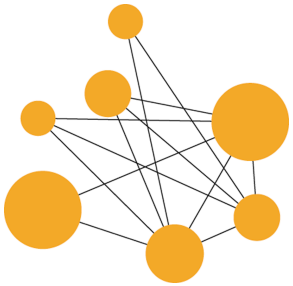
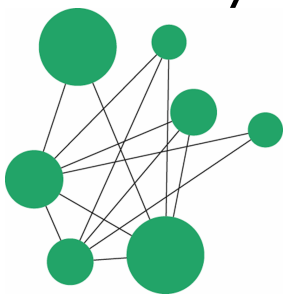

# Tutorial:
# Reviewing Methods sections

Kevin Bretonnel Cohen
Director, Biomedical Text Mining Group,
University of Colorado School of Medicine;
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

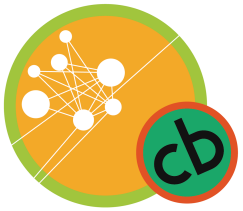
# Karën Fort, Margot Mieskes, and Aurélie Névéol


Kevin Bretonnel Cohen, UCSOM

Karën Fort, Sorbonne Université / Loria

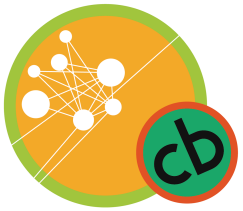
Margot Mieskes, h_da Darmstadt

Aurélie Névéol, Université Paris Saclay, CNRS, LIMSI




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

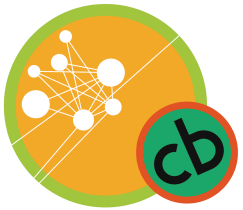
# COHEN, FORT, NÉVÉOL, AND MIESKES (2020) HOW TO REVIEW A PAPER ABOUT NATURAL LANGUAGE PROCESSING




We are talking about the Materials and Methods section because…

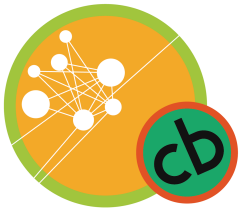
# THE METHODS ARE HOW THE PAPER LOOKS FOR THE ANSWER TO ITS QUESTION

# PRIMARY REVIEWING QUESTION: COULD THE METHODOLOGY ANSWER THE QUESTION POSED IN THE PAPER?



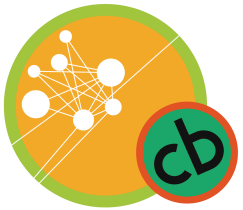

# (Partial!) checklist for the Methods section

- ❑ No confounds
- ❑ No confirmation bias
- ❑ Complete documentation
- ❑ Availability of code and data
- ❑ Manipulation experiments



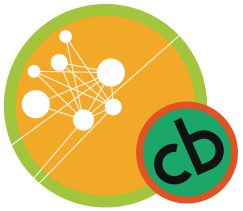

There are **many** kinds of papers, and you review them differently—see Karën Fort's lecture in this tutorial

# I AM FOCUSING HERE ON PAPERS THAT ARE BASED ON A PERFORMANCE CLAIM



# What *is* the methodology, **exactly?**

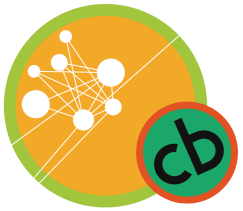
- ***In order to know whether or not a procedure could answer an experimental question, we must first agree on what the procedure actually was.***



# A good Methods section explains **why** these methods were used

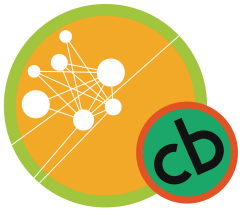
As a first step, we perform sentence alignment of the corpus. This step is necessary **because** there is never a full one-to-one correspondence between the sentences of two parallel documents.

Deléger, L., Merabti, T., Lecrocq, T., Joubert, M., Zweigenbaum, P., & Darmoni, S. (2010). A twofold strategy for translating a medical terminology into French. In *AMIA Annual Symposium Proceedings* (Vol. 2010, p. 152). American Medical Informatics Association.



