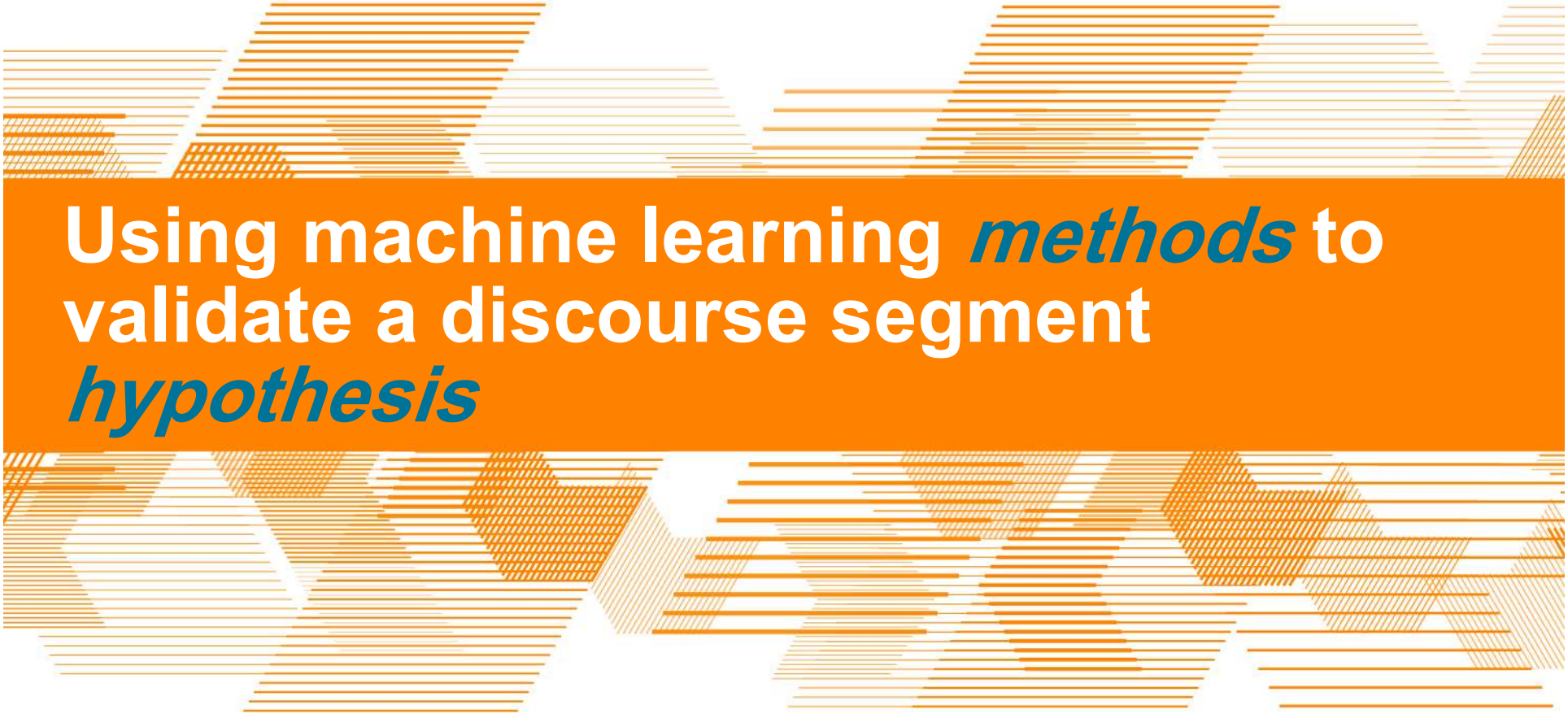
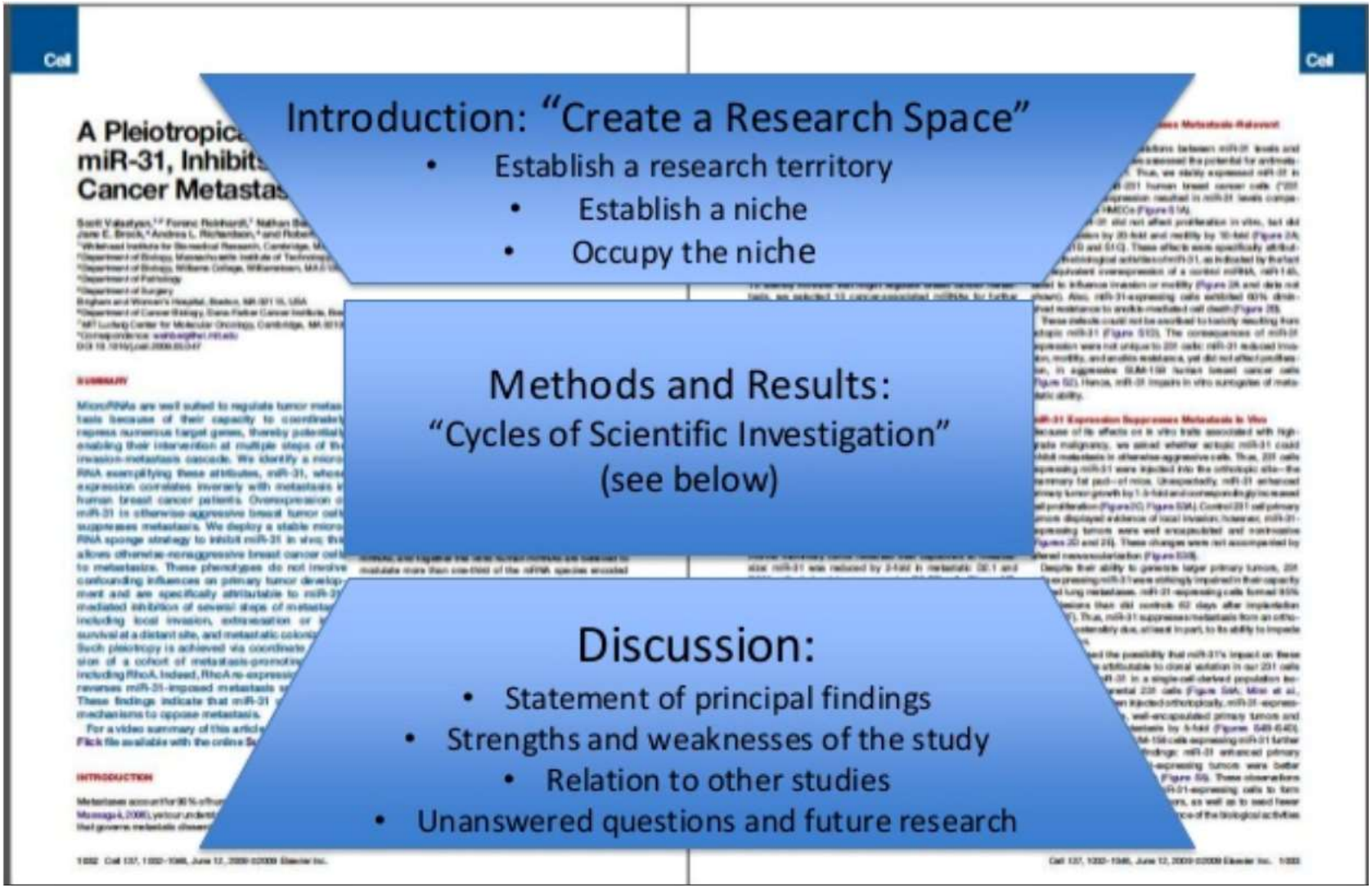


