CS11-711 Advanced NLP

Research Skills & Experimental Design

Sean Welleck

https://cmu-l3.github.io/anlp-fall2025/

https://github.com/cmu-l3/anlp-fall2025-code

Many 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

Eval \ Model	"Galleon"	"Dreadnought"	Difference
MATH	65.5%	63.0%	+2.5%
HumanEval	83.6%	87.7%	−3.1%
MGSM	75.3%	78.0%	−2.7%

- What can we conclude?

Statistical background on evaluation

- Suppose an eval consists of *N* independently drawn questions, $q_1, \ldots, q_N$
- Let $\bar{s} = \frac{1}{n}\sum_i s_i$ be the average of observed model scores s_i
- Let μ be the unobserved true underlying score, $\mu = \mathbb{E}[s]$

Statistical background on evaluation

- By the law of large numbers, we can estimate $\mu \approx \bar{s}$
- By the central limit theorem, the standard error of the estimator can be estimated as:
- $$SE_{CLT} = \sqrt{Var(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_i - \bar{s})^2\right)/n}$$
- $$SE_{Bernoulli} = \sqrt{\bar{s}(1-\bar{s})/n}$$

Confidence interval

- $CI_{95\%} = \bar{s} \pm 1.96 \times SE$
- We can report:
  - Number of questions N

- The standard error or a confidence interval

	# Questions	"Galleon"	"Dreadnought"
MATH	5,000	65.5% (0.7%)	63.0% (0.7%)
HumanEval	164	83.6% (3.2%)	86.7% (3.0%)
MGSM	2,500	75.3% (0.9%)	78.0% (0.9%)

Code example

Clustered questions

- We assumed that questions are drawn independently, but often they are not
- For instance, we may have a single math problem translated into multiple languages (MGSM)
- We can account for such “clustering” of question using a different standard error estimator:

$$\mathrm{SE}_{\text{clustered}} = \left(\mathrm{SE}^2_{\text{C.L.T.}} + \frac{1}{n^2} \sum_c \sum_i \sum_{j \neq i} (s_{i,c} - \bar{s})(s_{j,c} - \bar{s}) \right)^{1/2}$$

Clustered questions

	# Questions	# Clusters	"Galleon"	"Dreadnought"
DROP	9,622	588	87.1 (0.8)	83.1 (0.9)
RACE-H	3,498	1,045	91.5% (0.5%)	82.9% (0.7%)
MGSM	2,500	250	75.3% (1.6%)	78.0% (1.5%)

	$SE_{clustered}$	$SE_{C.L.T.}$	Ratio
DROP	(1.34)	(0.44)	3.05
RACE-H	(0.51%)	(0.46%)	1.10
MGSM	(1.62%)	(0.86%)	1.88

Comparing models: unpaired

- Difference of means: $\hat{\mu}_{A-B} = \hat{\mu}_A - \hat{\mu}_B$
    - Null hypothesis: difference of means is 0
- Standard error: $SE_{A-B} = \sqrt{SE_A^2 + SE_B^2}$
- Confidence interval: $CI_{A-B,95\%} = \hat{\mu}_{A-B} \pm 1.96 \times SE_{A-B}$
    - If this doesn't include 0, the result is statistically significant
- Compute z score: $z_{A-B} = \hat{\mu}_{A-B} / SE_{A-B}$
    - Standardizes the difference
- Get associated *p*-value
    - Probability of observing this difference under the null hypothesis
- If below a threshold (e.g., $p < 0.01$), reject the null hypothesis

Code example

Comparing models: paired

- Evaluate both systems on the same examples
  - Suppose we have access to all of the evaluations, (x, y_A, y_B)
  - Then we can use a “paired” test that typically has reduced variance.

$$\mathrm{SE}_{A-B,\mathrm{paired}} = \sqrt{\mathrm{Var}(s_{A-B})/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_{A-B,i} - \bar{s}_{A-B})^2\right)/n}$$

Comparing models: paired

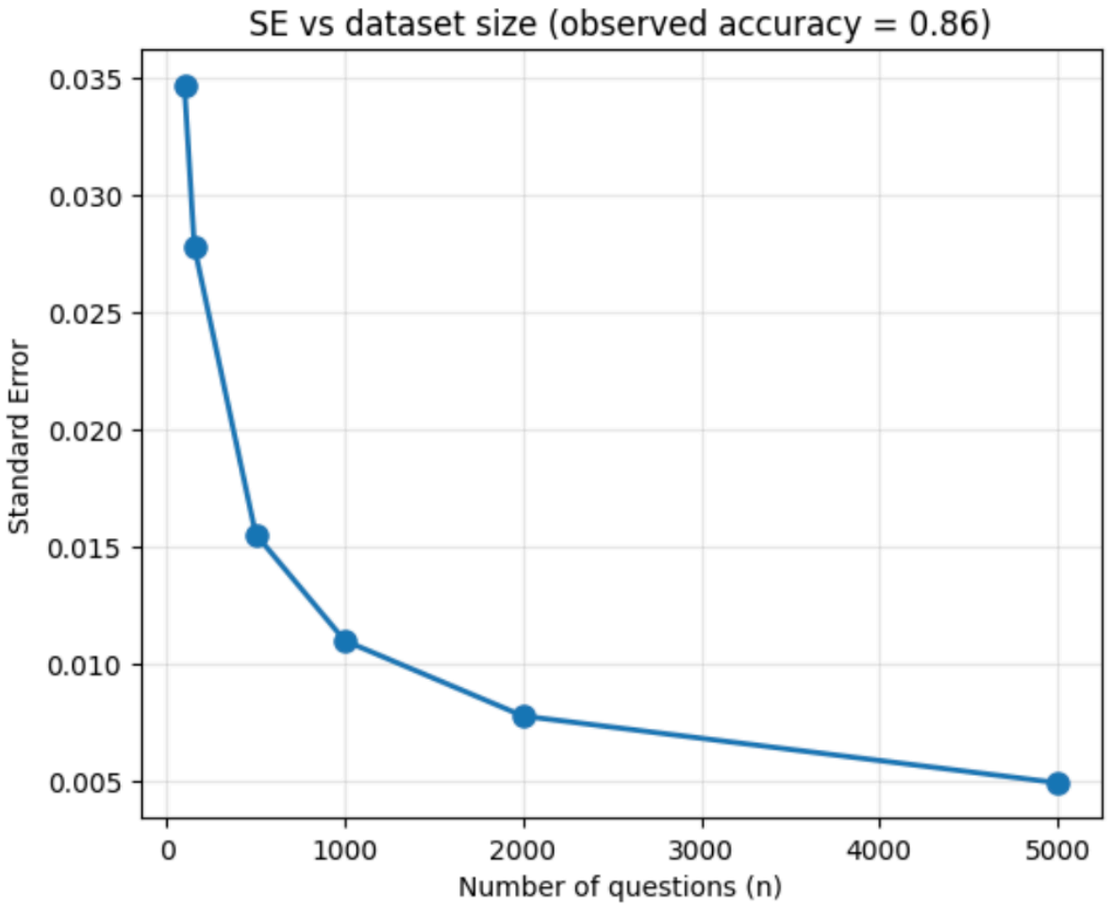
Eval	Model	Baseline	Model – Baseline	95% Conf. Interval	Correlation
MATH	Galleon	Dreadnought	+2.5% (0.7%)	(+1.2%, +3.8%)	0.50
HumanEval	Galleon	Dreadnought	−3.1% (2.1%)	(−7.2%, +1.0%)	0.64
MGSM	Galleon	Dreadnought	−2.7% (1.7%)	(−6.1%, +0.7%)	0.37