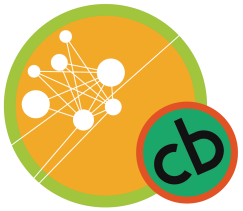
# A good Methods section explains **why** these methods were used

This paper takes a [method] approach **because** [characteristics of the task/data/use case/etc.].



# The Methods section sets up the Results section

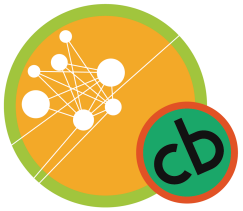
| If you see... | Then look for... |
| --- | --- |
| • Experiment: how dependent on the size of the dataset? | • A growth curve |
| • Experiment: how much effect does tokenization have? | • A table showing performance across tokenizers |
| • Experiment: how stable/variable are the results across folds? | • Plot of folds and/or standard deviation |



# A good Methods section tells you enough to repeat the experiments…

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A link to a downloadable source code, with specification of all dependencies, including external libraries
- ❑ Description of computing infrastructure used
- ❑ Average runtime for each approach
- ❑ Number of parameters in each model
- ❑ Corresponding validation performance for each reported test result
- ❑ Explanation of evaluation metrics used, with links to code

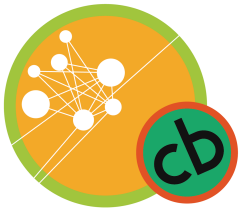
Source: https://2020.emnlp.org/call-for-papers




# …to repeat the parameter space search…

- Bounds for each hyperparameter
- Hyperparameter configurations for best-performing models
- Number of hyperparameter search trials
- The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)
- Expected validation performance, as introduced in Section 3.1 in * Dodge et al, 2019, or another measure of the mean and variance as a function of the number of hyperparameter trials.

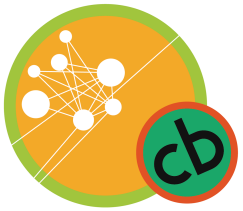
Source: https://2020.emnlp.org/call-for-papers




# …and for all datasets used:

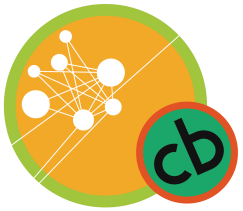
- Relevant statistics such as number of examples
- Details of train/validation/test splits
- Explanation of any data that were excluded, and all pre-processing steps
- A link to a downloadable version of the data
- For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

Source: https://2020.emnlp.org/call-for-papers




# Exercise:
# What is missing?

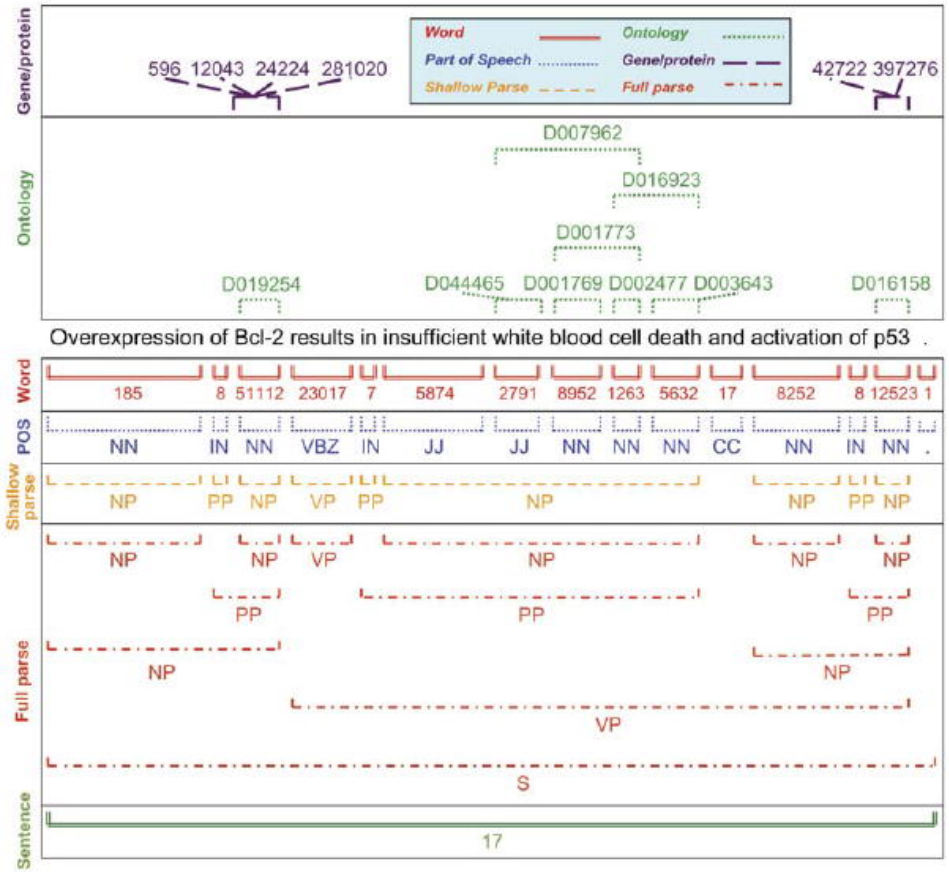
- The transcripts were tokenized, part-of-speech tagged, and subjected to shallow parsing to mark the boundaries of noun phrases.
- The texts were preprocessed.



