CS11-711 Advanced NLP

NLP Experimental Design

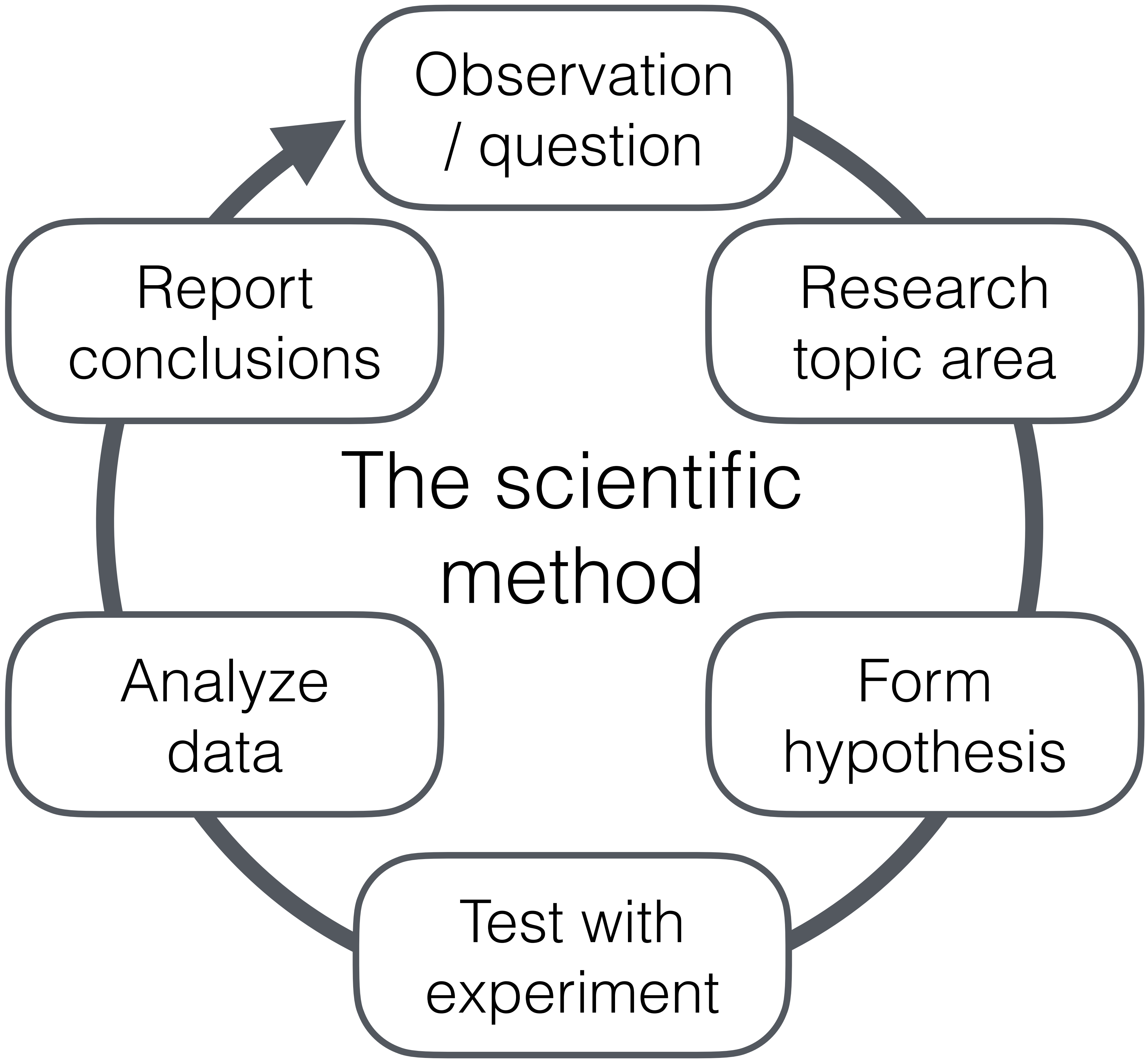
Sean Welleck

Carnegie Mellon University
Language Technologies Institute

https://cmu-l3.github.io/anlp-spring2025/

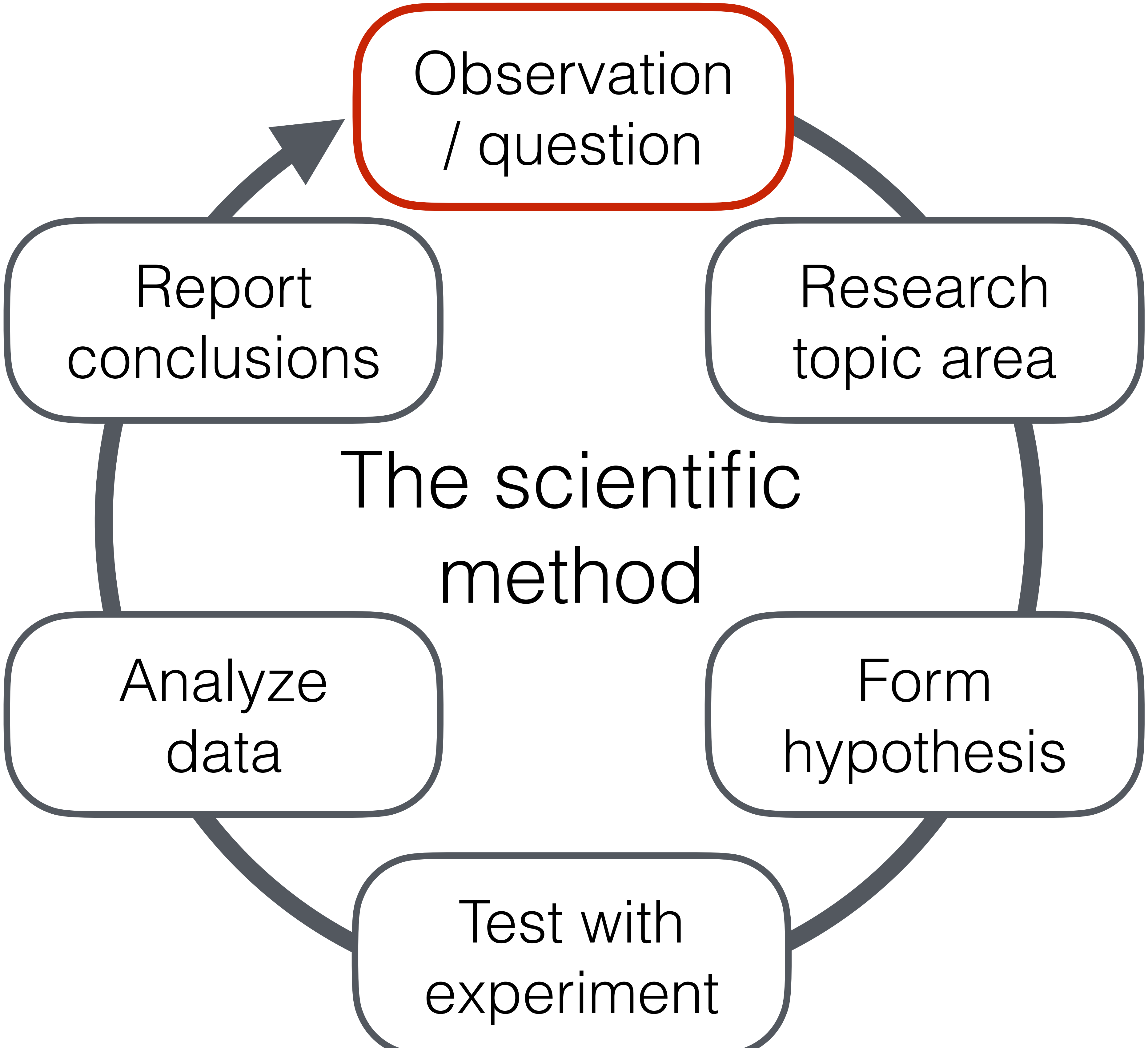
Most slides from Graham Neubig from Fall 2024