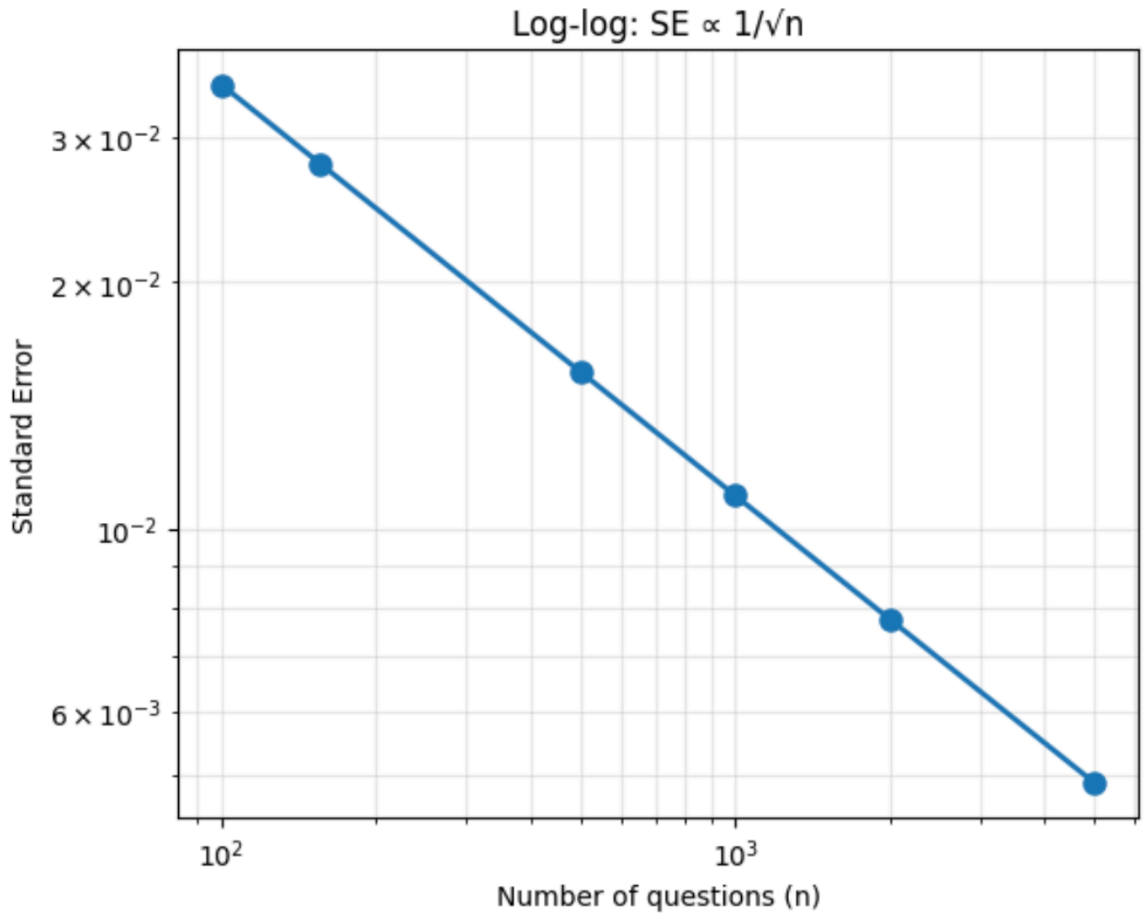
Code example

Variance reduction

- Recall that the estimator is:
  - $\hat{\mu} = \sum_{i=1}^{n} s_i / n$
- Then the variance is $Var(\hat{\mu}) = Var(s)/n$
- To reduce variance:
  - Increase number of questions n
  - If we are using stochastic decoding, sample more outputs and take the average as the score.

