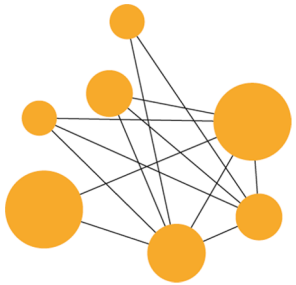
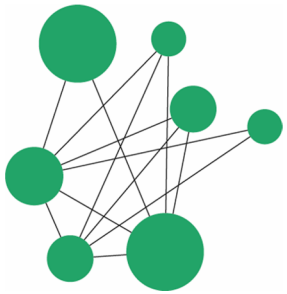
Reviewing Data Science Research: Evaluating *Results* sections

With examples from machine learning and natural language processing

Kevin Bretonnel Cohen
Director, Biomedical Text Mining Group,
University of Colorado School of *Medicine*;
Emeritus D'Alembert Chair in Natural Language Processing for the Biomedical Domain, Université Paris-Saclay





kevin.cohen@gmail.com
http://compbio.ucdenver.edu/Hunter_lab/Cohen

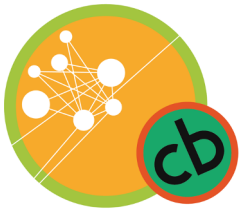
Karën Fort, Margot Mieskes, and Aurélie Névéol

Kevin Bretonnel Cohen,
UCSOM

Karën Fort,
Sorbonne
Université / Loria

Margot Mieskes,
h_da Darmstadt

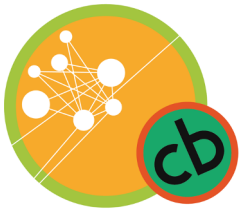
Aurélie Névéol,
Université
Paris Saclay,
CNRS, LIMSI



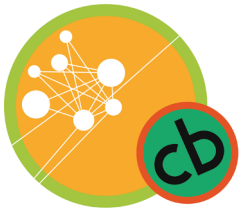
WE ARE TALKING ABOUT "RESULTS" SECTIONS BECAUSE THEY ARE A NEAR-UNIVERSAL SECTION OF PAPERS IN OUR FIELD

Typical elements of a Results section

- Tables
- Figures
- …that convey **values and relationships** between values
- Error analysis

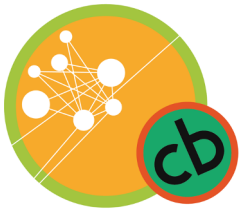


THIS IS OFTEN WHERE YOU CAN BE THE MOST HELPFUL TO THE AUTHORS



Exercise:
Strength, or weakness?

1. The assertion that the system achieves state-of-the-art performance is only supported by a single figure of merit. See *Empirical Methods for Artificial Intelligence* for why this is an important weakness of the paper.
2. The assertion that the system achieves state-of-the-art performance is supported by multiple metrics. See *Empirical Methods for Artificial Intelligence* for why this is a strength of the paper.



A checklist for the *Results* section

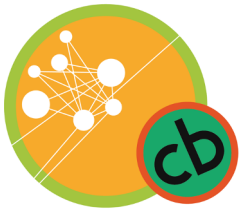
- ❑ Metric is appropriate for the task
- ❑ Baselines are non-trivial
- ❑ Measures of dispersion are reported
- ❑ Graphs:
  - ❑ Axes are not truncated
  - ❑ Axes are labelled
- ❑ Parameteric statistics used only with normally distributed data
- ❑ Error analysis is present
- ❑ Error analysis is non-trivial
- ❑ Alternative analyses are considered

Measures of Central Tendency: mean, median, mode

MEAN/MEDIAN/MODE NEED MEASURES OF DISPERSION

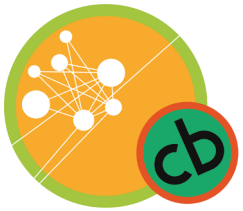
Measures of central tendency with their typical measures of dispersion

Measure of central tendency	Measure of dispersion
Mean	Standard deviation
Median	Interquartile range
Mode	Range?

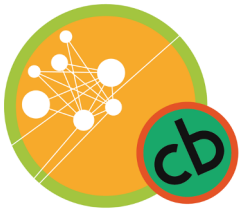


Exercise:
Strength, or weakness?

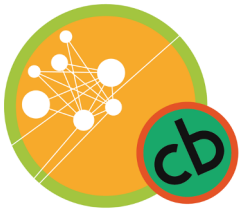
1. The work is evaluated via cross-validation, but does not report measures of dispersion. Without those measures of dispersion, it is difficult to have confidence in the stability of the results.
2. The work reports measures of dispersion, which increases confidence that we understand the stability of the results (or lack thereof).



...EVEN WHEN THERE IS NO HYPOTHESIS TEST!

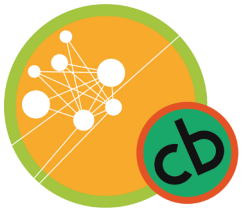


PARAMETRIC HYPOTHESIS TESTS NEED NORMALLY DISTRIBUTED DATA



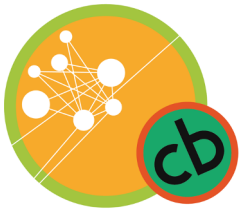
Exercise:
Strength, or weakness?