Acknowledgements on Graham's slides: thanks to Shaily Bhatt, Jordan Boyd-Graber, Joe Brucker, Hal Daume, Derguene Mbaye, Rajaswa Patil for content suggestions included here



Credit: Adapted From Wikipedia (Efbrazil)

Identifying Good Research Directions



Why Do We Research?

- **Applications-driven Research:** I would like to make a useful system, or make one work better.
- **Curiosity-driven Research:** I would like to know more about language, or the world viewed through language.
- NLP encompasses both, sometimes in the same paper

Examples of Application-driven Research

- Pang et al. (2002) propose a task of *sentiment analysis*, because "labeling these articles with their sentiment would provide succinct summaries to readers".
- Reddy et al. (2019) propose a task of *conversational question answering* because "an inability to build and maintain common ground is part of why virtual assistants usually don't seem like competent conversational partners."
- Gehrmann et al. (2018) propose a method of *bottom-up abstractive summarization* because "NN-based methods for abstractive summarization produce outputs that are fluent but perform poorly at content selection."
- Kudo and Richardson (2018) propose a *method for unsupervised word segmentation* because "language-dependent processing makes it hard to train multilingual models, as we have to carefully manage the configurations of pre- and post-processors per language."

Examples of Curiosity-Driven Research

- Rankin et al. (2017) ask what is the *difference between the language of real news with that of satire, hoaxes, and propaganda*?

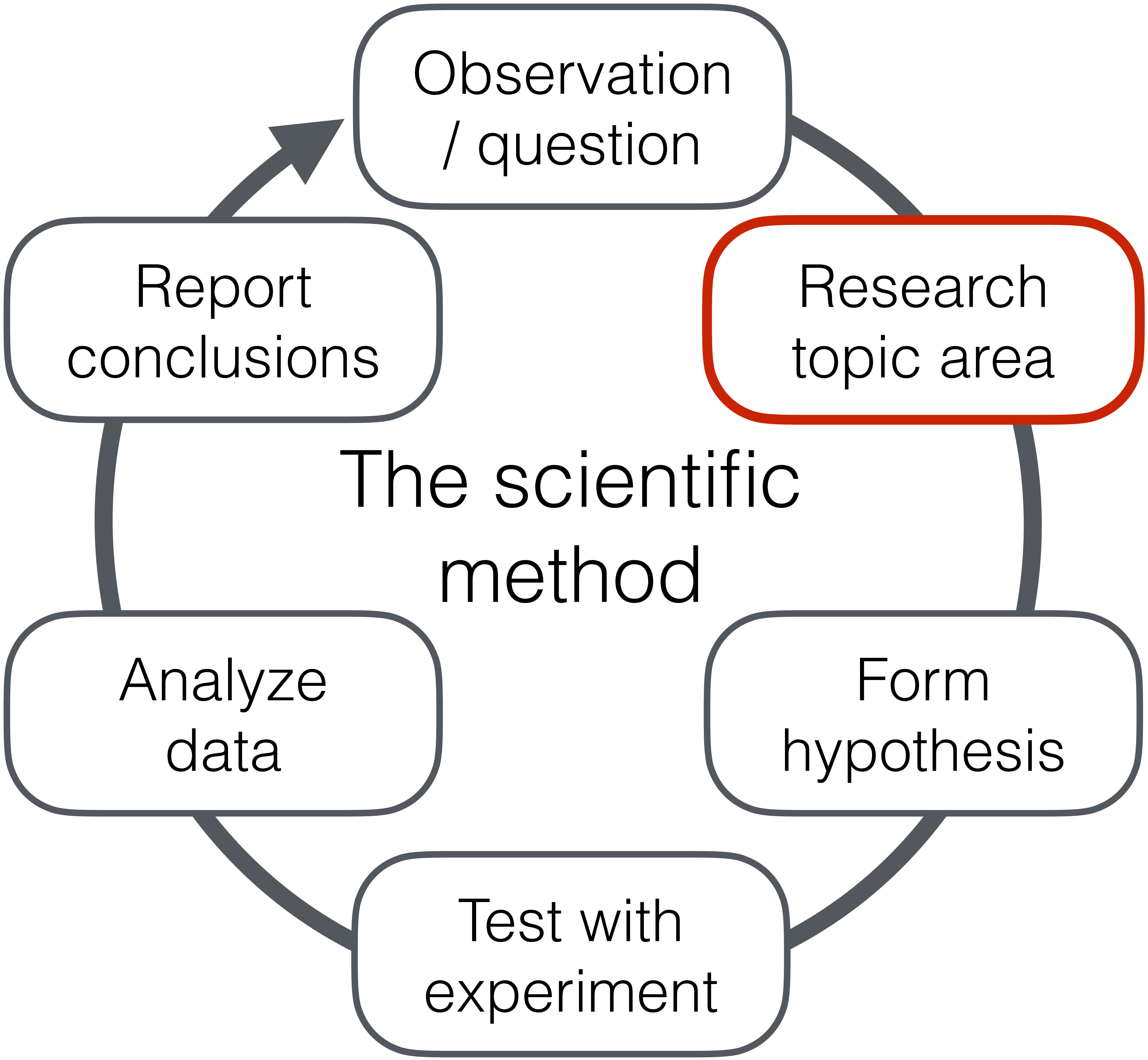
- Cotterell et al. (2018) ask "*are all languages equally hard to language model*?"

- Tenney et al. (2019) quantify *where specific types of linguistic information are encoded in BERT*.

How Do We Get Research Ideas?