Variance reduction



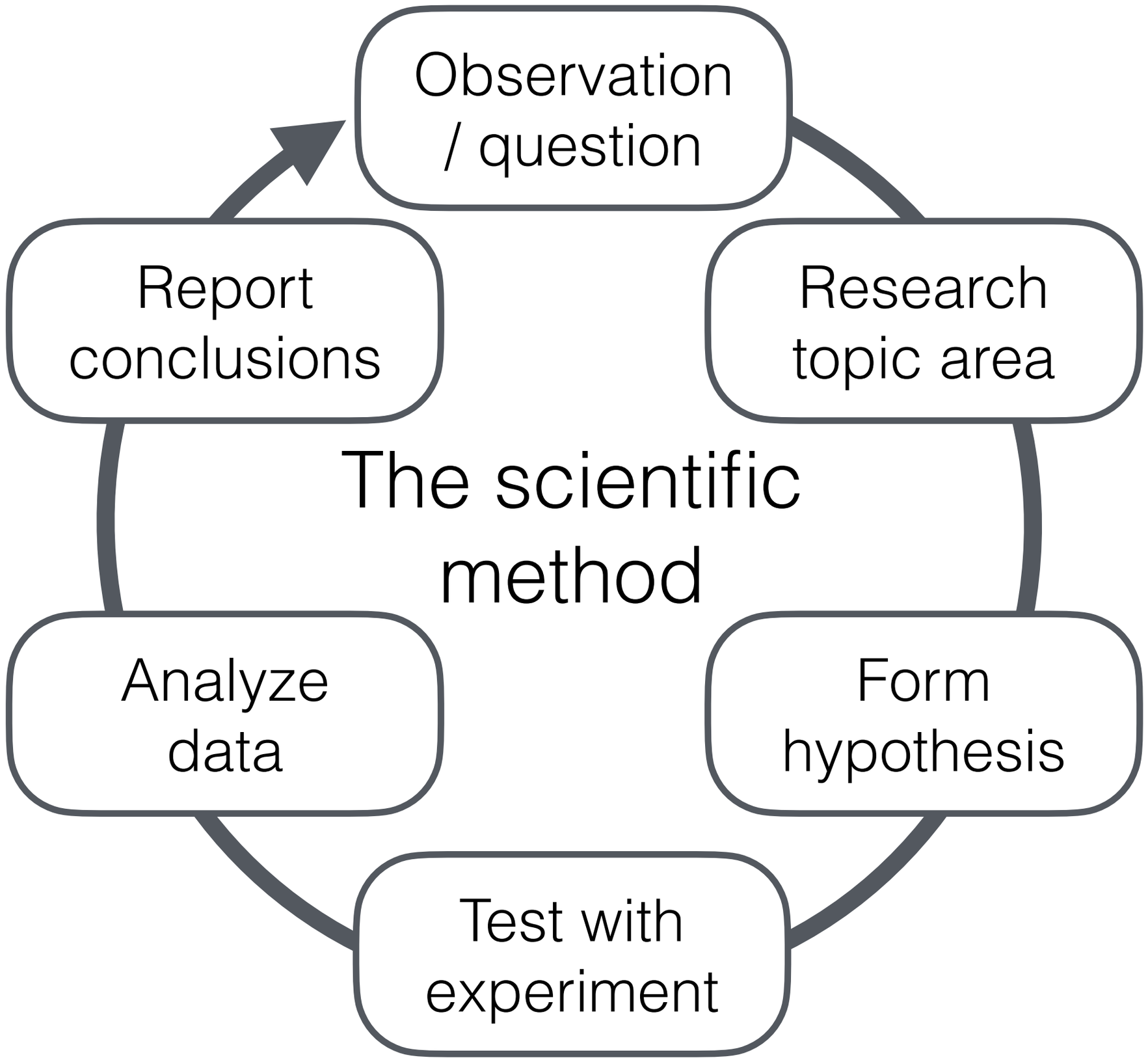



```
n=  100: SE=0.0347
n=  156: SE=0.0278
n=  500: SE=0.0155
n= 1000: SE=0.0110
n= 2000: SE=0.0078
n= 5000: SE=0.0049
```

This lecture

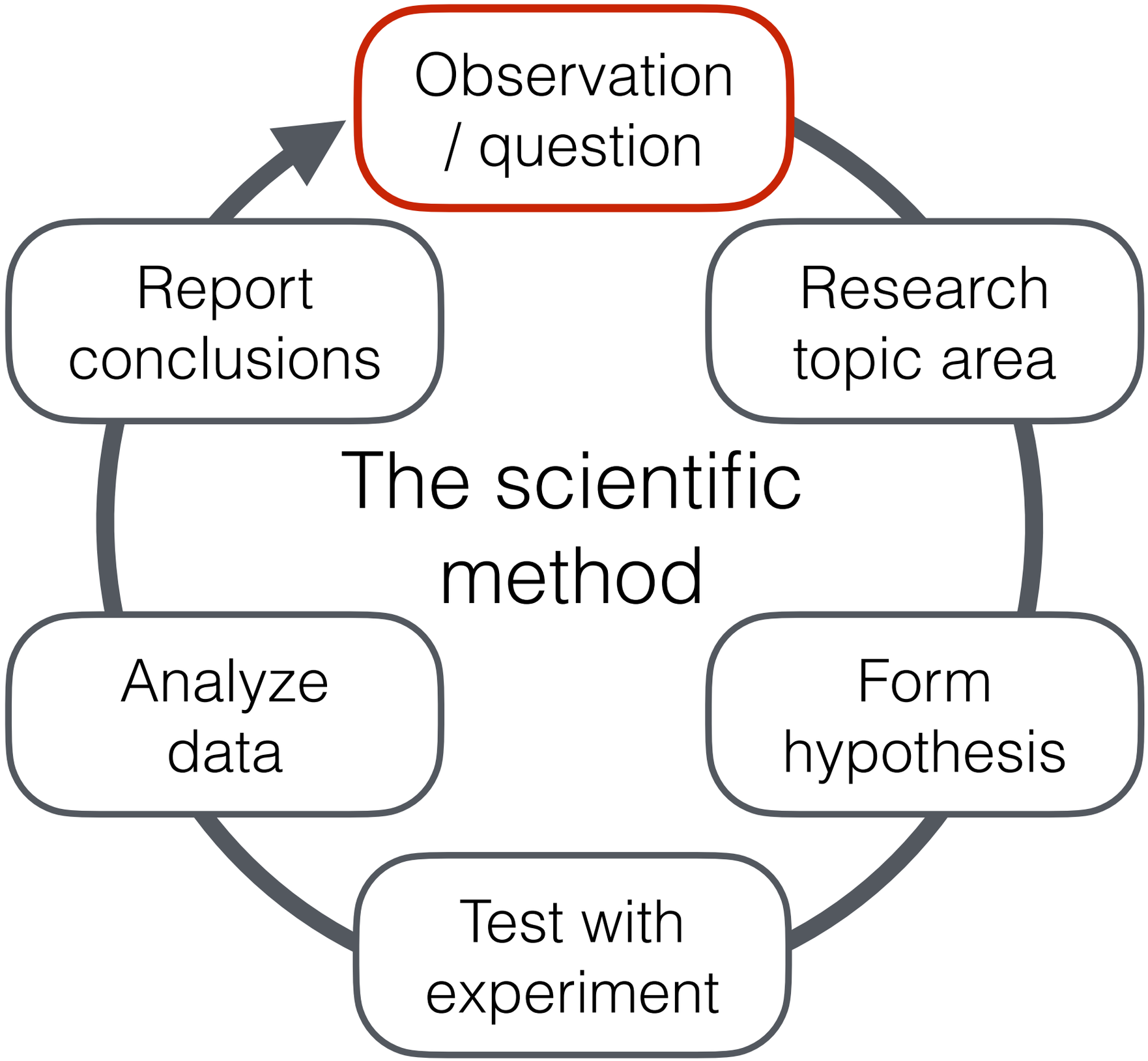
- What can I conclude from a study?