Using machine learning *methods* to validate a discourse segment *hypothesis*

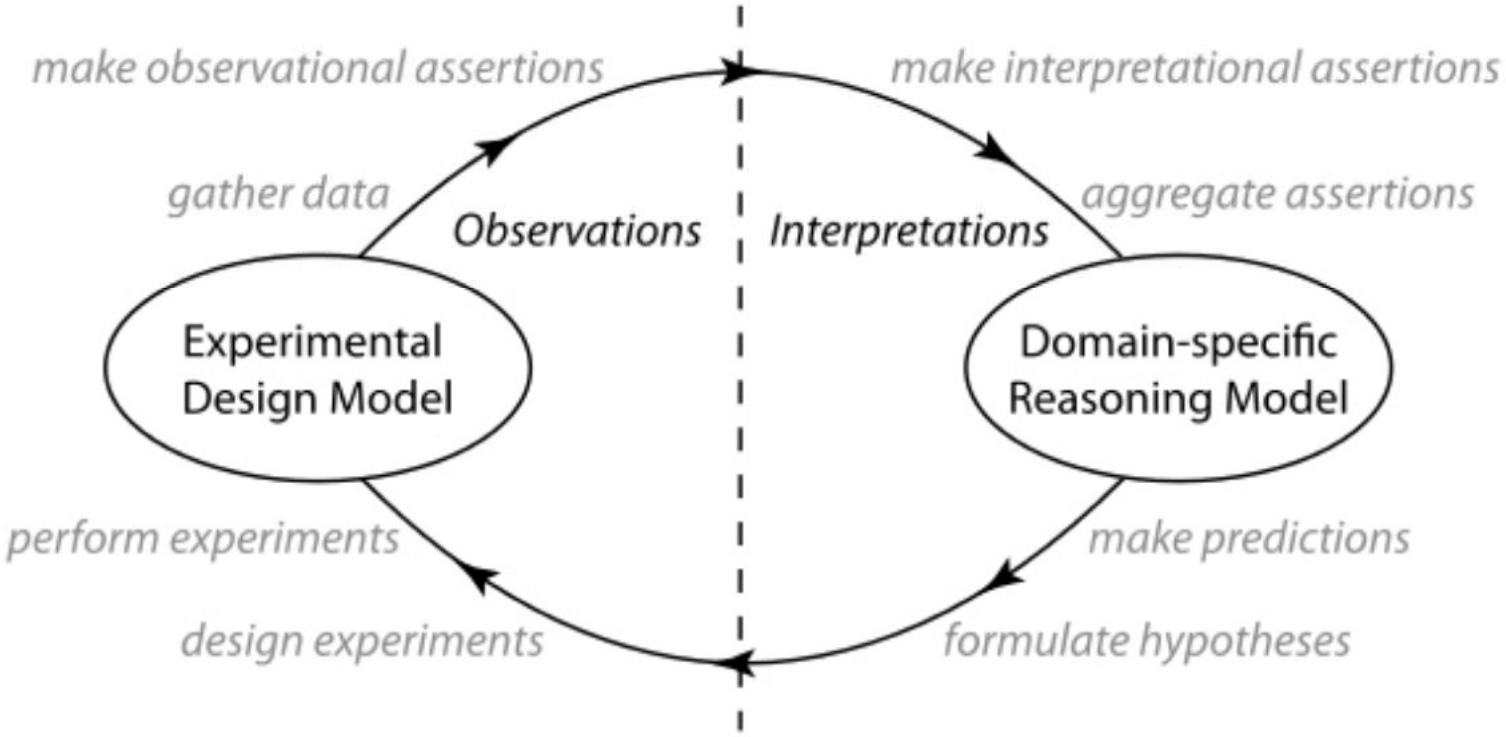
Jessica Cox, Corey A Harper, & Anita de Waard -- SAVE-SD

April 24, 2018

The IMRaD Structure of A Paper



Rubber hits the road in Results: Cycles of Scientific Investigation



The Narrative Structure of Research Articles, Or, Why Science is Like a Fairy Tale

Anita de Waard, VP Research Data Collaborations
Research Data Management Services, Elsevier

Similar to a fairy tale…

Story Grammar		**The Story of Goldilocks and the Three Bears**
Setting	***Time***	Once upon a time
	Character	a little girl named Goldilocks
	Location	She went for a walk in the forest. Pretty soon, she came upon a house.
Theme	***Goal***	She knocked and, when no one answered,
	Attempt	she walked right in.
Episode	***Name***	At the table in the kitchen, there were three bowls of porridge.
	Subgoal	Goldilocks was hungry.
	Attempt	She tasted the porridge from the first bowl.
	Outcome	This porridge is too hot! she exclaimed.
	Attempt	So, she tasted the porridge from the second bowl.
	Outcome	This porridge is too cold, she said