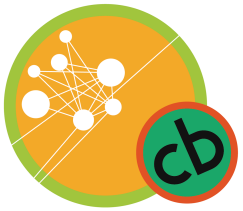
# Exercise:
# Which is better?

1. The training and test files of the corpus were sentence split, tokenized, part-of-speech tagged and lemmatized with the TTL tool (Ion, 2007). The corpus was automatically syntactically annotated with UDPipe based on the romanian-ud-ro2.0-170801.udpipe. The format of the corpus is *cupt* (Ramisch et al., 2018). https://www.aclweb.org/anthology/W19-5103.pdf
2. The concatenated corpus was tokenized, Part-of-Speech-tagged and lemmatized using the TreeTagger (www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.192.5438&rep=rep1&type=pdf
3. The text was tokenized, part-of-speech tagged and stemmed.

# Exercise: How many things need to be documented in this Methods section?



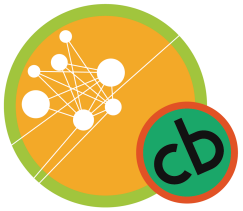

Source: Preslav Nakov and Marti Hearst




# Look for recognizable experimental methodologies

- Exploratory studies
- Assessment studies
- Manipulation experiments

Cohen, Paul R. *Empirical methods for artificial intelligence*. MIT Press, 1995. *Over 1,000 citations as of June 2020!*



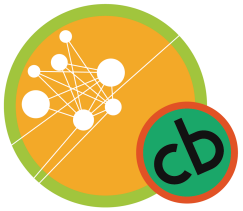

# Exercise:
# Find the confounds

## Experiment & Results

- **System 1:** Text is tokenized with NLTK Version 3.0's GENIA model, stemmed with the Tartarus implementation of the Porter stemmer, and documents are classified with a MALLET random forest
- **System 2:** Documents are classified with a MALLET support vector machine
- **Finding:** (2) outperformed (1)

## Conclusion

- Support vector machines outperform random forests



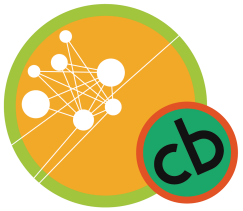

# Exercise:
# Find the confounds

## Experiment & Results

- **System 1:** Text is tokenized with NLTK Version 3.0's GENIA model, stemmed with the Tartarus implementation of the Porter stemmer, and documents are classified with a MALLET random forest
- **System 2:** Documents are classified with a MALLET support vector machine
- **Finding:** (1) outperformed (2)

## Conclusion

- Random forests outperform support vector machines



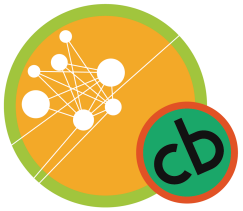
# Watch for confirmation bias

"**Confirmation bias** is the tendency to search for, interpret, favor, and recall information that confirms or supports one's prior personal beliefs or values.[1]"