1. The work is evaluated via statistical hypothesis testing using a parametric test, but does not demonstrate that the data is normally distributed.
2. The work demonstrates that the distribution of the data is appropriate for the hypothesis test.

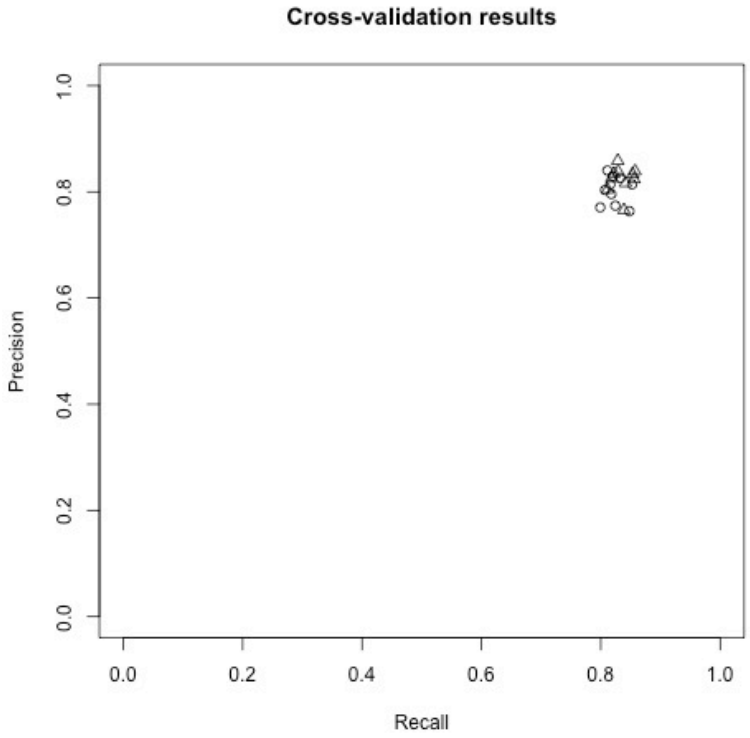


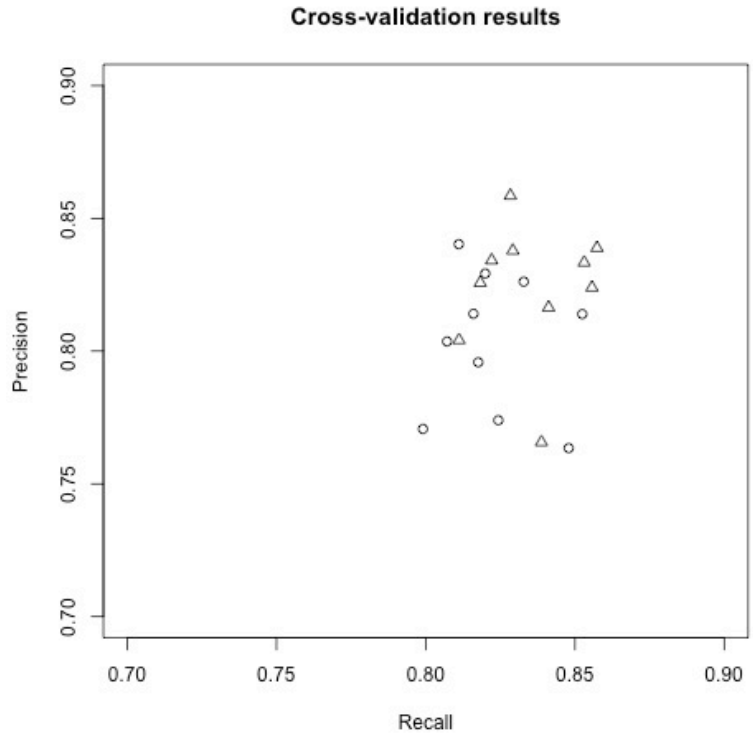
For example: left to right, top to bottom, axes to quadrants