Paper Grammar	**The AXH Domain of Ataxin-1 Mediates Neurodegeneration through Its Interaction with Gfi-1/Senseless Proteins**
Background	The mechanisms mediating SCA1 pathogenesis are still not fully understood, but some general principles have emerged.
Objects of study	the Drosophila Atx-1 homolog (dAtx-1) which lacks a polyQ tract,
Experimental setup	studied and compared in vivo effects and interactions to those of the human protein
Research goal	Gain insight into how Atx-1's function contributes to SCA1 pathogenesis. How these interactions might contribute to the disease process and how they might cause toxicity in only a subset of neurons in SCA1 is not fully understood.
Hypothesis	Atx-1 may play a role in the regulation of gene expression
Name	dAtX-1 and hAtx-1 Induce Similar Phenotypes When Overexpressed in Files
Subgoal	test the function of the AXH domain
Method	overexpressed dAtx-1 in flies using the GAL4/UAS system (Brand and Perrimon, 1993) and compared its effects to those of hAtx-1.
Results	Overexpression of dAtx-1 by Rhodopsin1(Rh1)-GAL4, which drives expression in the differentiated R1-R6 photoreceptor cells (Mollereau et al., 2000 and O'Tousa et al., 1985), results in neurodegeneration in the eye, as does overexpression of hAtx-1[82Q]. Although at 2 days after eclosion, overexpression of either Atx-1 does not show obvious

Discourse Segment Type (DST) Classification

Discourse Segment Type	Definition	Example
Goal	Research goal	*To examine the role of endogenous TGF-β signaling in restraining cell transformation,*
Fact	A known fact, a statement taken to be true by the author.	*Sustained proliferation of cells in the presence of oncogenic signals is a major leap toward tumorigenicity.*
Result	The outcome of an experiment	*Two largely overlapping constructs encoded both miRNA-371 and 372 (miR-Vec-371&2).*
Hypothesis	A claim proposed by the author	*These miRNAs could act on a factor upstream of p53 as a cellular suppressor to oncogenic RAS.*
Method	Experimental method	*We examined p53 mutations in exons five to eight in the primary tumors.*
Problem	An unresolved or contradictory issue	*The mechanism underlying this effect and its conservation to other tissues is not known.*
Implication	An interpretation of the results	*[This indicates that] miR-372/3 acts as a molecular switch.*

ARGUMENTATION IN THE RESULTS SECTION:

1. Importantly, our results so far indicate that the expression of miR-372&3 did not reduce the activity of RASV12, as these cells were still growing faster than normal cells and were tumorigenic, for which RAS activity is indispensable (Hahn et al, 1999 and Kolfschoten et al, 2005).
2. To shed more light on this aspect, we examined the effect of miR-372&3 expression on p53 activation in response to oncogenic stimulation.
3. We used for this experiment BJ/ET cells containing p14ARFkd because, following RASV12 treatment, in those cells p53 is still activated but more clearly stabilized than in parental BJ/ET cells (Voorhoeve and Agami, 2003), resulting in a sensitized system for slight alterations in p53 in response to RASV12.
4. Figure 4A shows that following RASV12 stimulation, p53 was stabilized and activated, and its target gene, p21cip1, was induced in all cases, indicating an intact p53 pathway in these cells.

Reg-clause	Fact	Goal	Method	Result	Implication

Hypothesis (not shown)	Problem (not shown)

Potential Applications

- Text Summarization
- Hypothesis Formulation
- Citation Analysis
- Identifying Methods and Protocols
  - For extraction to Lab Notebooks
- Knowledge Extraction
  - Is extracted data a hypothesis, claim, or fact?
- Figure and Table Interpretation
  - Which figures represent your ”results”

Networks of Claims and Evidence

Claim:

- sustained miR-31 activity is necessary to prevent the acquisition of aggressive traits by both tumor cells and untransformed breast epithelial

Evidence: Method:

- We transiently inhibited miR-31 in noninvasive MCF7-Ras cells with either antisense oligonucleotides or miRNA sponges.

Evidence: Result:

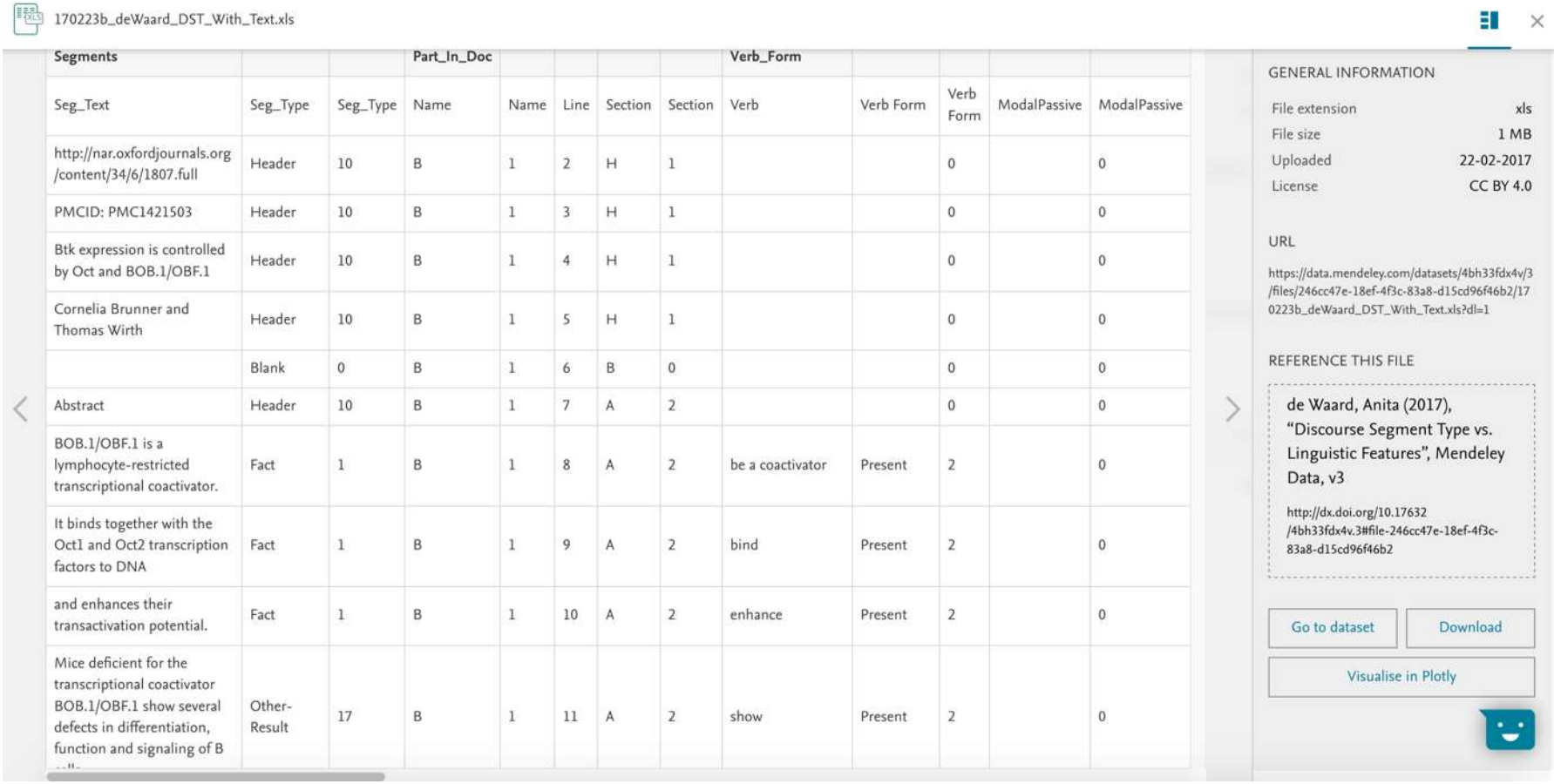
- Both approaches inhibited miR-31 function by >4.5-fold (Figure S7A).
- Suppression of miR-31 enhanced invasion by 20-fold and motility by 5-fold, but cell viability was unaffected by either inhibitor (Figure 3A; Figure S7B).
- The miR-31 sponge reduced miR-31 function by 2.5-fold, but did not affect the activity of other known antimetastatic miRNAs (Figures S8A and S8B).

Is it pertinent? -> Probably

Is it true? -> Sounds likely!

Is it new, but in agreement with what I know? -> Check/know

Dataset of “Discourse Segment Type vs. Linguistic Features”



170223b_deWaard_DST_With_Text.xls

Segments			Part_In_Doc					Verb_Form				
Seg_Text	Seg_Type	Seg_Type	Name	Name	Line	Section	Section	Verb	Verb Form	Verb Form	ModalPassive	ModalPassive
http://nar.oxfordjournals.org/content/34/6/1807.full	Header	10	B	1	2	H	1			0		0
PMCID: PMC1421503	Header	10	B	1	3	H	1			0		0
Btk expression is controlled by Oct and BOB.1/OBF.1	Header	10	B	1	4	H	1			0		0
Cornelia Brunner and Thomas Wirth	Header	10	B	1	5	H	1			0		0
	Blank	0	B	1	6	B	0			0		0
Abstract	Header	10	B	1	7	A	2			0		0
BOB.1/OBF.1 is a lymphocyte-restricted transcriptional coactivator.	Fact	1	B	1	8	A	2	be a coactivator	Present	2		0
It binds together with the Oct1 and Oct2 transcription factors to DNA	Fact	1	B	1	9	A	2	bind	Present	2		0
and enhances their transactivation potential.	Fact	1	B	1	10	A	2	enhance	Present	2		0
Mice deficient for the transcriptional coactivator BOB.1/OBF.1 show several defects in differentiation, function and signaling of B	Other-Result	17	B	1	11	A	2	show	Present	2		0

GENERAL INFORMATION

File extension	xls
File size	1 MB
Uploaded	22-02-2017
License	CC BY 4.0

URL

https://data.mendeley.com/datasets/4bh33fdx4v/3/files/246cc47e-18ef-4f3c-83a8-d15cd96f46b2/170223b_deWaard_DST_With_Text.xls?dl=1

REFERENCE THIS FILE

de Waard, Anita (2017), “Discourse Segment Type vs. Linguistic Features”, Mendeley Data, v3

http://dx.doi.org/10.17632/4bh33fdx4v.3#file-246cc47e-18ef-4f3c-83a8-d15cd96f46b2

Go to dataset | Download

Visualise in Plotly

https://data.mendeley.com/datasets/4bh33fdx4v/3

Features / Feature Groups

Frequently Used Verb	
Top 10 Verb	”Show” verb

Verb Tense				
Future	Gerund	Past	Past participle	Past perfect
Past progressive	Present	Present perfect	Present progressive	To-infinitive

Verb Class				
Cause and effect	Change and growth	Discourse verb	Interpretation	Investigation
None	Observation	Prediction	Procedure	Properties

Modality Marker				
Modal	Verb class interpretation	Ruled by verb class interpretation	Reference internal	Reference external
First person	Modal significant_ly	Possible_ility_ly	Potential_ly	UN_Likely
Sum_Adverbs_YesNO				

Summary Statistics

Class Distribution

```
In [6]: df.Seg_Type_name.value_counts()

Out[6]: Result                625
        Implication           335
        Method                321
        Blank                 316
        Reg-Implication       205
        Fact                  202
        Hypothesis            197
        Other-Result          167
        Goal                  138
        Header                134
        Other-Implication     120
        Problem                97
        Reg-Hypothesis         93
        Intratextual           72
        Reg-Result             61
        Other-Fact             53
        Other-Method           31
        Other-Hypothesis       26
        Intertextual           14
        Reg-Problem            10
        Reg-Fact                7
        Other-Goal              6
        Reg-Goal                5
        Other-Problem           3
                                1
        Name: Seg_Type_name, dtype: int64
```