**Common signs:**

- Results are uniformly positive
- Results are unsurprising
- Results are exactly what the Introduction would predict

https://en.wikipedia.org/w/index.php?title=Confirmation_bias&oldid=963946258

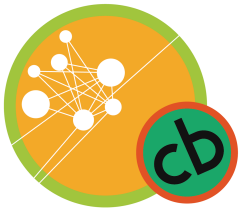



# Watch for YOUR OWN confirmation bias!

The social process of peer review is thought to mitigate the effect of individual scientists' biases, even though the peer review process itself may be susceptible to such biases [95][96][89][97][98] Confirmation bias may thus be especially harmful to objective evaluations regarding nonconforming results since biased individuals may regard opposing evidence to be weak in principle and give little serious thought to revising their beliefs.[88] Scientific innovators often meet with resistance from the scientific community, and research presenting controversial results frequently receives harsh peer review.[99]

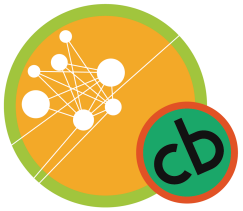
https://en.wikipedia.org/w/index.php?title=Confirmation_bias&oldid=963946258

# SCORING THE METHODS SECTION



# X is better than Y

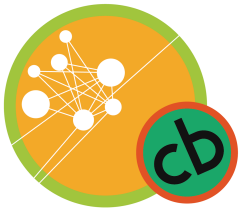
| X | Y |
|---|---|
| Simple | Complex |
| Available | Unavailable |
| Fast | Slow |
| Multiple languages/genres/… | Single language/genre/… |
| Multiple metrics | Single metric |
| Blind test set | Cross-validation |
| | |




Hand, David J. "Classifier technology and the illusion of progress." *Statistical Science* (2006): 1-14.

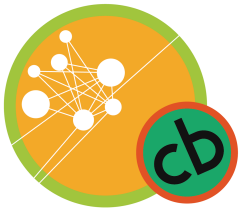
Sculley, David, et al. "Machine learning: The high interest credit card of technical debt." (2014).

# SIMPLE IS BETTER THAN COMPLEX



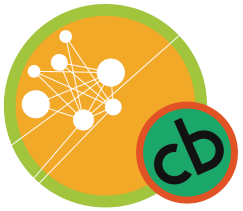

Pedersen, Ted. "Empiricism is not a matter of faith." *Computational Linguistics* 34.3 (2008): 465-470.

# AVAILABLE IS BETTER THAN UNAVAILABLE



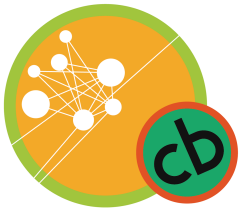

# A good Methods section makes the following available:

- ❑ Code for the system
- ❑ Code for running the experiments
- ❑ Intermediate results
- ❑ Code for the analysis



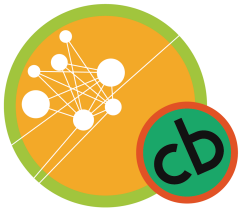

For “State Of The Art” papers, …

# FAST IS BETTER THAN SLOW




# Watch for confirmation bias

- Unsurprising findings, e.g. paper on a popular approach finds that the approach works well
- A paper that sets out to get state-of-the-art performance reports state-of-the-art performance
- A paper that reports only the best set of results



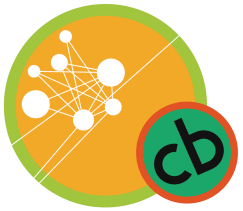
# Exercise: Which paper should get a higher score for its methodology?

**Paper #1**

The information extraction pipeline consists of a convolutional neural network followed by a recurrent neural network with bidirectional long short-term memory and optimization across a wide range of number of layers and nodes per layer.

**Paper #2**

The information extraction pipeline consists of a recurrent neural network with bidirectional long short-term memory and the same number of layers and nodes per layer that were used in the related work.



# Exercise:
# Which paper should get a higher score for its methodology?

**Paper #1**

The information extraction pipeline consists of a

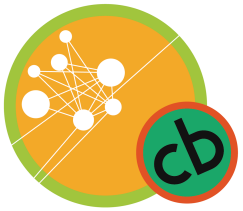
range of number of layers and nodes per layer.