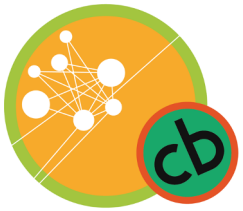
DEVELOP A SYSTEM FOR LOOKING AT GRAPHS



Left to right, top to bottom, axes to quadrants

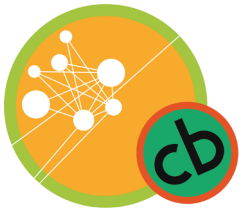




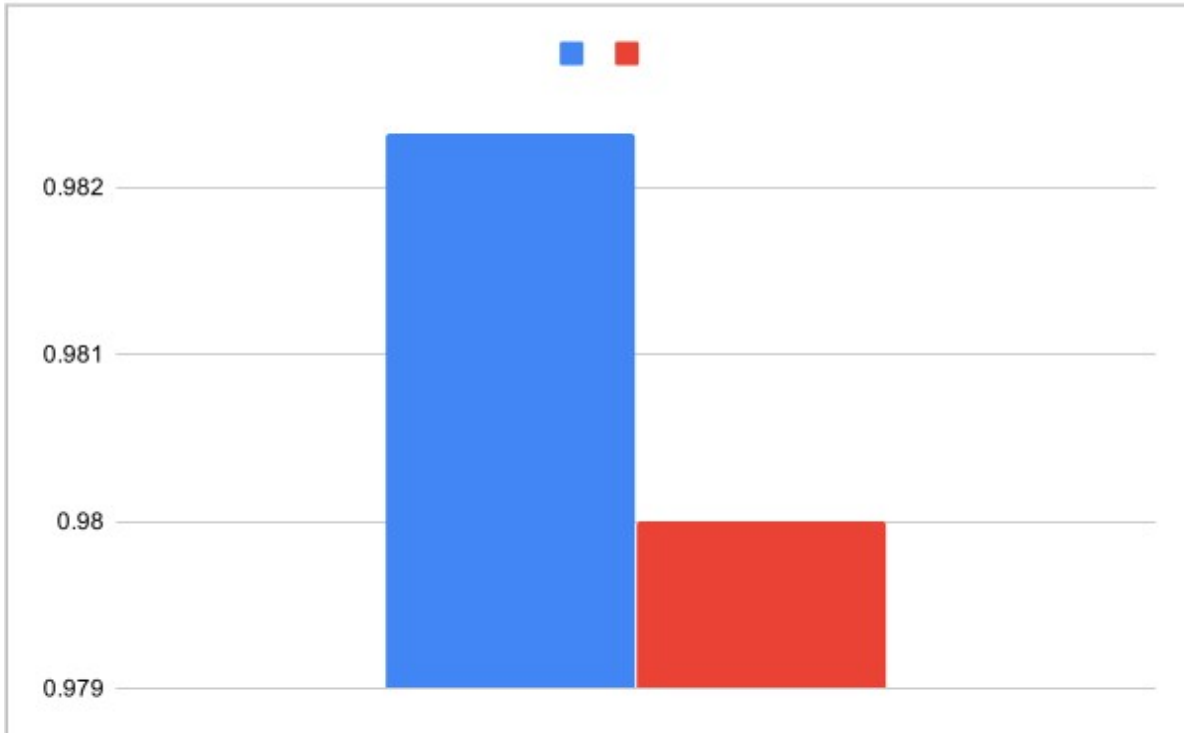


When the actual values are important, use a table

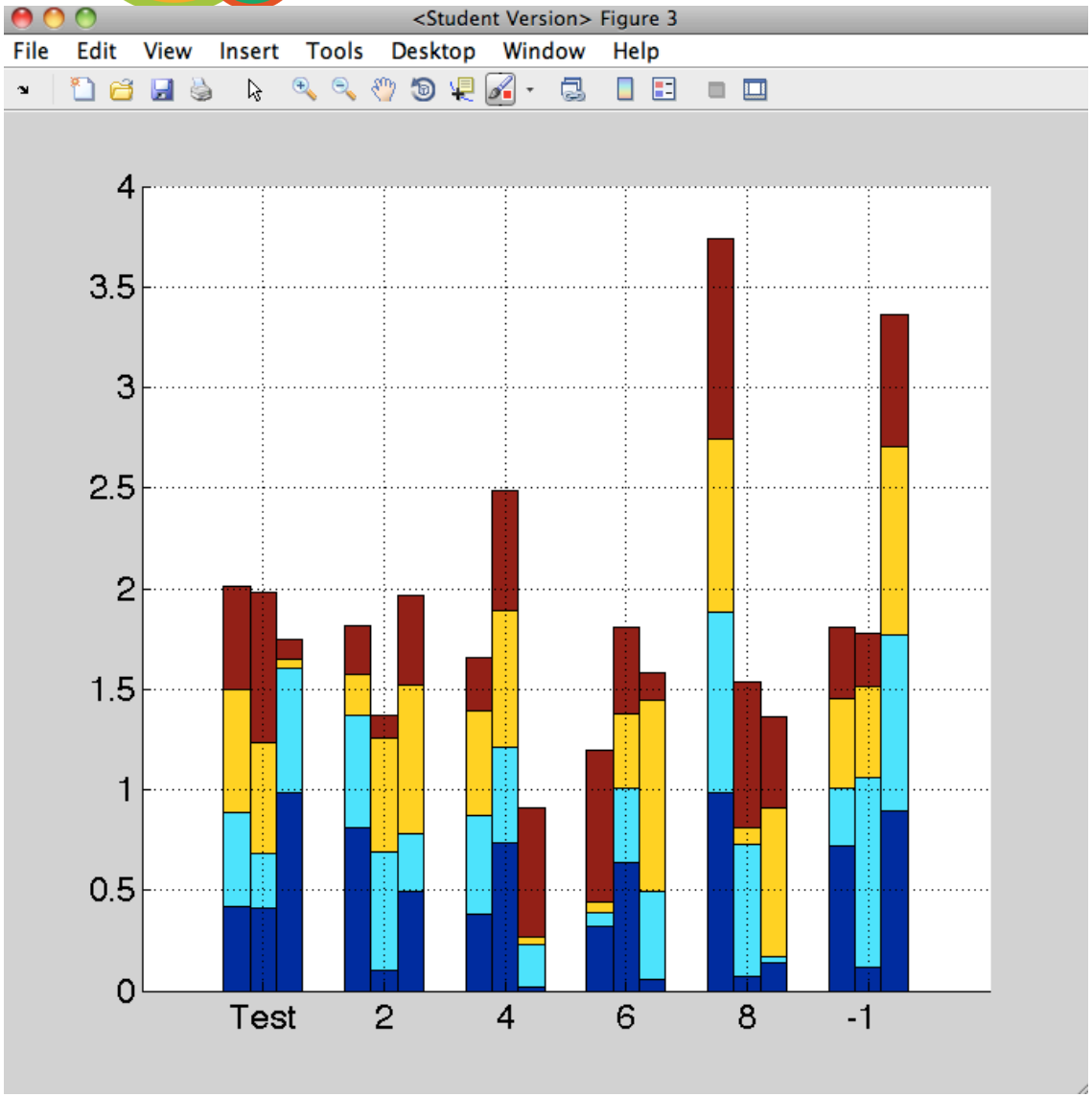
HOW TO BE HELPFUL TO THE AUTHOR



