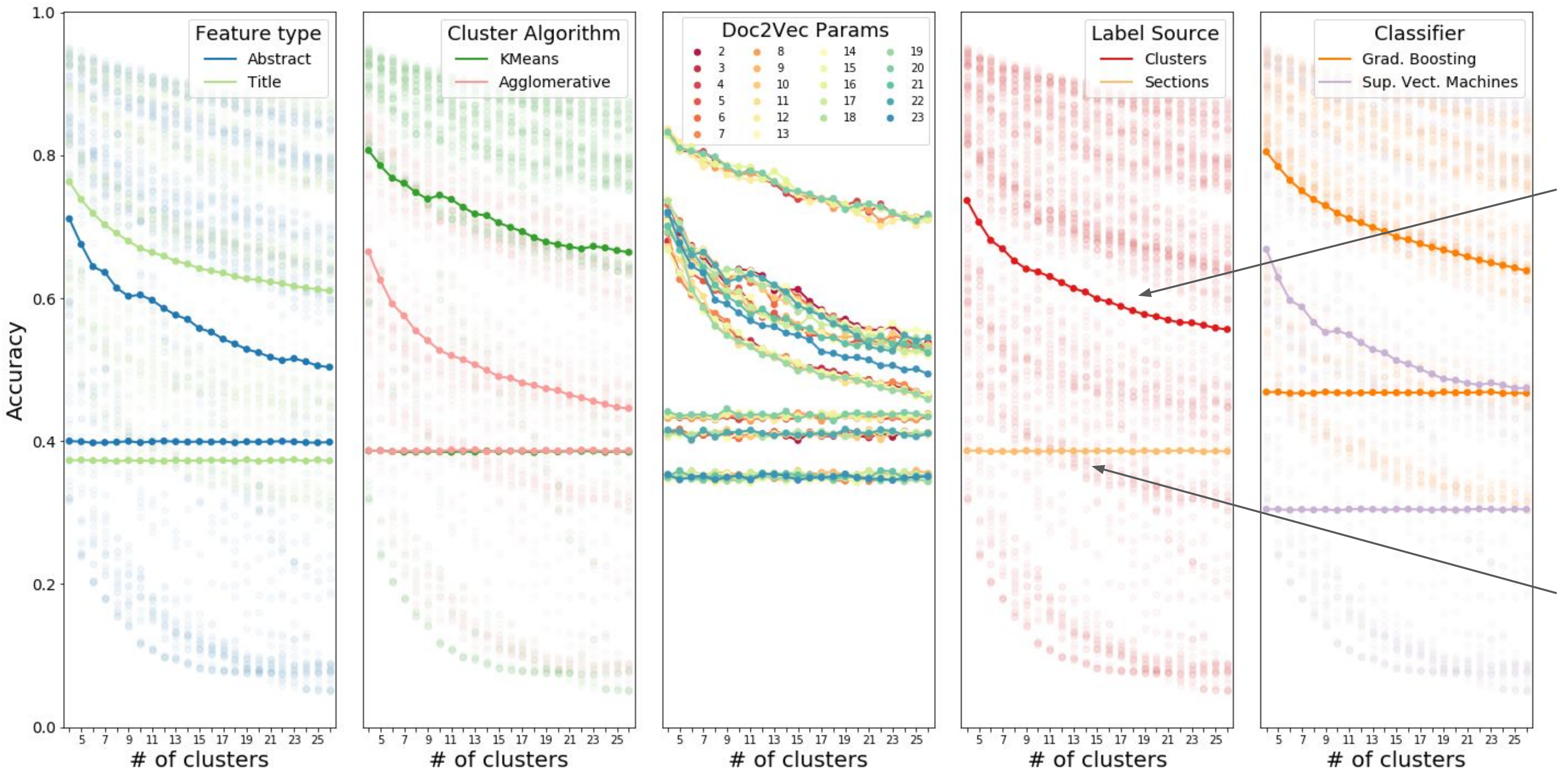
Supervised Classifiers

Experiment 3: Overview

Given training, are section labels the most intuitive labels for abstracts?