Collapsed and Filtered

- Subcategories normalized to parent
- *Blank* and *Header* rows removed
- *Intertextual* and *Intratextual* removed

```
In [24]: df1['Seg_Type_name'].value_cc

Out[24]: Result          853
         Implication     660
         Method          352
         Hypothesis      316
         Fact            262
         Goal            149
         Problem         110
         Intratextual     72
         Intertextual     14
         Name: Seg_Type_name, dtype: i
```

Verb Class & Modality Distributions

```
In [7]: df.Verb_Class.value_counts()

Out[7]: Cause and Effect      716
        Interpretation        536
                              447
        Properties            330
        Procedure             315
        None                  217
        Investigation         201
        Change and Growth     153
        Observation           148
        Discourse verb        105
        Prediction             70
        Discourse Verb          1
        Name: Verb_Class, dtype: int64

In [8]: df.Modal.value_counts()

Out[8]: 0    3052
        1     186
              1
        Name: Modal, dtype: int64
```

Machine Learning Models

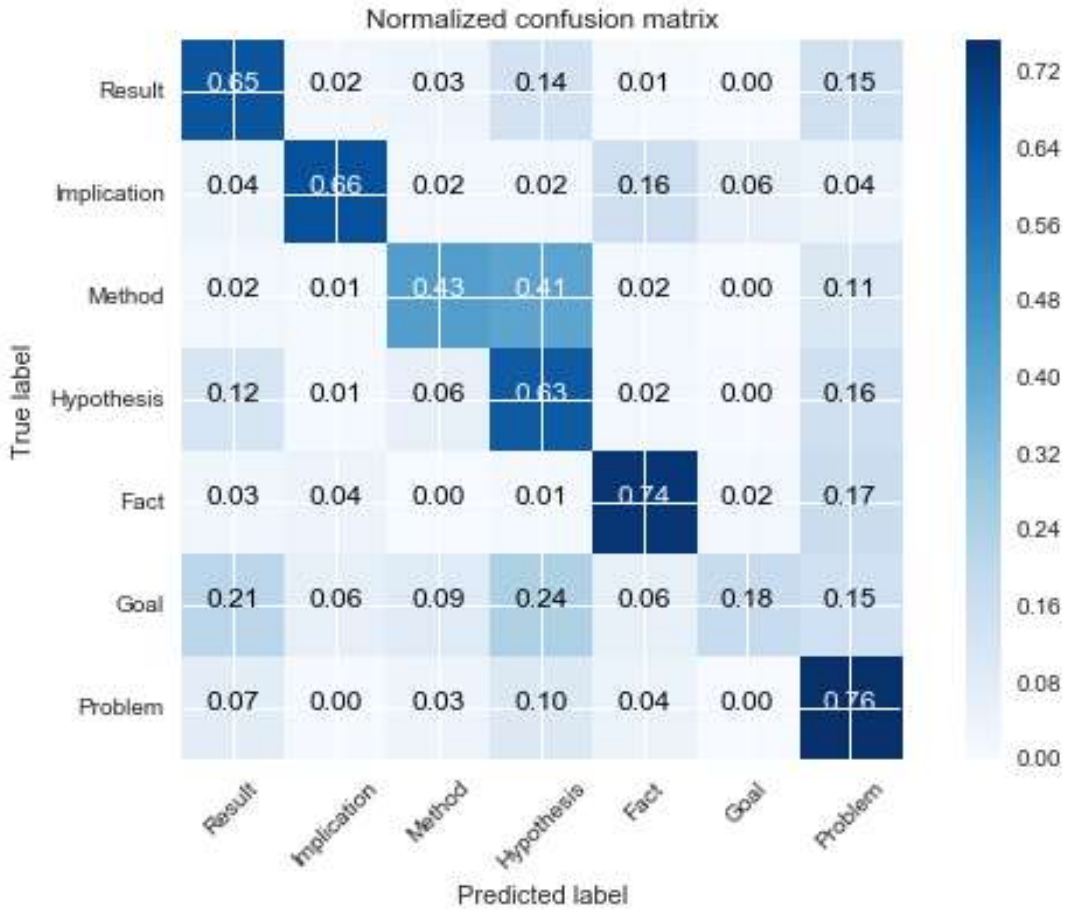
Methods Explored

- Baseline: Logistic Regression, Decision Tree, Random Forest
- Same classifiers using *Class Balancers*
- SciDT Neural Network for comparison
- Feature Reduction / Ablation
- Mini-experiment with binary targets and limited feature set

Baseline Random Forest Results

Accuracy score 0.6427688504326329

	precision	recall	f1-score	support
Fact	0.53	0.67	0.59	88
Goal	0.79	0.60	0.68	50
Hypothesis	0.56	0.44	0.49	93
Implication	0.61	0.60	0.61	210
Method	0.78	0.75	0.77	106
Problem	0.60	0.18	0.28	33
Result	0.67	0.77	0.72	229
avg / total	0.64	0.64	0.64	809


Normalized confusion matrix
Result 0.65 0.02 0.03 0.14 0.01 0.00 0.15
Implication 0.04 0.66 0.02 0.02 0.16 0.06 0.04
Method 0.02 0.01 0.43 0.41 0.02 0.00 0.11
Hypothesis 0.12 0.01 0.06 0.63 0.02 0.00 0.16
Fact 0.03 0.04 0.00 0.01 0.74 0.02 0.17
Goal 0.21 0.06 0.09 0.24 0.06 0.18 0.15
Problem 0.07 0.00 0.03 0.10 0.04 0.00 0.76
True label
Result Implication Method Hypothesis Fact Goal Problem
Predicted label
0.72 0.64 0.56 0.48 0.40 0.32 0.24 0.16 0.08 0.00

RandomUnderSampler	Undersamples the majority classes by randomly picking samples
Tomeklinks	**Undersamples the majority classes by removing Tomek's links**
ClusterCentroids	Under samples the majority classes by replacing a cluster of the majority samples by the cluster centroid of a KMeans algorithm
CondensedNearestNeighbor	Under samples the majority classes using the condensed nearest neighbor method
OneSidedSelection	Uses one-sided selection method on majority classes
InstanceHardnessThreshold	Samples with lower probabilities are removed from the majority class
RandomOverSampler	Randomly generates new samples from the minority classes
SMOTE	Synthetic Minority Oversampling Technique; generates new samples of minority class by interpolation
SMOTEborderline	Generates new samples of minority class specific to the borders between two classes.
SMOTEborderline2	Generates new samples of minority class specific to the borders between two classes.
SMOTETomek	Combines use of SMOTE on minority class and Tomek Links on majority class
SMOTEENN	Combines use of SMOTE on minority class and Edited

Class balancers had minimal to no effect

Appendix 5.3. Accuracy, precision, recall and F1 scores of all 36 models tested.