Exercise:
Is this figure good, bad, or mediocre? Why?



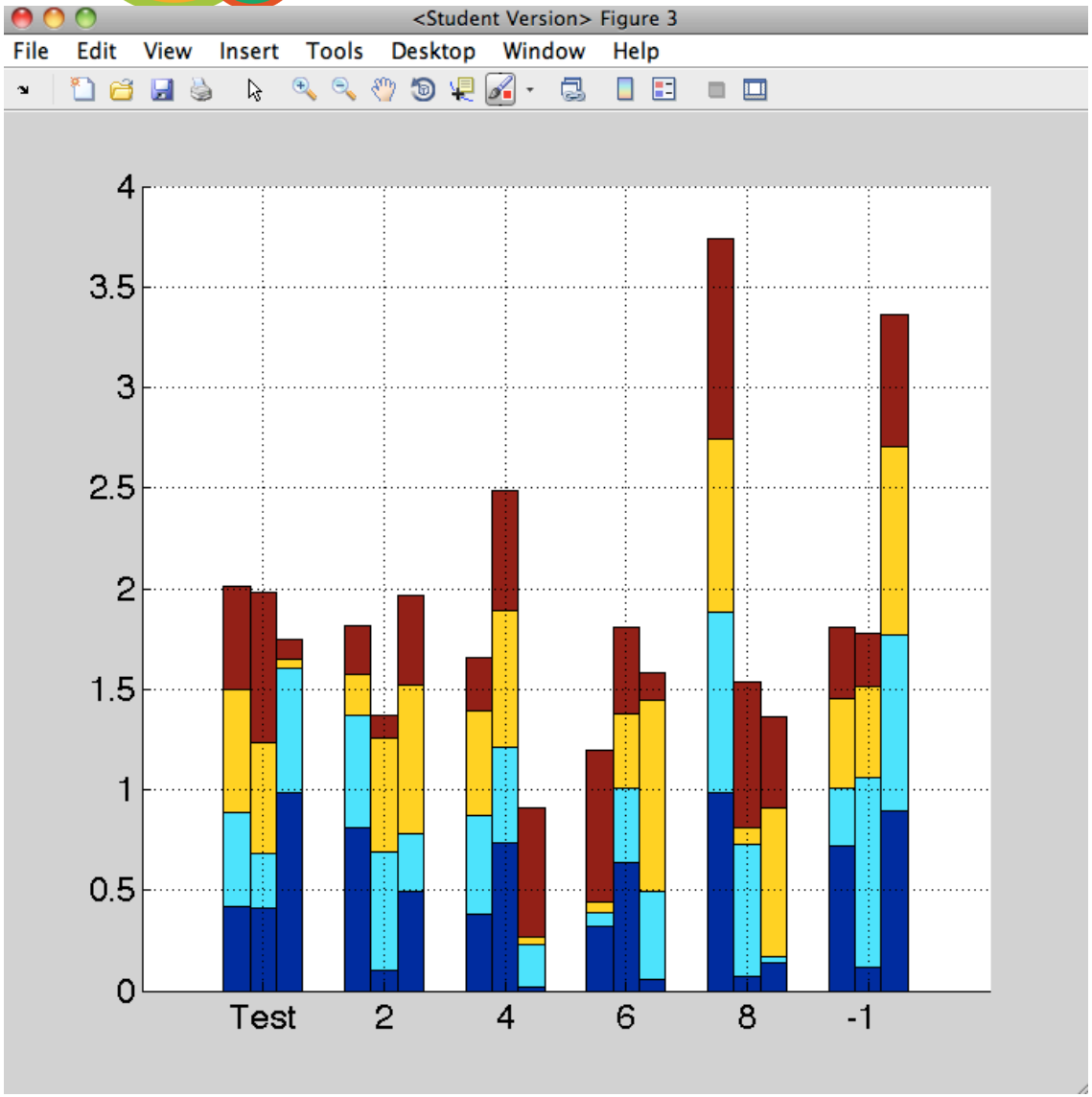
Clarity counts! Exercise…



How many independent variables are being graphed?

https://www.mathworks.com/matlabcentral/fileexchange/32884-plot-groups-of-stacked-bars

Exercise…



How many independent variables are being graphed?