- Turn a concrete understanding of existing research's failings to a higher-level experimental question.
  - **Bottom-up Discovery** of research ideas
  - Great tool for incremental progress, but may preclude larger leaps
- Move from a higher-level question to a lower-level concrete testing of that question.
  - **Top-down Design** of research ideas
  - Favors bigger ideas, but can be disconnected from reality
    - Solving a problem that is not actually a problem
    - Using a method that doesn't actually fit because you chose the method beforehand

Identifying Good Research Directions



Research Survey Methods

- **Keyword search**
- Find **older/newer papers**
- Read **abstract/intro**
- Read **details of most relevant papers**
- [Make a short summary?]

Some Sources of Papers in NLP

https://arxiv.org/

https://scholar.google.com/

https://openreview.net/

- NeurIPS*: https://neurips.cc/
- ICLR*: https://iclr.cc/
- COLM*: https://colmweb.org
- TMLR*: https://jmlr.org/tmlr/
- ICML: https://icml.cc/
- ACL/NAACL/EMNLP/etc.: https://aclanthology.org/

*Reviews available on OpenReview

ACL Anthology

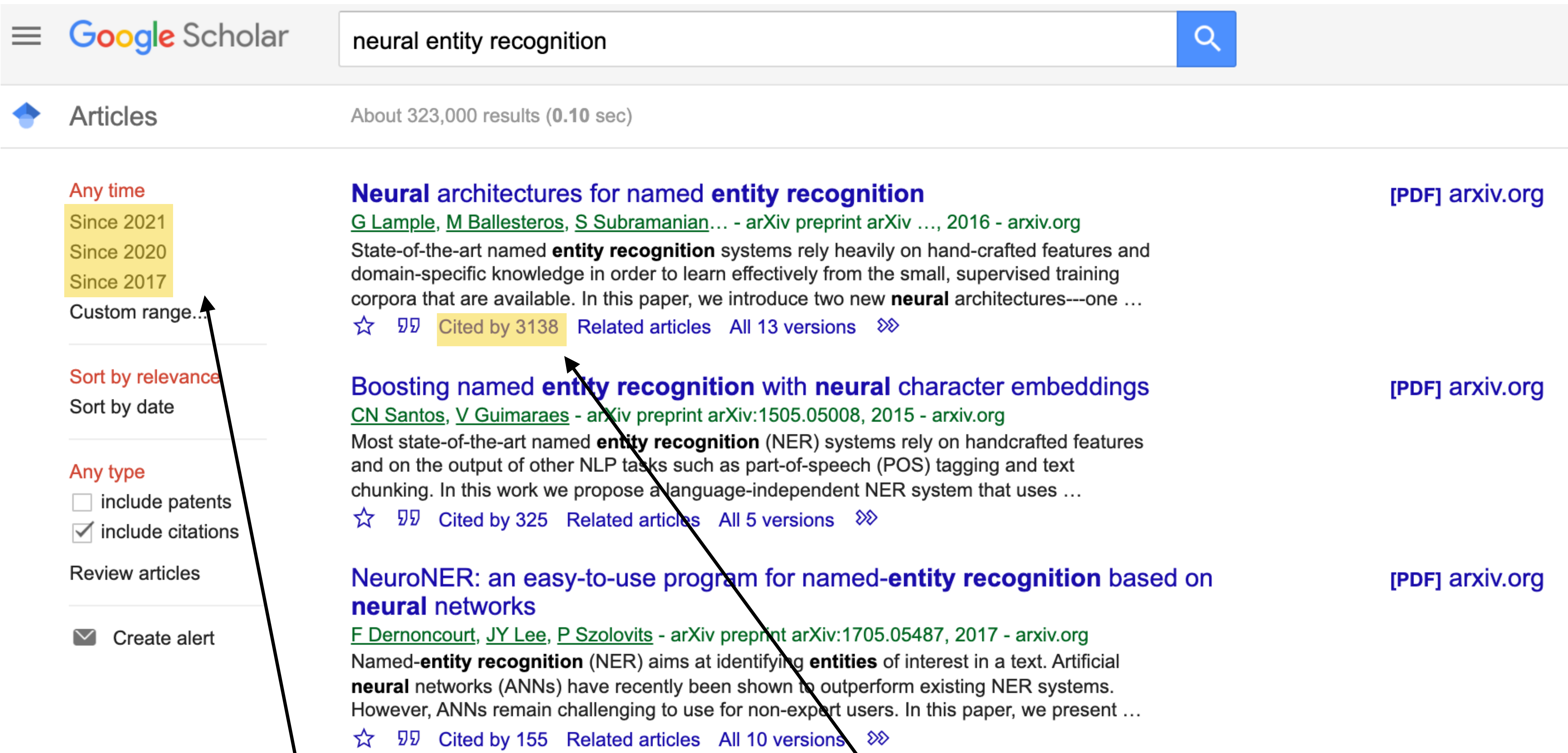
- Covers many prestigious venues in NLP
- Start with past 3-5 years of several top venues (e.g. ACL, EMNLP, NAACL, TACL)

ACL Events

Venue	2021 – 2020	2019 – 2010	2009 – 2000	1999 – 1990
AACL	20			
ACL	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90
ANLP			00	97 94 92
CL	20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90
CoNLL	20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97
EACL	21	17 14 12	09 06 03	99 97 95 93 91
EMNLP	20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96
Findings	21 20			
NAACL	21	19 18 16 15 13 12 10	09 07 06 04 03 01 00	
SemEval	21 20	19 18 17 16 15 14 13 12 10	07 04 01	98
*SEM	21 20	19 18 17 16 15 14 13 12		
TACL	21 20	19 18 17 16 15 14 13		
WMT	20	19 18 17 16 15 14 13 12 11 10	09 08 07 06	
WS	20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90
SIGs	ANN | BIOMED | DAT | DIAL | EDU | EL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | S			

Google Scholar

- Allows for search of papers by keyword



View recent papers

View papers that cite this one

Finding Older Papers

- Often as simple as following references

References

Akbik, A.; Bergmann, T.; and Vollgraf, R. Pooled contextualized embeddings for named entity recognition.

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th COLING*, 1638–1649.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th ICML-Volume 70*, 233–242. JMLR. org.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv e-prints*.

Baluja, S., and Fischer, I. 2017. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint arXiv:1703.09387*.

Borthwick, A.; Sterling, J.; Agichtein, E.; and Grishman, R. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Sixth Workshop on Very Large Corpora*.

Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; and Liu, S. 2018. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on EMNLP*, 182–192.

Chen, L., and Moschitti, A. 2019. Transfer learning for sequence labeling using source model and target data.

Chiu, J. P., and Nichols, E. 2016. Named entity recognition with bidirectional lstm-cnns. *TACL* 4:357–370.

chinese word segmentation with bi-lstms. In *Proceedings of the 2018 Conference on EMNLP*, 4902–4908.