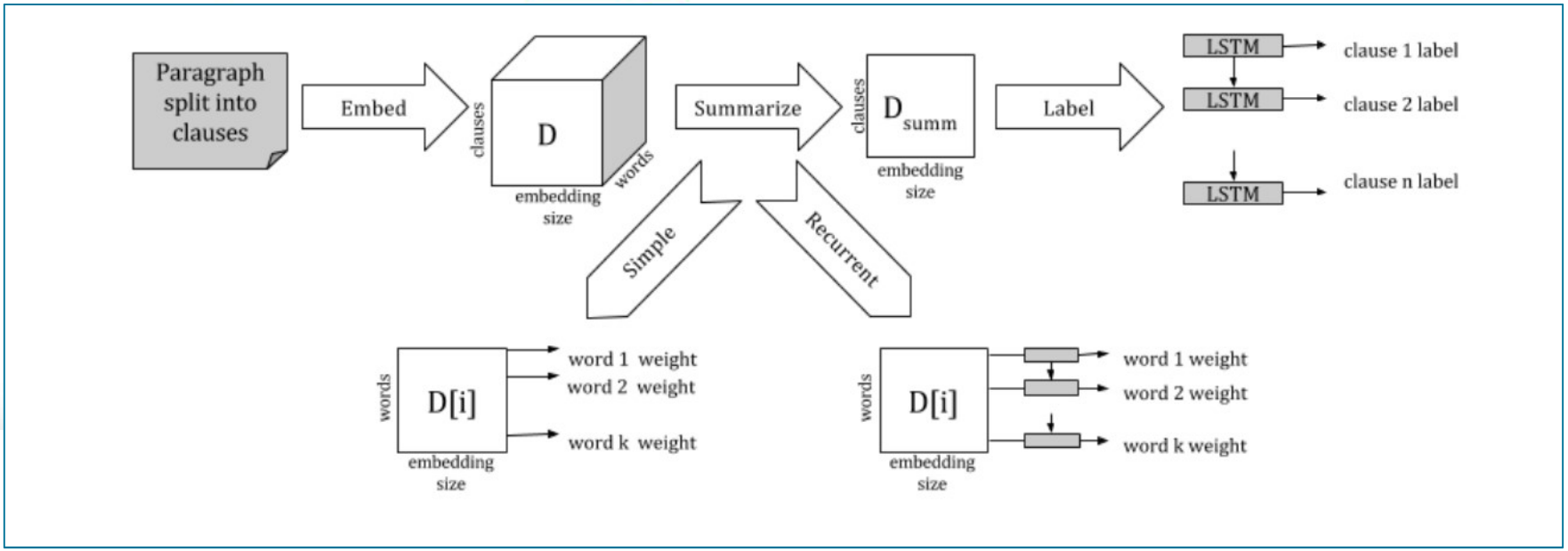
Classifier	Class Balancer	Accuracy	Precision	Recall	F1
LR	No Class Balancer	0.62	0.68	0.63	0.64
DTC	No Class Balancer	0.64	0.64	0.64	0.64
RFC	No Class Balancer	0.64	0.65	0.65	0.64
LR	RandomUnderSampler	0.58	0.64	0.58	0.59
DTC	RandomUnderSampler	0.55	0.64	0.55	0.56
RFC	RandomUnderSampler	0.57	0.63	0.56	0.57
LR	Tomeklinks	0.63	**0.68**	0.63	0.64
DTC	Tomeklinks	**0.64**	0.64	**0.64**	0.64
RFC	Tomeklinks	**0.64**	0.64	**0.64**	0.64
LR	ClusterCentroids	0.55	0.64	0.55	0.55
DTC	ClusterCentroids	0.35	0.48	0.35	0.32
RFC	ClusterCentroids	0.38	0.47	0.38	0.35

Scientific Discourse Tagger

Experiment Segmentation in Scientific Discourse as Clause-level Structured Prediction using Recurrent Neural Networks

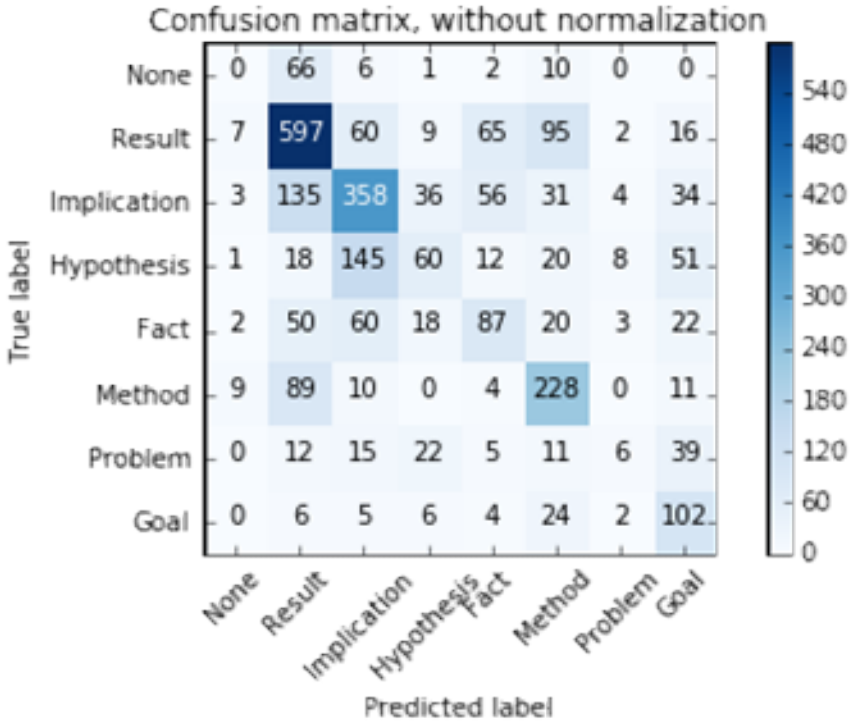
Pradeep Dasigi[1], Gully A.P.C. Burns[2], Eduard Hovy[1], and Anita de Waard[3]