**Paper #2**

The information extraction pipeline consists of a recurrent

work.

Simple is better than complicated, so Paper #2 should get a higher score for methodology than Paper #1.



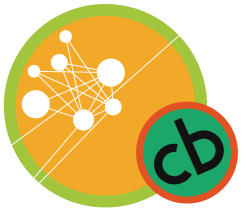

# Exercise: Which paper should get a higher score for methodology?

**Paper #1**

The pipeline consists of a convolutional neural network followed by a recurrent neural network. The pipeline code and libraries are available at [ANONYMIZED FOR SUBMISSION].

**Paper #2**

The pipeline consists of a convolutional neural network followed by a recurrent neural network. The pipeline code and libraries are available by request from the authors.




# Exercise: Which paper should get a higher score for methodology?

**Paper #1**

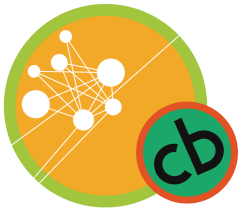
The pipeline consists of a convolutional neural network

**Paper #2**

The pipeline consists of a convolutional neural network

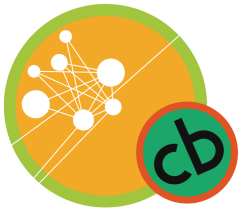
Available is better than unavailable, and a public repository is more available than "available on request," so Paper #1 should get a higher score for methodology than Paper #2.

Pedersen, Ted. "Empiricism is not a matter of faith." *Computational Linguistics* 34.3 (2008): 465-470.



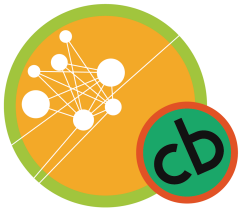
# X is better than Y

| X | Y |
|---|---|
| Simple | Complex |
| Available | Unavailable |
| Fast | Slow |
| Multiple languages/genres/… | Single language/genre/… |
| Multiple metrics | Single metric |
| Blind test set | Cross-validation |
| Multiple tasks | Single task |



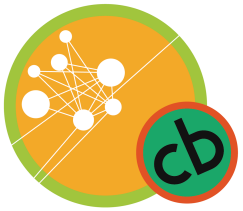
# Exercise: Which methods section should get a higher score? "To approach the problem, we…"

| Paper #1 | Paper #2 |
| --- | --- |
| …used data from the English, French, Chinese, and Russian sections of the OntoNotes and of the WMT 2020 corpora. A naïve Bayes classifier was used (all code and parameter settings are available on GitHub at [ANONYMIZED]). Performance was measured against a blind test set (see GitHub for the sampling procedure) using 5 standard metrics (see Table 3). | …used data from the English section of the OntoNotes corpus. A convolutional neural network was followed by a recurrent neural network with bidirectional long short-term memory and attention. Performance was measured by cross-validation. After random splitting into folds, the folds were post-hoc balanced for positive and negative exemplars. Table 3 reports the mean F-measure. |



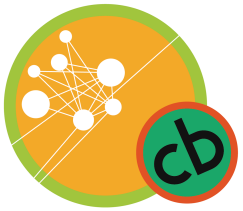

# What you need to document in natural language processing (and why)


- Kevin Bretonnel Cohen, *Writing about data science research: With examples from machine learning and natural language processing* (2021), Cambridge University Press
- Dodge, Jesse, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. "Show your work: Improved reporting of experimental results." (2019).
- Fokkens, Antske, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. "Offspring from reproduction problems: What replication failure teaches us." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1691-1701. 2013.
- Pedersen, Ted. "Empiricism is not a matter of faith." *Computational Linguistics* 34.3 (2008): 465-470.
- Joelle Pineau, https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

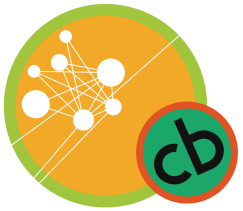



Source: Kenneth Church

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




# INSIGHTS TYPICALLY COME FROM GOOD METHODOLOGY AND FROM ERROR ANALYSIS




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

# 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