1. Group (Test/2/4/etc.)
2. 3 bars in each group
3. Stack within each bar

https://www.mathworks.com/matlabcentral/fileexchange/32884-plot-groups-of-stacked-bars

Clarity counts!

This table is extremely difficult to parse because of its awkward formatting. The required eye travel is across a row, but the eye naturally travels down a column due to the fact that column spacing is relatively larger than row spacing.

Burden of significance.
Is the decimal required? The visual burden of high-significance data should be weighed against the desired degree of information delivery.

Table 2. Frequency of repeat pairing

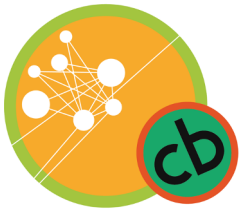
Amino acid (no. of proteins)	%	Self	Decreasing frequency (%, first column) of pairs (<5% not shown)																	
A (1117)	18.7	A	22.4	Q	15.4	G	12.6	S	7.8	P	5.0	N								
D (511)	15.3	D	20.9	N	12.0	E	10.2	Q	8.0	T	7.6	S	7.6	G	6.5	A				
E (681)	28.9	E	9.8	Q	9.6	N	9.0	D	8.1	S	6.3	P	6.0	A	5.6	K	5.5	T		
F (22)	13.6	F	22.7	N	13.6	K	9.1	D	9.1	L	9.1	P								
G (1126)	28.0	G	16.1	Q	15.2	A	11.3	S	7.6	P	5.8	N								
H (309)	8.9	H	21.5	Q	14.2	A	12.4	S	12.3	N	10.4	G	7.2	T	5.2	P				
K (382)	23.8	K	36.0	N	10.0	E	8.5	S	6.0	D										
L (96)	11.5	L	16.7	A	14.6	E	13.5	Q	9.4	P	7.3	G	6.2	S						
N (2098)	37.2	N	15.8	Q	12.6	T	9.9	S	6.6	K	5.1	D								
P (718)	28.3	P	13.0	Q	12.1	A	11.9	G	10.0	S	6.0	E								
Q (2172)	30.3	Q	15.2	N	11.5	A	11.0	S	8.8	T	8.8	G								
R (126)	15.1	R	14.3	G	11.1	P	11.1	E	10.3	S	8.3	A	7.1	T	5.2	D				
S (1460)	22.1	S	16.3	Q	14.3	N	10.8	T	9.7	A	8.7	G								
T (1064)	20.3	T	24.8	N	17.9	Q	14.8	S												
V (21)	28.6	V	19.0	S	15	A	9.5	P	9.5	R										

Incidental formatting.
Why is the number associated with each amino acid in brackets, rather than a column? If it's not important, it should be removed.

Misguided sight line.
The author wants us to visually scan along a row. This table's column spacing is larger than row spacing and therefore the eye scans down, not across.

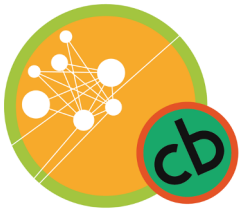
Source: http://circos.ca/presentations/articles/vis_tables1/img/bad-table-03.png

Matthews, Janice R., and Robert W. Matthews. *Successful scientific writing: a step-by-step guide for the biological and medical sciences*. Cambridge University Press, 2014.



Exercise:
Strength, or weakness?

1. The paper does not specify whether or not the hypothesis test is directional, making it difficult to determine whether or not the hypothesis testing is reasonably conservative.
2. The paper explains how the hypothesis test is the appropriate one for this experiment.



Watch for confounds

The results show that a small change in tokenization strategy can improve a mediocre 2006 TREC genomics submission (MAP average: 29%) to the top quarter of the submissions (36%-54%). Normalization and splitting compounds to multiple terms shows to be very beneficial for the tested IR models which assume term independence in both queries and documents. We expect that incorporation of proximities of related terms in the retrieval model will even further improve retrieval performance.

Trieschnigg, Dolf, Wessel Kraaij, and Franciska de Jong. "The influence of basic tokenization on biomedical document retrieval." *SIGIR* 2007.

Table 4. Effect of data balance, holding all other factors constant.

POSITIVE INSTANCES	NEGATIVE INSTANCES	*F*-MEASURE
100	100	0.82 ± 0.03
100	200	0.80 ± 0.03
100	300	0.74 ± 0.04
100	400	0.70 ± 0.04

Cohen, K. B., Glass, B., Greiner, H. M., Holland-Bouley, K., Standridge, S., Arya, R., ... Pestian, J., & Glauser, T. (2016). Methodological issues in predicting pediatric epilepsy surgery candidates. *Biomedical Informatics Insights*.