[1]Language Technologies Institute , Carnegie Mellon University , 5000 Forbes Avenue , Pittsburgh, PA 15213 , USA
[2]Information Sciences Institute , Viterbi School of Engineering , University of Southern California , Marina del Rey, CA 90292 , USA
[3]Elsevier Research Data Services , Jericho, VT 05465 , USA



https://github.com/edvisees/sciDT
https://arxiv.org/pdf/1702.05398.pdf

	precision	recall	f1-score	support
0	0.00	0.00	0.00	85
1	0.61	0.70	0.65	851
2	0.54	0.54	0.54	657
3	0.39	0.19	0.26	315
4	0.37	0.33	0.35	262
5	0.52	0.65	0.58	351
6	0.24	0.05	0.09	110
7	0.37	0.68	0.48	149
avg / total	0.49	0.52	0.49	2780



Feature Reduction and Forward Ablation

- Reduce to 13 Features: F1 increases .65 before dropping

Accuracy score 0.6477132262051916

	precision	recall	f1-score	support
Fact	0.43	0.81	0.56	88
Goal	0.69	0.72	0.71	50
Hypothesis	0.64	0.47	0.54	93
Implication	0.65	0.65	0.65	210
Method	0.86	0.76	0.81	106
Problem	0.27	0.24	0.25	33
Result	0.78	0.65	0.71	229
avg / total	0.68	0.65	0.65	809

Feature Reduction and Forward Ablation

- Forward Ablation to 9 Features before F1 levels off.

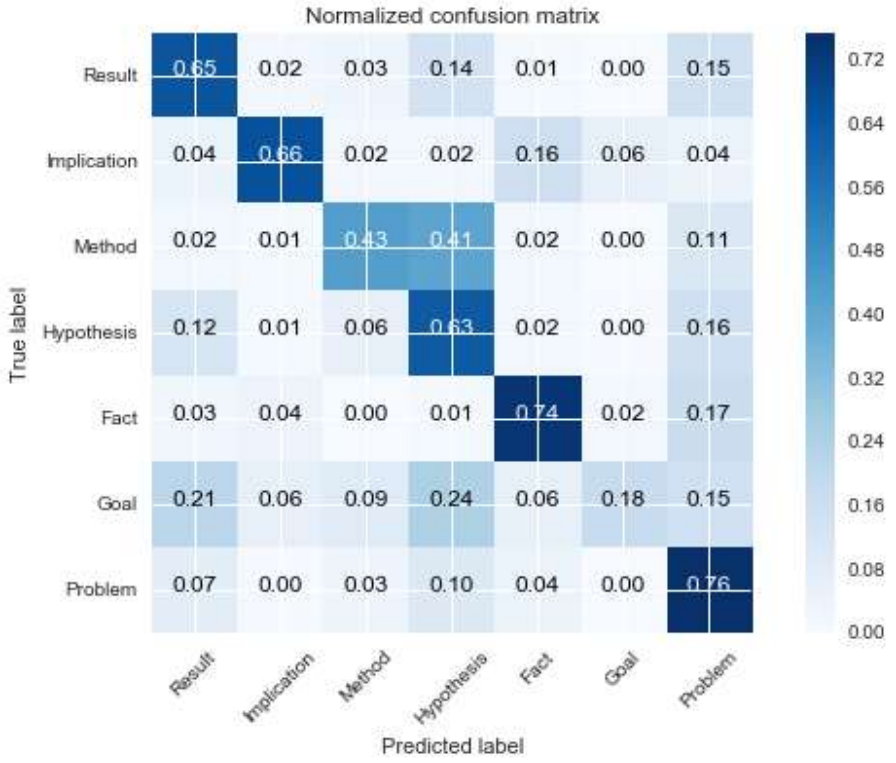
Accuracy score 0.6489493201483313

F Score 0.6445074039326106

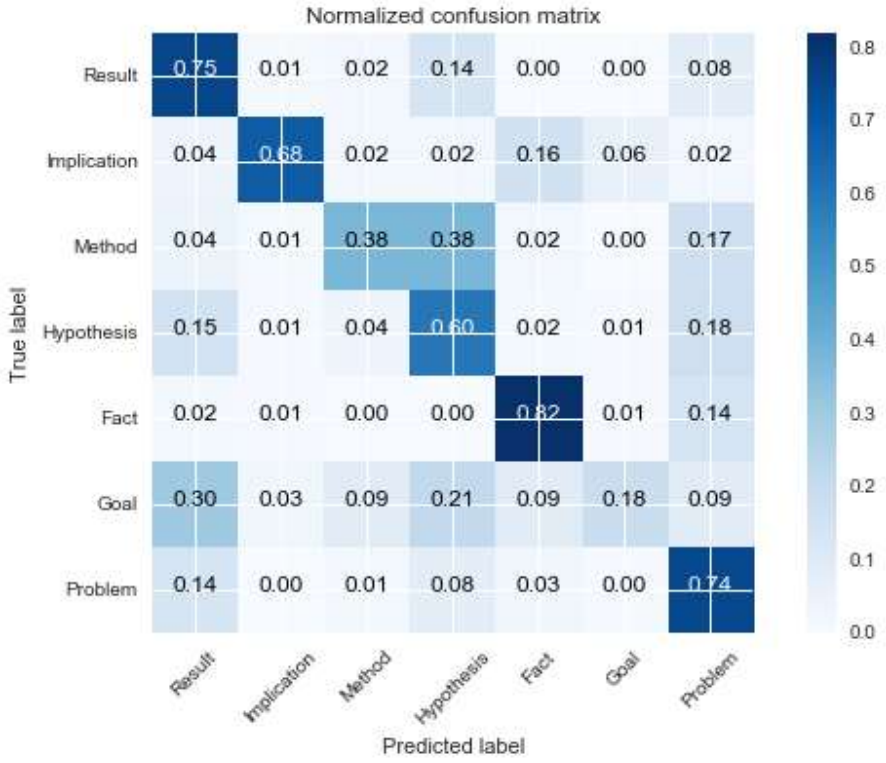
	precision	recall	f1-score	support
Fact	0.45	0.75	0.56	88
Goal	0.84	0.64	0.73	50
Hypothesis	0.67	0.38	0.48	93
Implication	0.61	0.69	0.64	210
Method	0.79	0.82	0.81	106
Problem	0.40	0.18	0.25	33
Result	0.74	0.68	0.71	229
avg / total	0.67	0.65	0.64	809