Manning, C. D. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, 171–189. Springer.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of NAACL*, volume 1, 2227–2237.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Reimers, N., and Gurevych, I. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Sang, E. F., and De Meulder, F. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

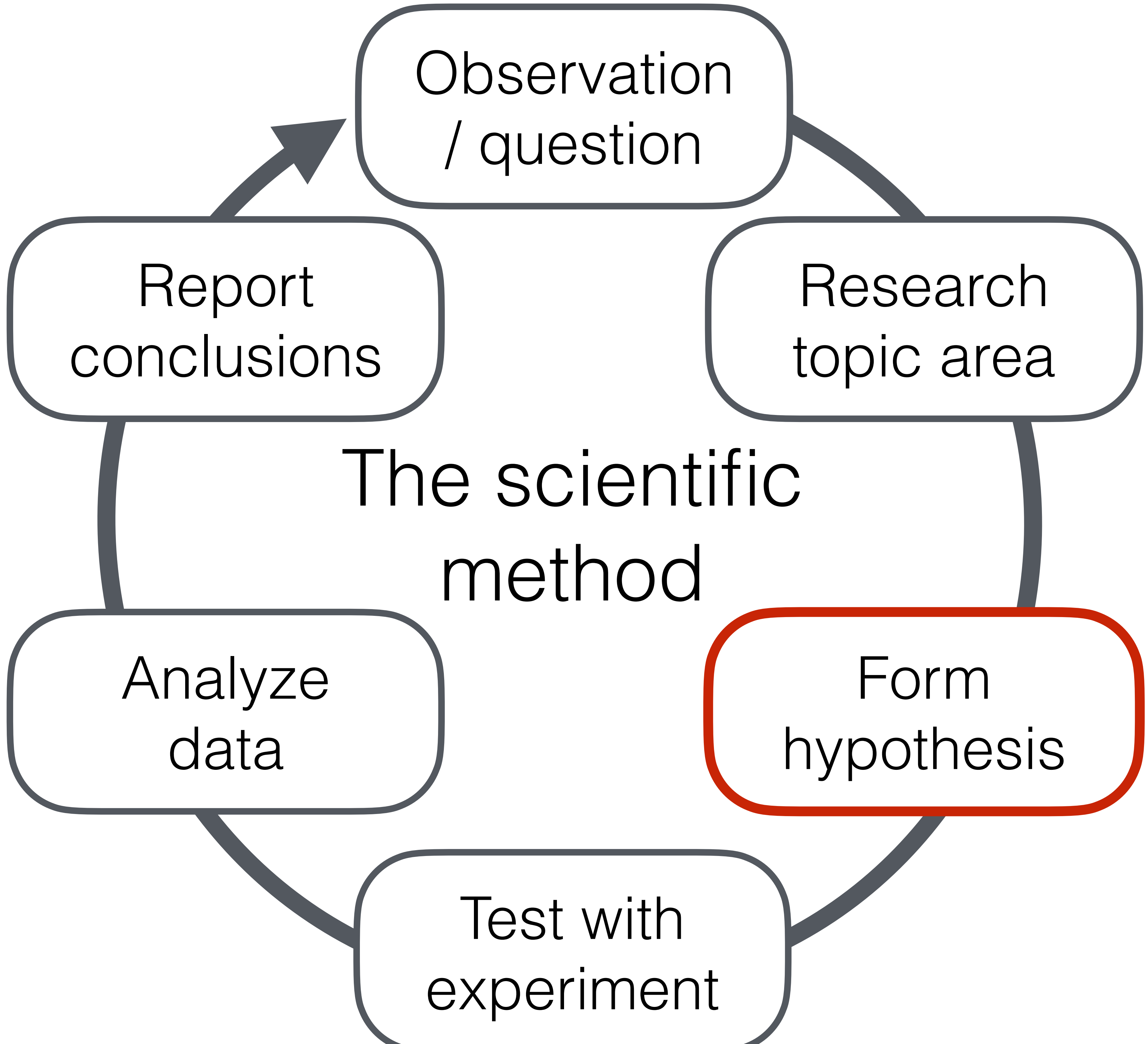
Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. In *Advances in NIPS*, 5014–5026.

Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. 2013. Ontonotes release 5.0 ldc2013t19. *LDC, Philadelphia, PA*.

The Ups and Downs of Pre-emptive Surveys

- Surveying extensively before doing research:
  - Prevents you from duplicating work
  - Increases your "toolbox" of methods
  - Constrains your thinking (see Varian 1994)

Identifying Good Research Directions



Devising Final Research Questions/Hypotheses

- **Research Question:**
  - One or several explicit questions regarding the thing that you want to know
  - "Yes-no" questions often better than "how to"
- **Hypothesis:**
  - What you think the answer to the question may be a-priori
  - Should be *falsifiable*: if you get a certain result the hypothesis will be validated, otherwise disproved

Curiosity-driven Questions + Hypotheses

Are All Languages Equally Hard to Language-Model?

Modern natural language processing practitioners strive to create modeling techniques that work well on all of the world's languages. Indeed, most methods are portable in the following sense: Given appropriately annotated data, they should, in principle, be trainable on any language. However, despite this crude cross-linguistic compatibility, it is unlikely that all languages are equally easy, or that our methods are equally good at all languages.

What makes a particular podcast broadly engaging? As a media form, podcasting is new enough that such questions are only beginning to be understood (Jones et al., 2021). Websites exist with advice on podcast production, including language-related tips such as reducing filler words and disfluencies, or incorporating emotion, but there has been little quantitative research into how aspects of language usage contribute to listener engagement.

Cotterell et al. (2018)

Reddy et al. (2018)

Application-driven Questions + Hypotheses

However, from these works, it is still not clear as to *when* we can expect pre-trained embeddings to be useful in NMT, or *why* they provide performance improvements. In this paper, we examine these questions more closely, conducting five sets of experiments to answer the following questions:

- Q1 Is the behavior of pre-training affected by language families and other linguistic features of source and target languages? (§3) — Yes?
- Q2 Do pre-trained embeddings help more when the size of the training data is small? (§4) — Yes?
- Q3 How much does the similarity of the source and target languages affect the efficacy of using pre-trained embeddings? (§5) — Not much?
- Q4 Is it helpful to align the embedding spaces between the source and target languages? (§6) — Yes?
- Q5 Do pre-trained embeddings help more in multilingual systems as compared to bilingual systems? (§7) — Unclear

Qi et al. (2018)

Although recent studies on ST have achieved promising results with end-to-end (E2E) models (Anastasopoulos and Chiang, 2018; Di Gangi et al., 2019; Zhang et al., 2020a; Wang et al., 2020; Dong et al., 2020), nevertheless, they mainly focus on sentence-level translation. One practical challenge when scaling up sentence-level E2E ST to the document-level is the encoding of very long audio segments, which can easily hit the computational bottleneck, especially with Transformers (Vaswani et al., 2017). So far, the research question of whether and how contextual information benefits E2E ST has received little attention.