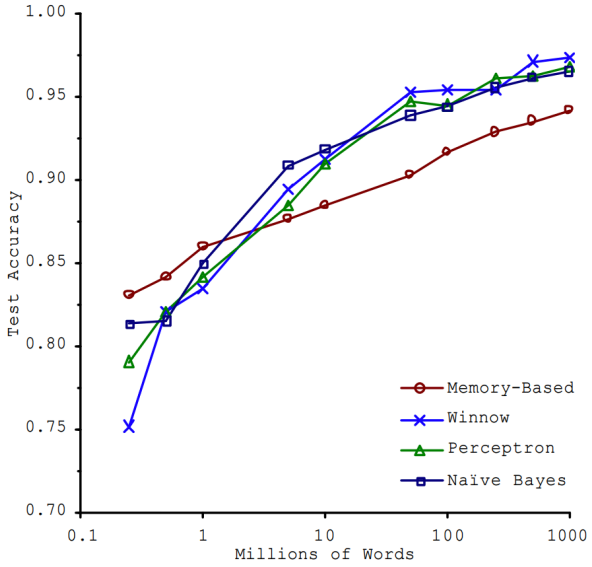
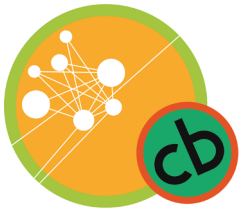


Figure 1. Learning Curves for Confusion Set Disambiguation

Banko, Michele, and Eric Brill. "Scaling to very very large corpora for natural language disambiguation." ACL 2001.



Watch for poorly-designed experiments

Bad ablation study

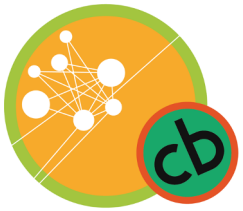
Feature set	F_1
Word embeddings	0.90
Word embeddings + Bag of words	0.90

Good ablation study

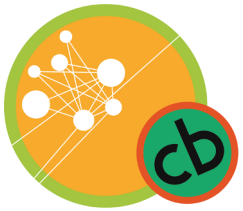
Feature set	F_1
Word embeddings	0.90
Bag of words	0.90
Word embeddings + Bag of words	0.90

Watch for over-optimizing

Thanks to TJ Callahan

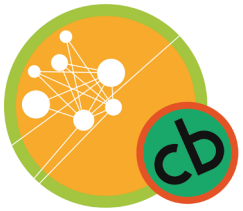


WHAT HAVEN'T I TALKED ABOUT—THAT YOU THOUGHT I WOULD?



Most frequent definition of "positive result" in machine learning: system under study out-performs some baseline/other system

PUBLICATION BIAS: TENDENCY TO PUBLISH PAPERS WITH "POSITIVE" RESULTS



Publication bias distorts the true state of the field

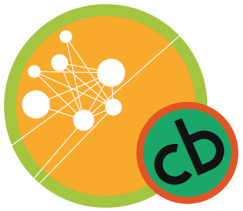
Publication bias:

- 25 papers report that deep learning out-performed any other approach
- 0 papers report that deep learning did not out-perform any other approach

No publication bias:

- 25 papers report that deep learning out-performed every other approach
- 25 papers reported that deep learning did not out-perform any other approach
- 50 papers report that deep learning out-performed some approaches, but not others

REVIEWS OF RESULTS SECTIONS ARE WHERE PUBLICATION BIAS PLAYS OUT



Two mechanisms of publication bias in data science

Rejecting papers that do not have state-o—the-art results

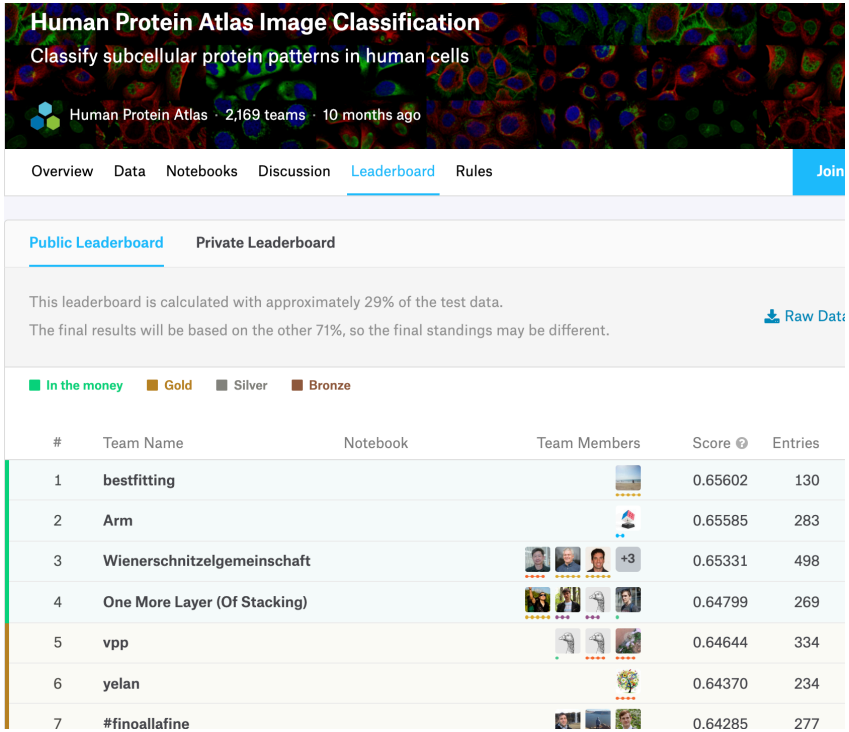
…even when they have interesting findings