- Next: how do I conduct a new study?



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*?

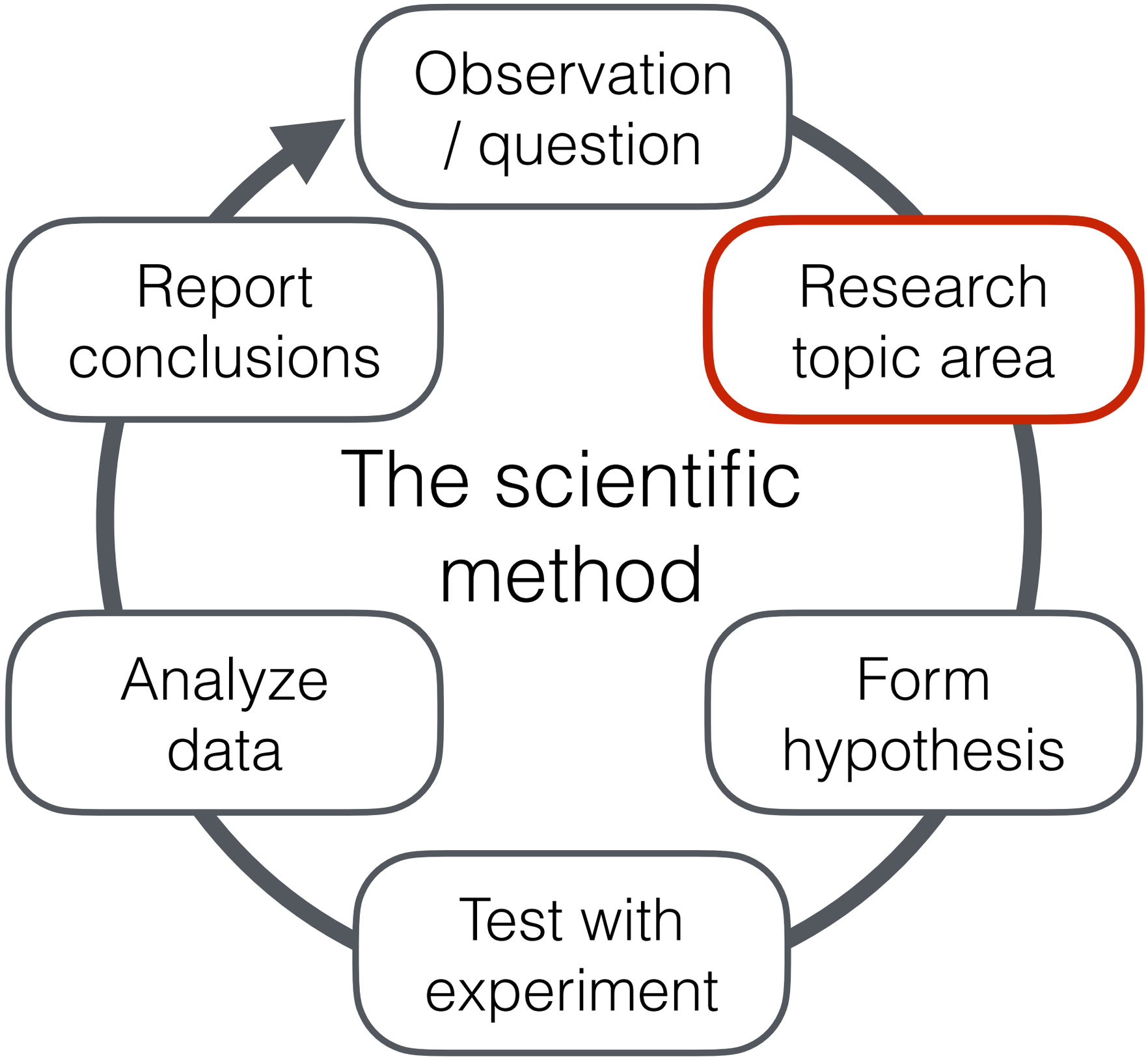
- Cotterell et al. (2018) ask "*are all languages equally hard to language model*?"