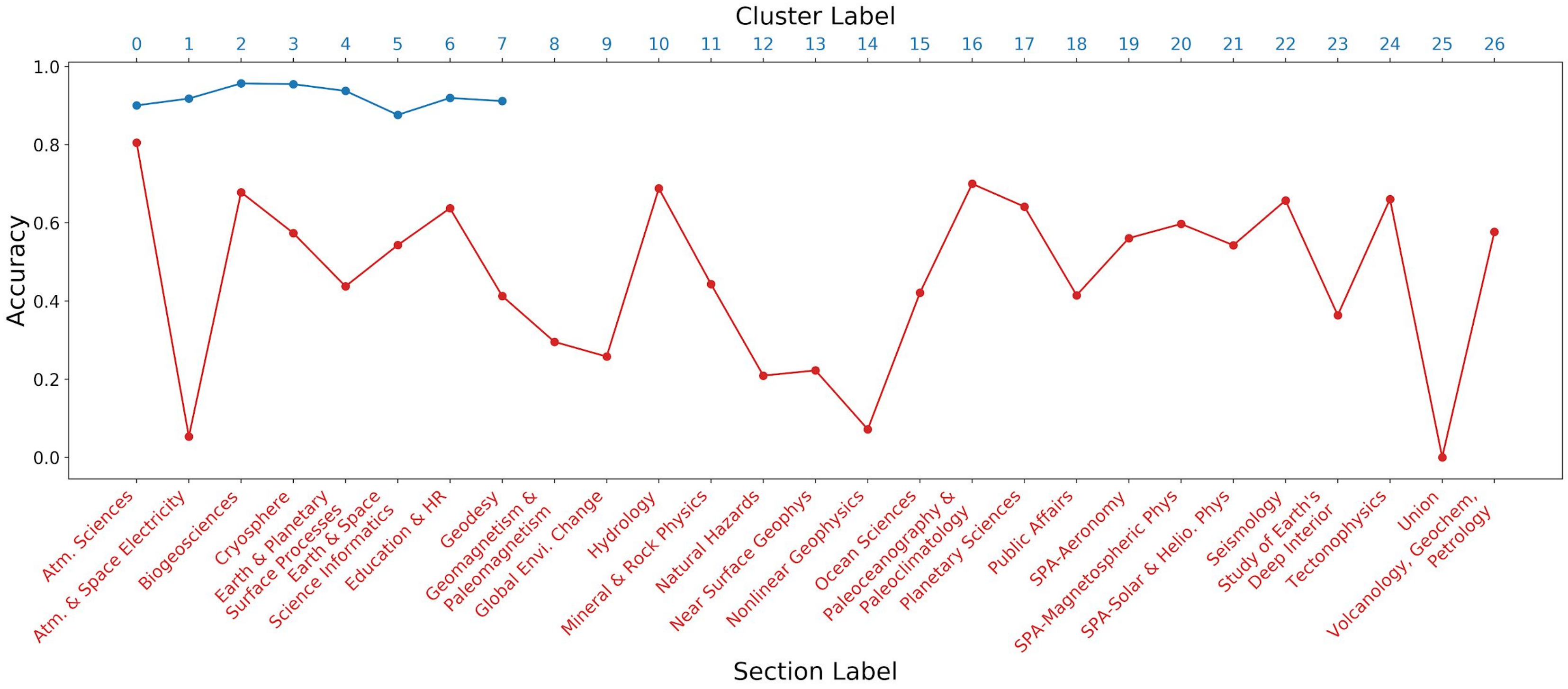
Results: Overall Classifier Accuracy



Classifier accuracy predicting cluster labels for different numbers of clusters

Classifier accuracy predicting section labels (reference lines)

Results: Classifier Accuracy by Label



Doc2Vec param set 8, K-means cluster labels (n=8), Gradient Boosting Classifier

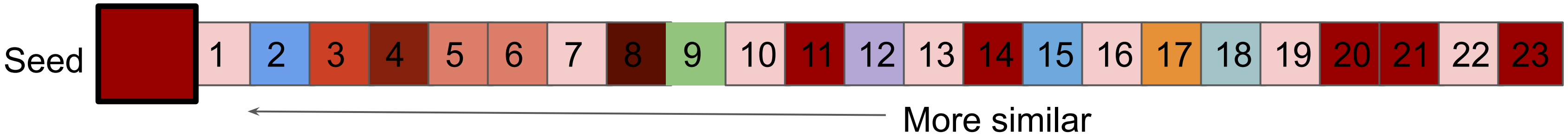
Final Thoughts

- “Similar abstracts” vary considerably with Doc2Vec parameterization
- Can tune parameters to improve d2v similarity score, but haven’t yet calibrated with a human so usefulness is unassessed
- Average number of sections featured in “all section” calculations ~15, suggesting that many abstracts may go undiscovered
- Section labels are useful, but potentially not the only or best way to describe the structure of the AGU program (seen by both clustering coherence and classifier study)
- Different models may perform better for different sections, sessions, or even abstracts,

Experiment 1: Calculations

For each model (call it model n):

- For each of the 25699 abstracts:
  1. Identify the 60 most similar abstracts to seed (subject to section limitation)



  2. Calculate the mean and standard deviation of those similarity scores

Model n	
Limitation: Seed section only	
Seed	Mean of similarity between seed and...
1	4, 11, 14, 20, 21, 23...
… 25699	...
Model avg.	

Model n	
Limitation: Seed section + 4 similar	
Seed	Mean of similarity between seed and...
1	1,3,4,5,6,7,8,10,11,13,14,16,19,20,21,22,23...
… 25699	...
Model avg.	

Model n	
Limitation: All sections	
Seed	Mean of similarity between seed and...
1	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23...
… 25699	...
Model avg.	