Minimal Change to Confusion Matrix

All Features



9 Features



Significant Features

Feature Class	Feature
Verb Tense	Past
Verb Tense	Present
Verb Tense	To-infinitive
Verb Class	Interpretation
Verb Class	Investigation
Verb Class	Procedure
Modality Marker	Modal
Modality Marker	Verb class interpretation
Modality Marker	Ruled by verb class interpretation

Verb Tense Mini-Experiment

Binary Verb Tense Features
Future
Gerund
Past
Past participle
Past perfect
Past progressive
Present
Present perfect
Present progressive
To-infinitive

Binary Class Labels
Result/Method
Fact/Implication

Table 7: Performance metrics of 3 models to evaluate segment type based on verb tense.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 score**
Logistic Regression	0.80	0.81	0.80	0.80
Decision Tree Classifier	0.81	0.82	0.81	0.81
Random Forest	0.81	0.82	0.81	0.81

Table 10. Random forest classifier model confusion matrix

True label	*Result/Method*	236	39
	Fact/impliImplication cation	84	278
		Result/Method	*Fact/Implication*
		Predicted label	

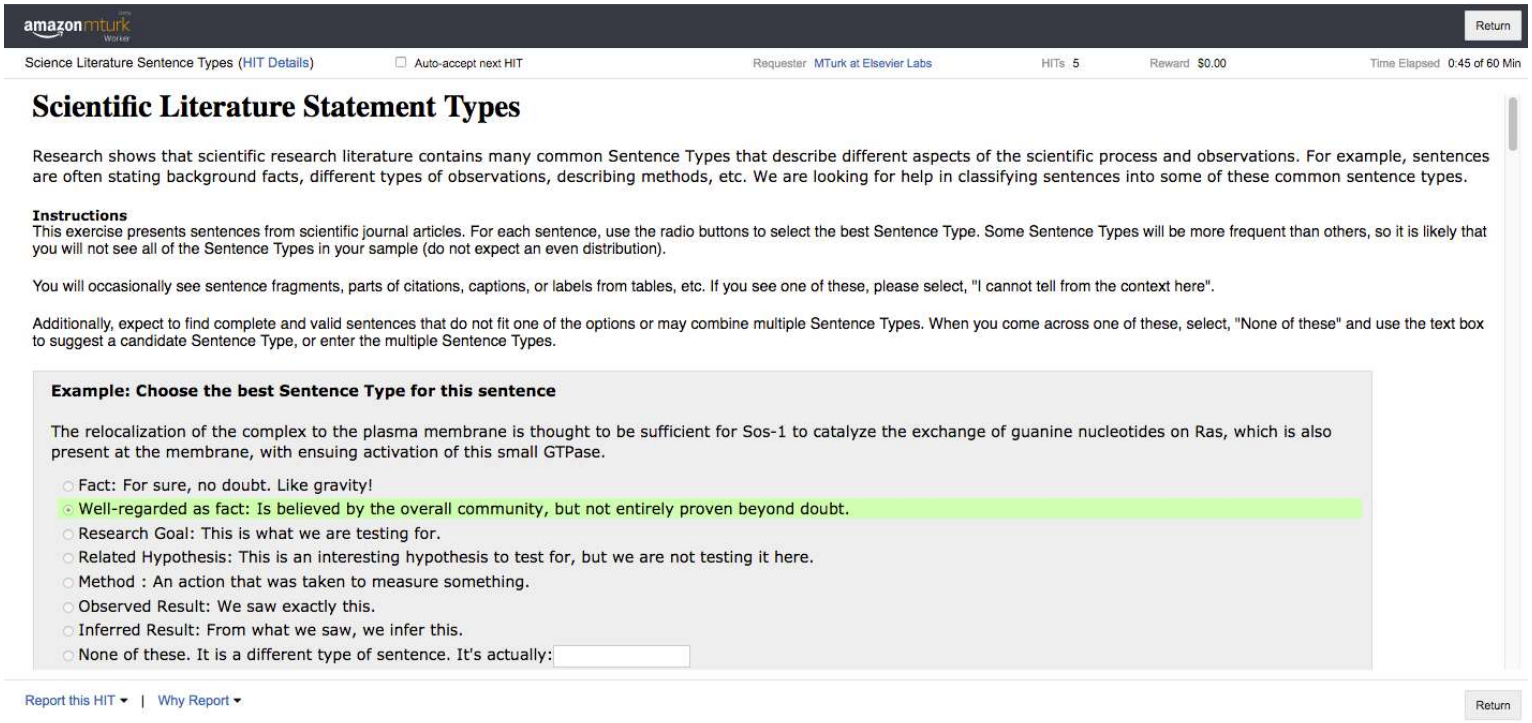
Conclusions And Future Work

Conclusions

- Verb tense, class, and modality are reasonably good predictors of discourse segment type
- Class prediction success doesn't line up with majority / minority classes
- Verb tense alone is very good at distinguishing Result/Method from Fact/Implication
- There are some limits to our methodology

Future Work

- Build larger training sets of segment types
  - Mechanical Turk experiment underway



Future Work

- Evaluate automated extraction of features
- Looking at Snorkel as a a tool for merging multiple noisy classifiers
  - Majority voting: F1 of .64
  - Generative model: F1 of .58
  - Based on Sujit Pal's post: http://sujitpal.blogspot.co.uk/2018/01/cleaning-up-noisy-labels-using-snorkels.html

Corey A Harper
@chrpr
c.harper@elsevier.com

Empowering Knowledge

We're hiring!