- 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/key results**
- Read **details of most relevant papers**

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

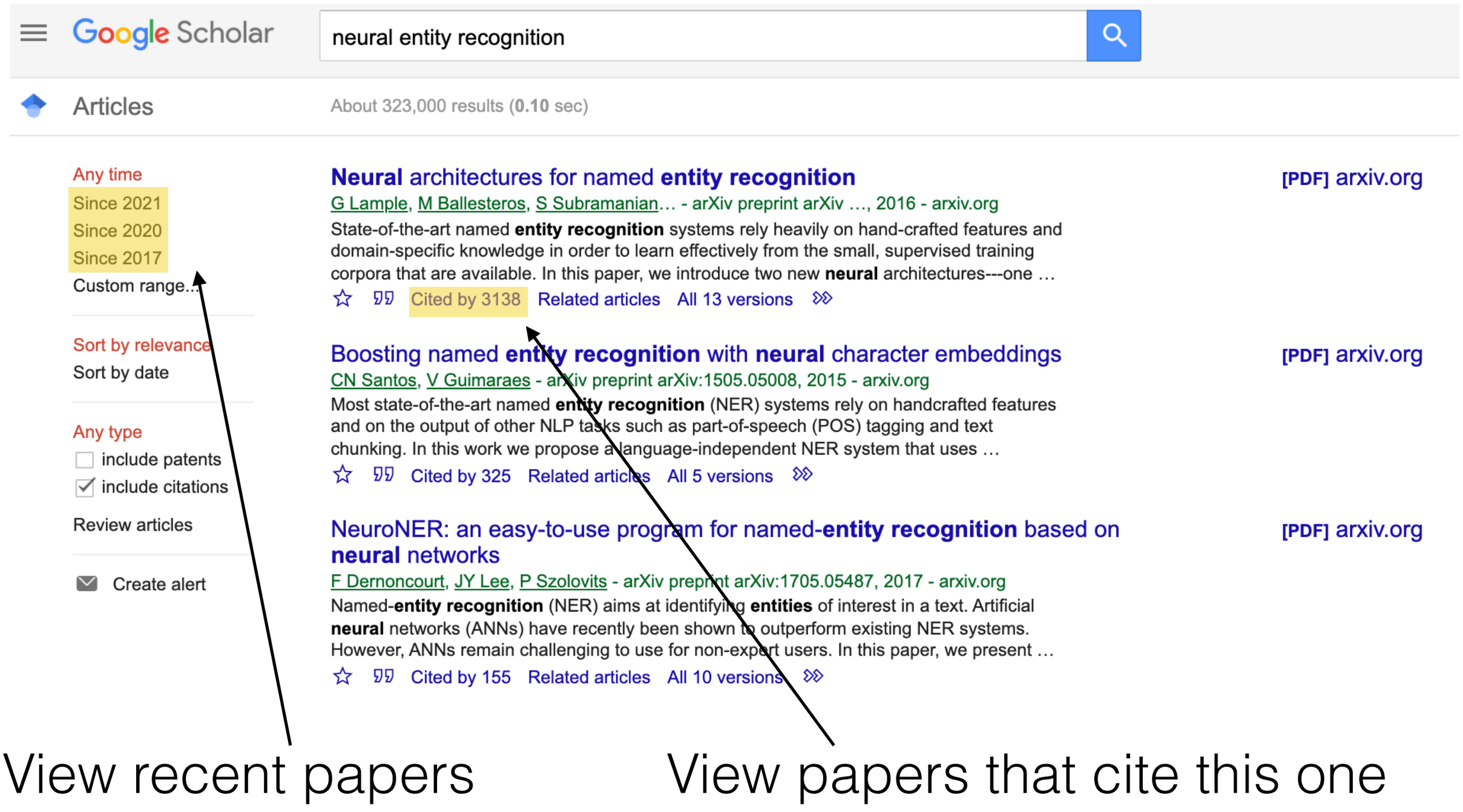
- Covers many prestigious venues in NLP
- Start with past 3-5 years of several top venues (e.g. ACL, EMNLP, NAACL, TACL)

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*.