Accepting papers that report state-of-the-art results

…even when the paper itself is otherwise deficient

To avoid publication bias, value validity over performance

Leaderboard model



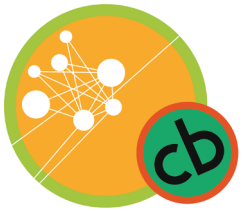
Smarter people than me have given this example—Ken Church, Christopher Manning, Bonnie Webber?

Deresultified model

Table 5. **Classification using Gene Ontology concepts** Five-fold cross validation performance of five binary classifiers when providing Gene Ontology concepts as features. Results from both unbalanced and balanced training sets are shown. The highest F-measure is bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

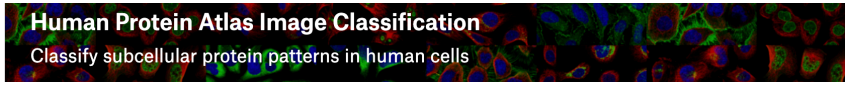
Classifier	GOA curated P/R/F	NLP abstracts P/R/F	NLP full-text P/R/F
Unbalanced Training			
Random			
OneR			
Naive Bayes			
Random Forest			
SMO			
LibSVM			
Balanced Training			
Random			
OneR			
Naive Bayes			
Random Forest			
SMO			
LibSVM			

Funk, Christopher S., Lawrence E. Hunter, and K. Bretonnel Cohen. "Combining heterogenous data for prediction of disease related and pharmacogenes." In *Pac. Symp. Biocomp. 2014*, pp. 328-339.



To avoid publication bias, value validity over performance

Leaderboard model



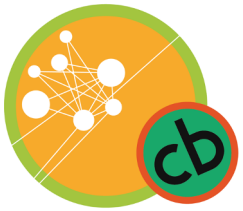
Deresultified model

Table 5. **Classification using Gene Ontology concepts**
Five-fold cross validation performance of five binary classi-

Button, Katherine S., Liz Bal, Anna Clark, and Tim Shipley. "Preventing the ends from justifying the means: withholding results to address publication bias in peer-review." (2016): 59.

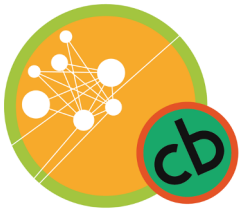
Smarter people than me have given this example—Ken Church, Christopher Manning, Bonnie Webber?

Funk, Christopher S., Lawrence E. Hunter, and K. Bretonnel Cohen. "Combining heterogenous data for prediction of disease related and pharmacogenes." In *Pac. Symp. Biocomp. 2014*, pp. 328-339.



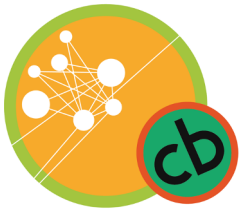
~~"Does not report state-of-the-art performance"~~

NEVER WRITE THIS IN A REVIEW!



Source: Kenneth Church

THE BETTER THE NUMBERS ARE, THE MORE IMPORTANT IT IS TO REJECT THE PAPER. WE CAN'T AFFORD PAPERS THAT REPORT RESULTS WITHOUT INSIGHTS.



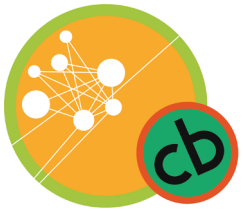
Exercise:
Ethics and the *Results* section

1.7.21 The current version of the Association for Computing Machinery Code of Ethics and Professional Conduct contains the following in its list of professional responsibilities: *2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.*

(a) If a research project involves many experiments with widely varying levels of performance, but the paper only shows the baseline and the best run, is that "comprehensive and thorough?"

(b) What would be some effective ways of showing the full range of performance levels?

Cohen (2021), *Writing about data science research.* Cambridge University Press.

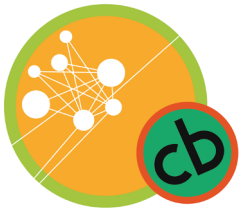


Exercise: Thinking about baselines

In the following extract from Resnik and Lin, what is meant by "independent of prior work in the literature," "reality check," and "fundamentally"?

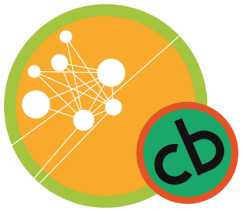
> One category of baselines can be defined independently of prior work in the literature; for example, choosing an answer at random, or always selecting an item's most frequent label in the training data, or applying something else that is equally obvious and simple. . . Generally [this] category [of baseline] can be viewed as a "reality check": if you cannot beat one of these baselines, most likely something is fundamentally wrong in your approach, or the problem is poorly defined.