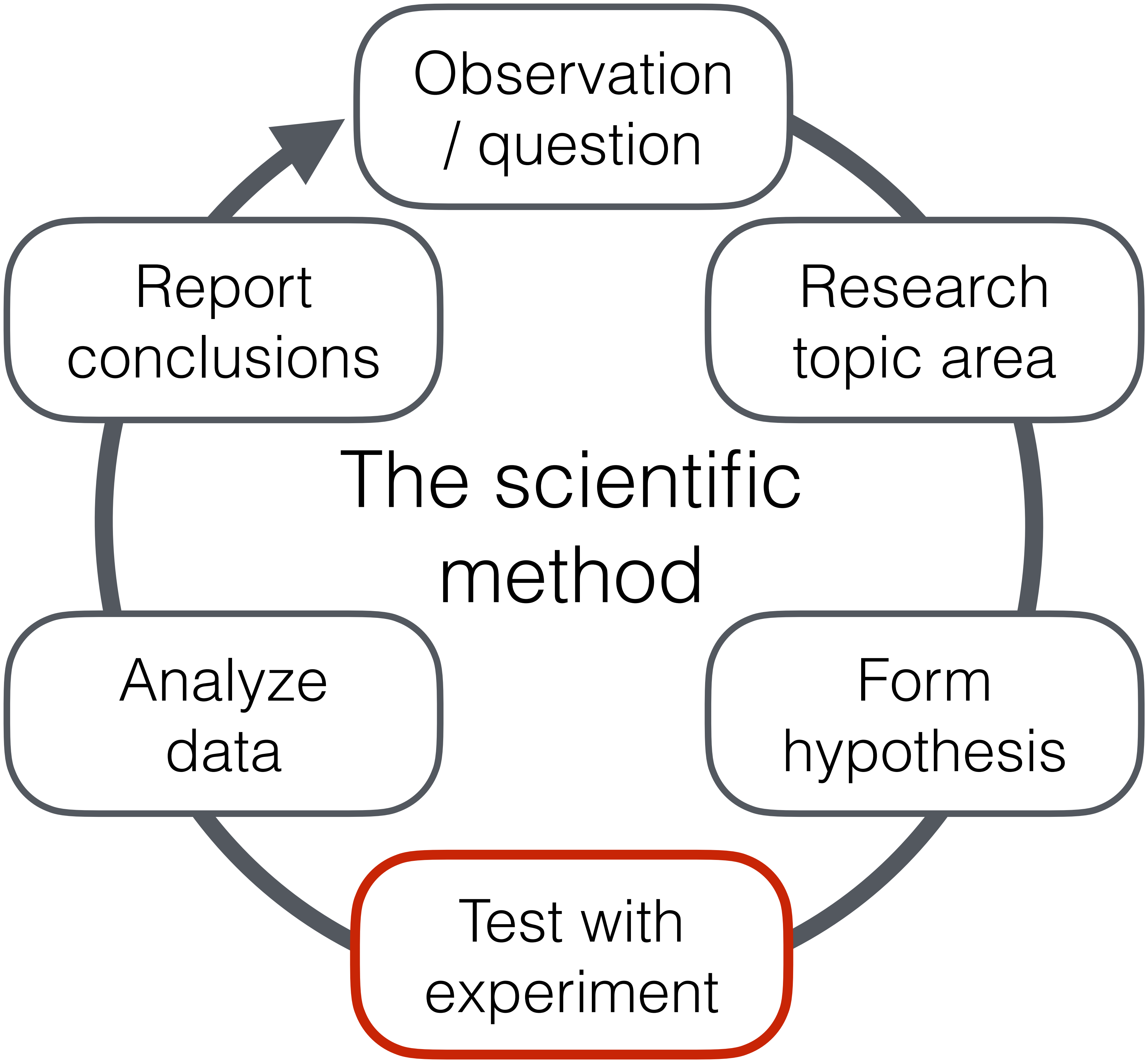
Probably will help?

Zhang et al. (2021)

Beware
"Does X Make Y Better?" "Yes"

- The above question/hypothesis is natural, but indirect
  - If the answer is "no" after your experiments, how do you tell what's going wrong?
- Usually you have an intuition about *why* X will make Y better (not just random)
- Can you think of other research questions/ hypotheses that confirm/falsify these assumptions

Performing Experiments



Running Experiments

- Find data that will help answer your research question
- Run experiments and calculate numbers
- Calculate significant differences and analyze effects

Obtaining Test Data

Finding Datasets

- If **building on previous work**, safest to start with same datasets
- If **answering a new question**
    - Can you repurpose other datasets to answer the question?
    - If not, you'll have to create your own

Dataset Lists

🤗 **Datasets**

https://github.com/huggingface/datasets

EUROPEAN LANGUAGE RESOURCES ASSOCIATION ELRA 2020 celebration 25 anniversary ELDA

http://www.elra.info/en/lrec/shared-lrs/

Papers With Code

https://paperswithcode.com/area/natural-language-processing

Annotating Data
(Tseng et al. 2020)

- Decide how much to annotate
- Sample appropriate data
- Create annotation guidelines
- Hire/supervise annotators
- Evaluate quality

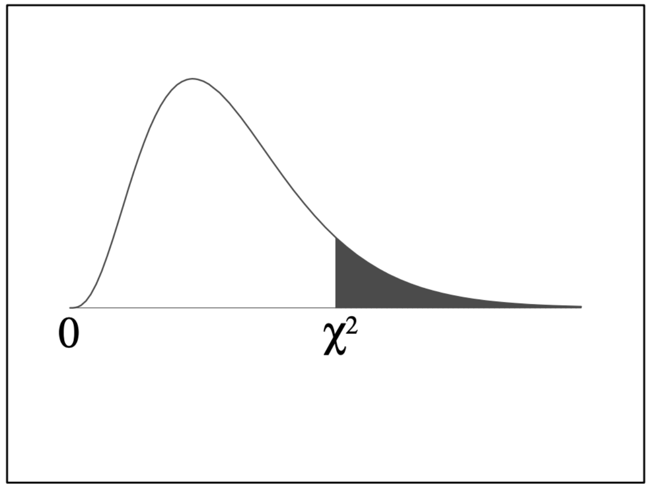
How Much Test/Dev Data Do I Need?