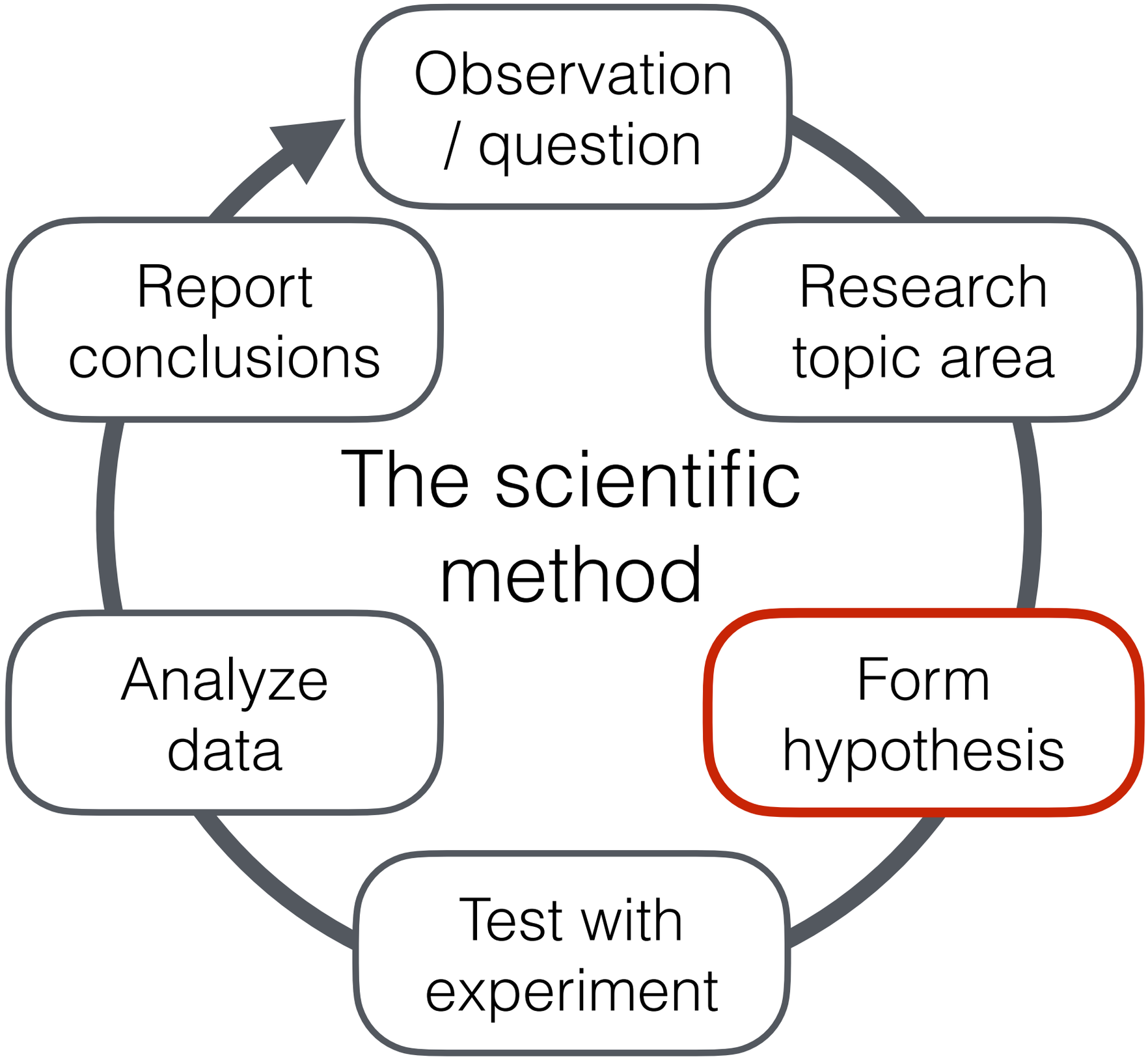
Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. In *Advances in NIPS*, 5014–5026.

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*.

Pros and Cons 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.

Cotterell et al. (2018)

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.

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)

Q2 Do pre-trained embeddings help more when the size of the training data is small? (§4)

Q3 How much does the similarity of the source and target languages affect the efficacy of using pre-trained embeddings? (§5)

Q4 Is it helpful to align the embedding spaces between the source and target languages? (§6)

Q5 Do pre-trained embeddings help more in multilingual systems as compared to bilingual systems? (§7)

Yes?

Yes?

Not much?

Yes?

Unclear

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?

Qi et al. (2018)

Zhang et al. (2021)

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

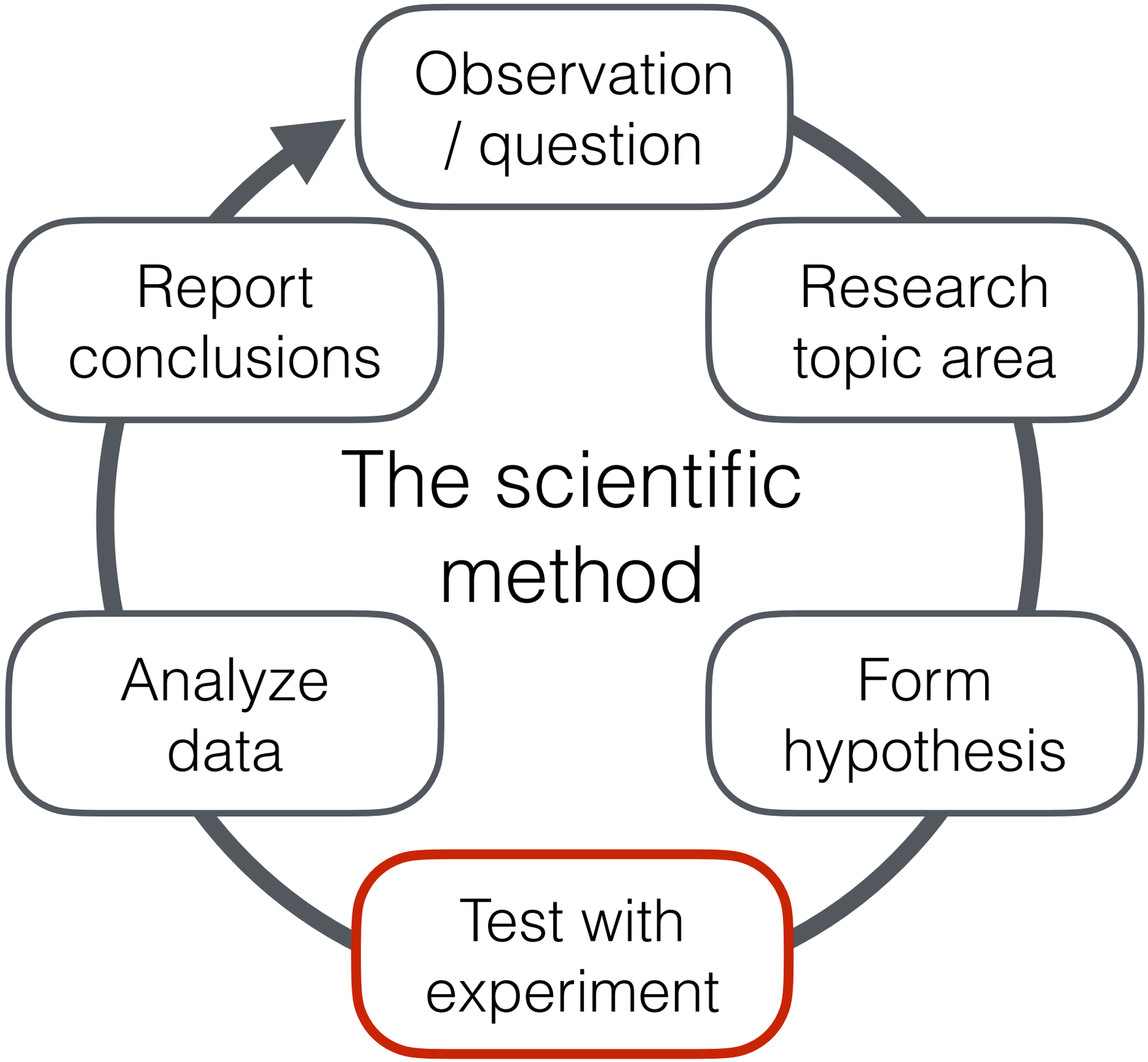
- The above question/hypothesis is natural, but indirect