Source: Cohen (2021), *Writing about data science research.* Cambridge University Press. Resnik and Lin is: Resnik, Philip, and Jimmy Lin (2010), "Evaluation of NLP Systems." In *The handbook of computational linguistics and natural language processing*.



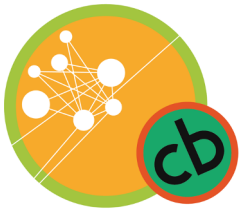
A checklist for the *Results* section

- ❑ Metric is appropriate for the task
- ❑ Baselines are non-trivial
- ❑ Measures of dispersion are reported
- ❑ Graphs:
  - ❑ Axes are not truncated
  - ❑ Axes are labelled
- ❑ Parameteric statistics used only with normally distributed data
- ❑ Error analysis is present
- ❑ Error analysis is non-trivial
- ❑ Alternative analyses are considered



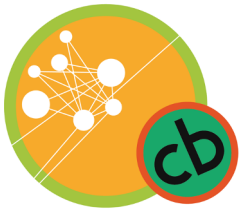
Publication bias

- Button, Katherine S., et al. "Preventing the ends from justifying the means: withholding results to address publication bias in peer-review." (2016) *BMC Psychology*.
- Church, Kenneth. ``Emerging trends: I did it, I did it, I did it, but..." Natural Language Engineering (2017).
- Cohen, Fort, Mieskes, and Néveol (2020) *How to review a paper about natural language processing*.
- Cohen, Kevin Bretonnel. Writing about data science research: With examples from machine learning and natural language processing. Cambridge University Press, summer/fall 2020. Basic principles of writing about the topics of most *ACL papers.
- Cohen, Paul R. Empirical methods for artificial intelligence.} Vol. 139. Cambridge, MA: MIT press, 1995. Principles of experimental design in machine learning and in artificial intelligence. Pedersen, Ted. ``Empiricism is not a matter of faith." Computational Linguistics (2008). The relationship between availability of code and data, and importance of results.



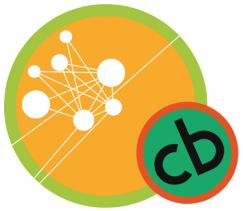
Why reviewers should not focus on performance

- Hand, David J. ``Classifier technology and the illusion of progress." Statistical science} (2006): 1-14. The dangers of overly complicated classifiers.
- Manning, Christopher D. ``Computational linguistics and deep learning." Computational Linguistics (2015). Problems of focusing on performance.
- Sculley, D. et al. ``Machine learning: The high interest credit card of technical debt." (2014). The dangers of sloppy machine learning work.
- D. Shatz. ”Is peer review conservative?" In *Peer review: A critical inquiry* (2004). The philosophy behind this issue.
- Steedman, Mark. ``On becoming a discipline." Computational Linguistics 34, no. 1 (2008): 137-144. The dangers of superficial evaluations and the importance of research that is not just about performance numbers.
- Wu, Stephen, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. "Negation’s not solved: generalizability versus optimizability in clinical natural language processing." PLoS ONE (2014). The danger of evaluation on a single dataset.
- Webber, Bonnie. ``Breaking news: Changing attitudes and practices." Computational Linguistics (2007). Problems of current reviewing practices in our field.



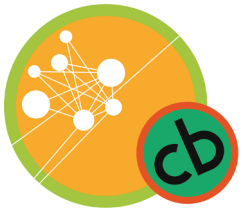
Statistical significance

- Berg-Kirkpatrick, Taylor, David Burkett, and Dan Klein. "An empirical investigation of statistical significance in NLP." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012.
- Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018, July). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1383-1392).
- Dror, R., Baumer, G., Bogomolov, M., & Reichart, R. (2017). Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, *5*, 471-486.
- Yeh, Alexander. "More accurate tests for the statistical significance of result differences." *Proceedings of the 18th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 2000.

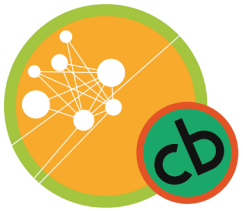


For in-depth discussion of all of the topics in this tutorial, see:

COHEN, FORT, NÉVÉOL, AND MIESKES (2020) HOW TO REVIEW A PAPER ABOUT NATURAL LANGUAGE PROCESSING



IF YOU HAD ALL THE ROOM IN THE WORLD, WHAT WOULD GO IN YOUR *RESULTS* SECTION?



IF YOU DON'T HAVE ALL THE ROOM IN THE WORLD, HOW SHOULD YOU DECIDE WHAT TO LEAVE OUT OF YOUR *RESULTS* SECTION?

Karën Fort, Margot Mieskes, and Aurélie Névéol

Kevin Bretonnel Cohen,
UCSOM

Karën Fort,
Sorbonne
Université / Loria

Margot Mieskes,
h_da Darmstadt

Aurélie Névéol,
Université
Paris Saclay,
CNRS, LIMSI