- Enough to have **statistically significant differences** (e.g. $p<0.05$) between methods
- How can I estimate how much is enough? **Power analysis** (see Card et al. 2020)
  - Make assumption about **effect size** between settings (e.g. expected accuracy difference between tested models)
  - Given effect size, significance threshold, determine how much data necessary to get significant effect in most trials

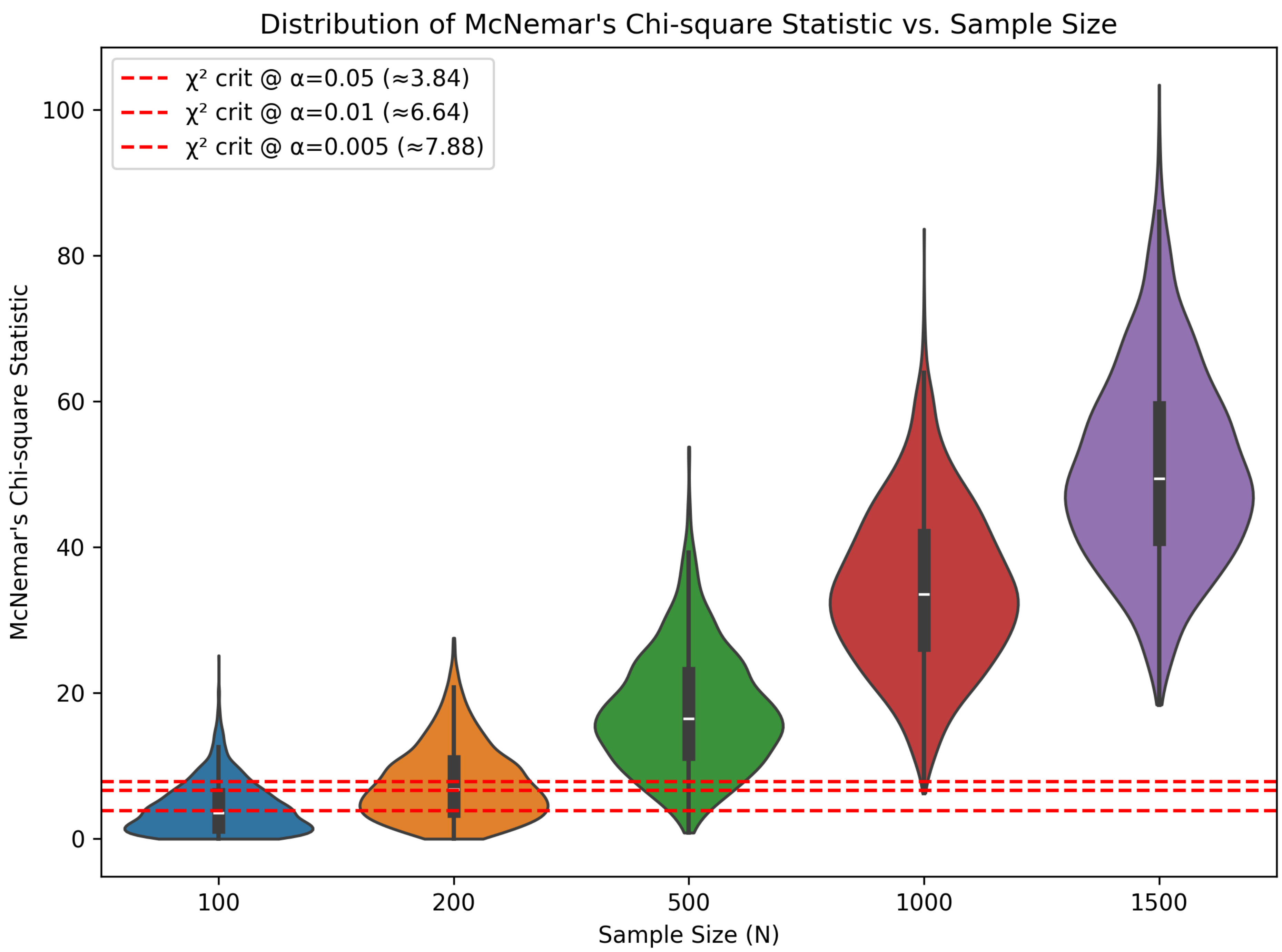
Power analysis: example

- Accuracy of M1: $\frac{40+10}{100} = 50\,\%$
- Accuracy of M2: $\frac{40+20}{100} = 60\,\%$
- McNemar's Test Statistic:
  - $\chi^2 = \frac{(10-20)^2}{10+20} \approx 3.333$
  - Based on how often M1 and M2 disagree on an item
  - $\chi^2_{0.05,1} \approx 3.841$
- Interpretation
  - Since 3.333 < 3.841, the difference in error patterns is not statistically significant at the 5% level.
- **Power:** The probability that the test will reject the null hypothesis (i.e., detect the difference) *if* the true difference exists. Typically, we aim for 80% or 90% power, which determines how large a sample we need.

	M2 Correct	M2 Incorrect	Total
M1 Correct	40	10	-
M1 Incorrect	20	30	-
Total	-	-	100