  - 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

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

Papers With Code

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

Annotating Data

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

Example

- Suppose we want to train a classifier to predict whether a movie review is “positive” or “not positive”

- We want to collect movie reviews, and ask human annotators to label them as positive or negative

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**
  - 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

How Much Test/Dev Data Do I Need?

- Null hypothesis: system A and system B perform equally
- Significance level α
  - Probability of falsely detecting a difference
- Power level $1 - \beta$
  - Probability of detecting a true difference when it exists
- Minimum detectable effect δ
  - The smallest difference we care about detecting
- Assume we evaluate system A and system B, either on a small amount of “pilot” data

How Much Test/Dev Data Do I Need?

- Number of questions required to achieve a Type I error rate α and Type II error rate β with minimum detectable effect δ:

$$n = (z_{\alpha/2} + z_{\beta})^2(\omega^2 + \sigma_A^2/K_A + \sigma_B^2/K_B)/\delta^2$$

- ω^2: across-question variance of the true performance difference between A and B
- σ_A^2, σ_B^2: average per-question variance with K samples
  - (e.g., using stochastic decoding and/or a stochastic evaluator)
- Assuming deterministic decoding and evaluation:

$$n = (z_{\alpha/2} + z_{\beta})^2\omega^2/\delta^2$$

How Much Test/Dev Data Do I Need?

- Example: collect a pilot set of 156 items
- Model A and model B achieve 0.86, with a cross-question variance $\hat{\omega}^2$ of 0.077
- Type 1 error α: 0.05
- Power $(1 - \beta)$: 0.8
- Minimum detectable difference δ: 0.03

$$n = (z_{\alpha/2} + z_{\beta})^2 \omega^2 / \delta^2$$

- => required *n*: 674

How Much Training Data Do I Need?

- More is usually better
- Collect in phases, fine-tune a model on increasing number of examples and evaluate marginal improvements
- Can do even better with intelligent data selection - active learning

Annotation Guidelines

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

What:

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

Assessing Annotation Quality

- Suppose multiple human raters label the data.
  - Example: label our n=674 examples as positive or negative
- We want to quantify how much the raters agree, beyond what we would expect by chance.
- Cohen's **Kappa Statistic** (Cohen 1960):

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

p_o: Observed agreement

p_e: Expected agreement

Assessing Annotation Quality

- Make a confusion matrix
- Observed agreement p_o
  - $p_o = \frac{C_{00} + C_{11}}{674}$
- Expected agreement p_e
  - $p_e = p_0^{(A)} p_0^{(B)} + p_1^{(A)} p_1^{(B)}$
    - e.g. $p_a^{(A)} = \frac{C_{00} + C_{01}}{674}$
- Cohen's κ
  - $\kappa = \frac{p_o - p_e}{1 - p_e}$

	B: Neg	**B: Pos**
A: Neg	C_00	C_01
A: Pos	C_10	C_11

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

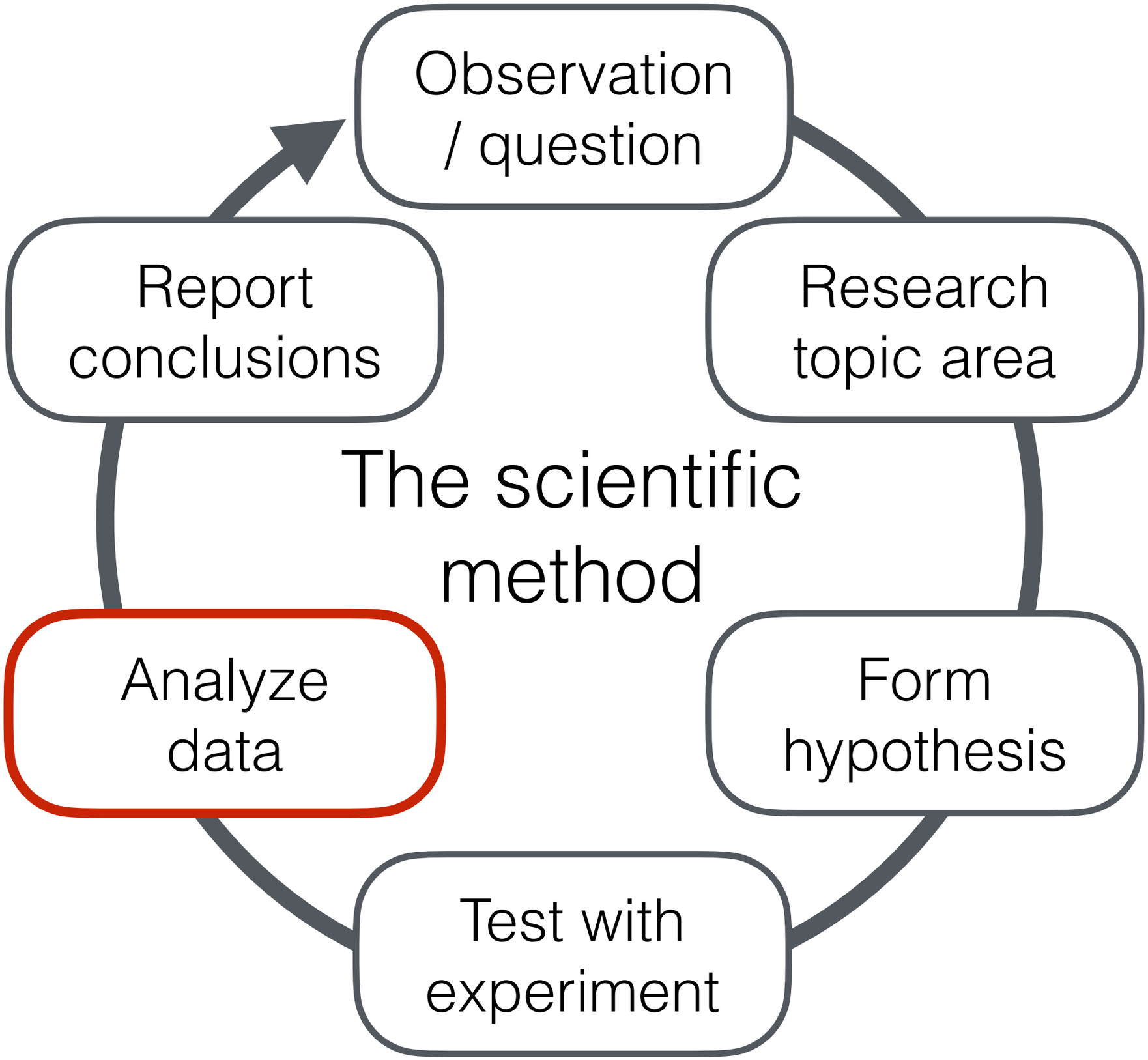
- If agreement statistics are low you may need to:
  - Revisit guidelines
  - Hire better annotators
  - Rethink whether task is possible