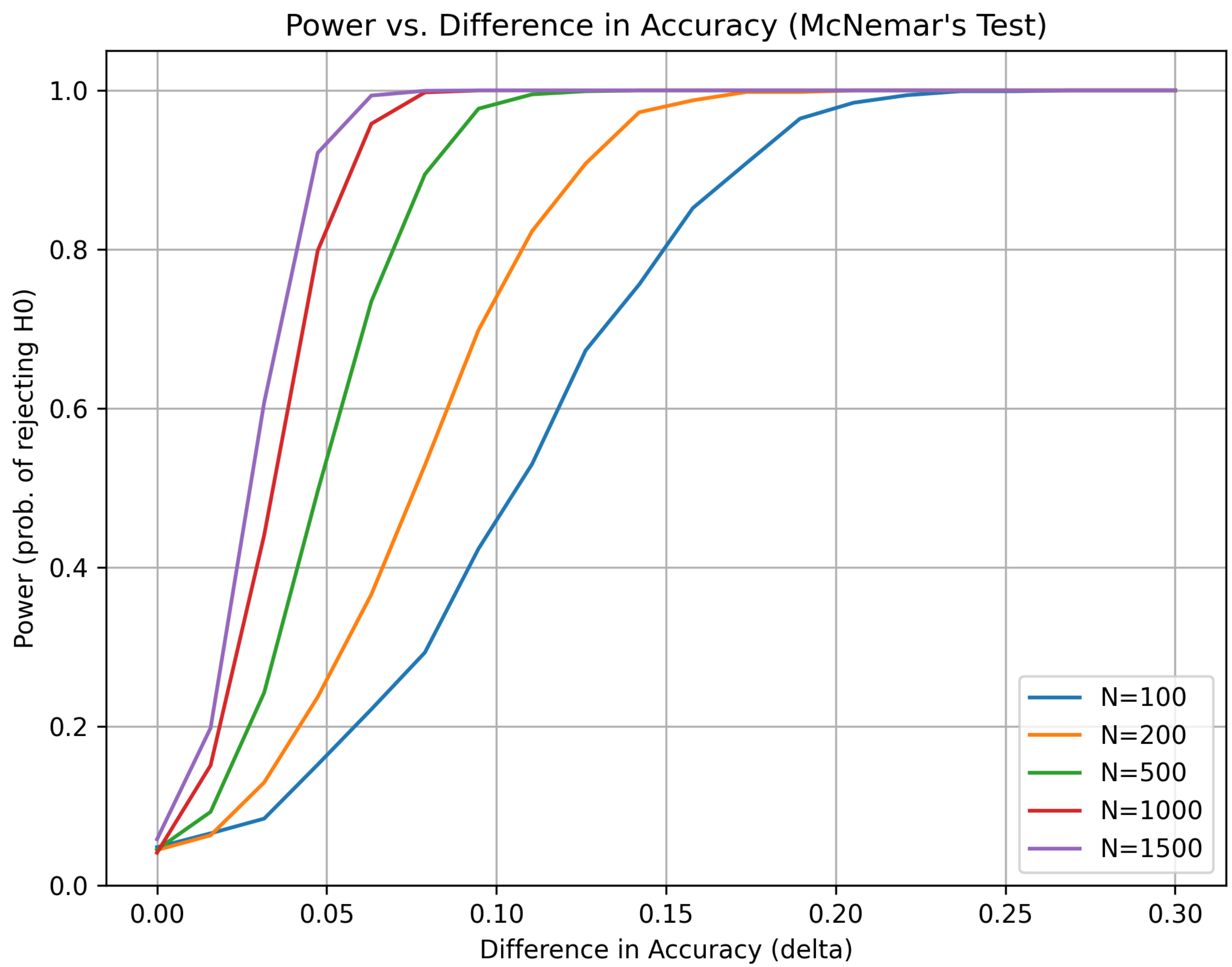


Power analysis: example



2,000 simulations

Power analysis: example



2,000 simulations

How Much Training Data Do I Need?

- More is usually better
- But recently reasonable perf. with few-shot, zero-shot transfer + pre-trained models (+prompting?)
- Can do even better with intelligent data selection - active learning

How Should I Sample Data?

- Coverage of the **domains** that you want to cover
- Coverage of the **language varieties, demographics** of users
- Documentation: **data statements for NLP** (Bender and Freidman 2018)

Curation Rationale	Speech Situation
Language Variety	Text Characteristics
Speaker Demographic	Recording Quality
Annotator Demographic	Other Comments

Annotation Guidelines

- Try to annotate yourself, create annotation guidelines, iterate.
- e.g. Penn Treebank POS annotation guidelines (Santorini 1990)

What:

2 LIST OF PARTS OF SPEECH WITH CORRESPONDING TAG 2

Adverb—RB

This category includes most words that end in *-ly* as well as degree words like *quite*, *too* and *very*, posthead modifiers like *enough* and *indeed* (as in *good enough*, *very well indeed*), and negative markers like *not*, *n't* and *never*.

Adverb, comparative—RBR

Adverbs with the comparative ending *-er* but without a strictly comparative meaning, like *later* in *We can always come by later*, should simply be tagged as RB.

Adverb, superlative—RBS

Difficult Cases:

4 Confusing parts of speech

This section discusses parts of speech that are easily confused and gives guidelines on how to tag such cases.

CC or DT

When they are the first members of the double conjunctions *both* ... *and*, *either* ... *or* and *neither* ... *nor*, *both*, *either* and *neither* are tagged as coordinating conjunctions (CC), not as determiners (DT).

EXAMPLES: Either/DT child could sing.

But:

Either/CC a boy could sing or/CC a girl could dance.
Either/CC a boy or/CC a girl could sing.
Either/CC a boy or/CC girl could sing.

Hiring Annotators

- **Yourself:** option for smaller-scale projects
- **Colleagues:** friends or other students/co-workers
- Online:
  - **Freelancers:** Through sites like UpWork
  - **Crowd Workers:** Through sites like Mechanical Turk
- Hire for a small job first to gauge timeliness/accuracy, then hire for bigger job!
- Note: *IRB approval* may be necessary for subjective tasks

Assessing Annotation Quality

- **Human Performance (Accuracy/BLEU/ROUGE):** Double-annotate some data, measure metrics
- Cohen's **Kappa Statistic** (Cohen 1960):

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - \boxed{p_o}}{1 - \boxed{p_e}}$$

Observed agreement

Expected agreement

Assessing Annotation Quality

- **Example**: 2 annotators classify 20 examples as positive or negative

	B: Pos	B: Neg	Total
A: Pos	8	2	10
A: Neg	3	7	10
Total	11	9	20

- $$p_e = \left(\frac{10}{20}\right)\left(\frac{11}{20}\right) + \left(\frac{10}{20}\right)\left(\frac{9}{20}\right)$$
  = 0.50
- $$p_o = \frac{8+7}{20}$$
  = 0.75
- $$\kappa = \frac{p_o - p_e}{1 - p_e}$$
  = 0.5

No agreement	<0
Slight	0-0.20
Fair	0.21-0.40
Moderate	0.41-0.60
Substantial	0.61-0.80
Almost perfect	0.81-1.0

[Landis & Koch] (Arbitrary, based on opinion)

Assessing Annotation Quality

- **Cohen Kappa**: 2 annotators
- **Fleiss' Kappa**: multiple annotator generalization
- **Krippendorff's Alpha**: more flexible (ordinal & interval data, varied number of annotators, missing data)