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)

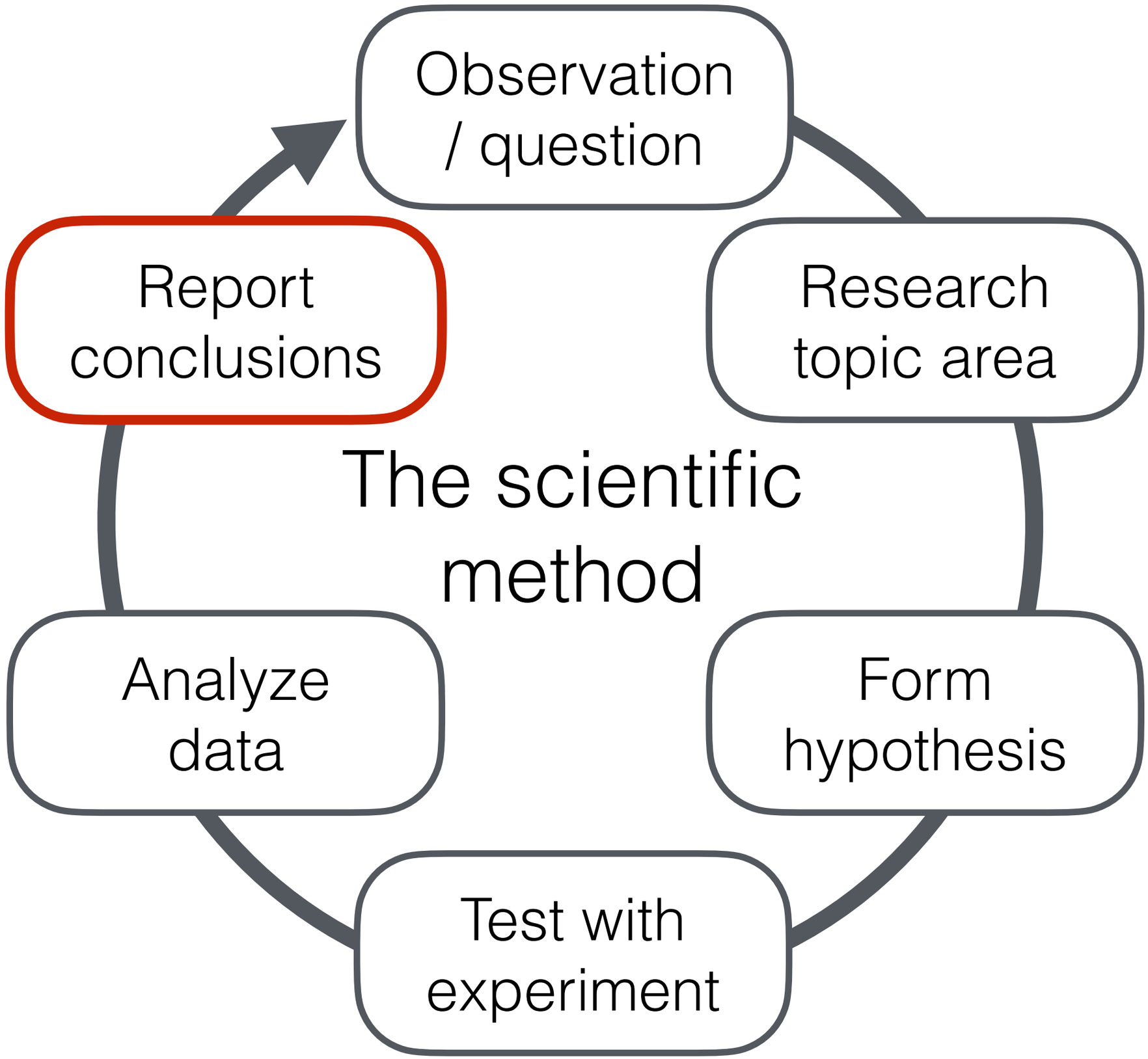
Analyzing Data



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?