Assessing Annotation Quality

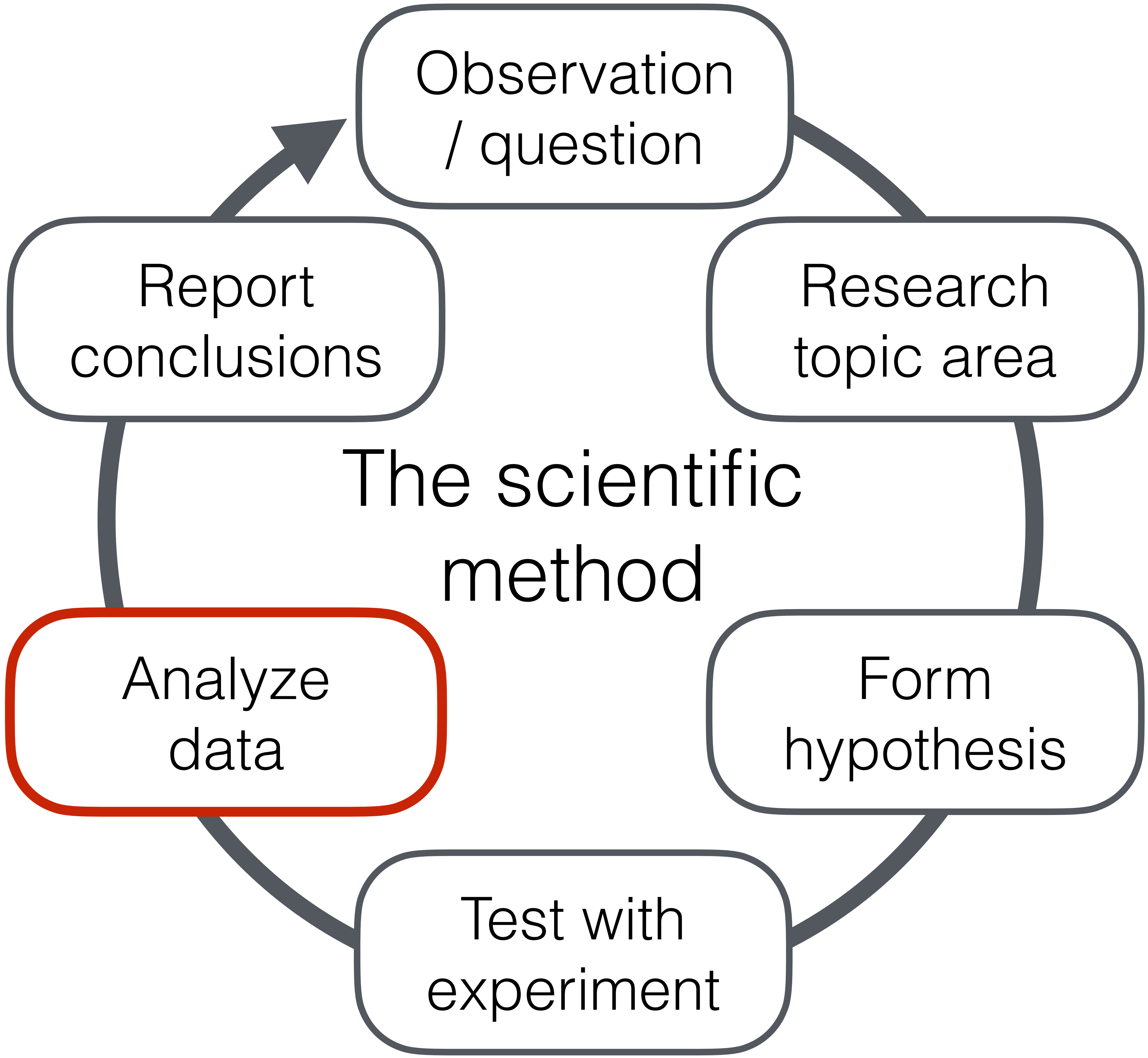
- If agreement statistics are low you may need to:
    - Revisit guidelines
    - Hire better annotators
    - Rethink whether task is possible

Other tips

Computational Resources

- **Online resources:**
  - Amazon Web Services (class credits)
  - Google Cloud/Colab + TPU Research Cloud (TPU)
- **Build your own:**
  - Commodity GPUs RTX 3090 (24GB), A6000 (48GB)

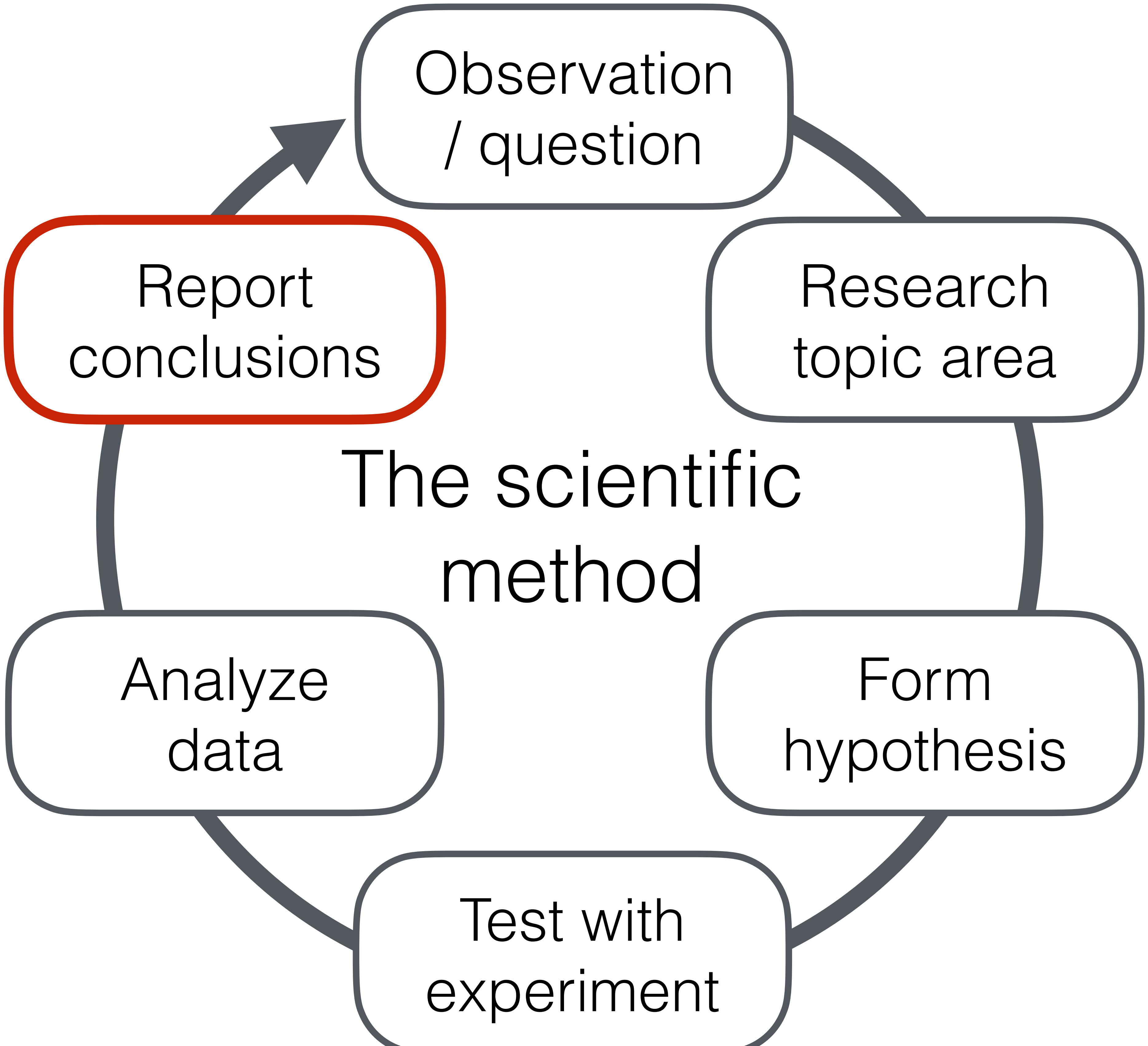
Analyzing Data



Data Analysis

- Look at the data, of course!
- Quantitative analysis
- Qualitative analysis

Reporting Conclusions



Paper Writing Process

- Too much for a single class, but highly recommend

How to Write a Great Research Paper
Simon Peyton-Jones

https://www.microsoft.com/en-us/research/academic-program/write-great-research-paper/

Questions?