Reminder

- HW1 grade update
- HW2 due 09/30
- Paper presentation sign-up due 09/30
  - Or I will randomly assign papers:)
- Project proposal due 10/07

CS678-002 Advanced NLP

Experimental Design and Model Interpretability

Antonis Anastasopoulos & Ziyu Yao

https://cs.gmu.edu/~antonis/course/cs695-fall20/

With many slides from Graham Neubig's CMU NN4NLP course & Greg Durrett CS388@UT Austin

Outline

Cluster Tutorial

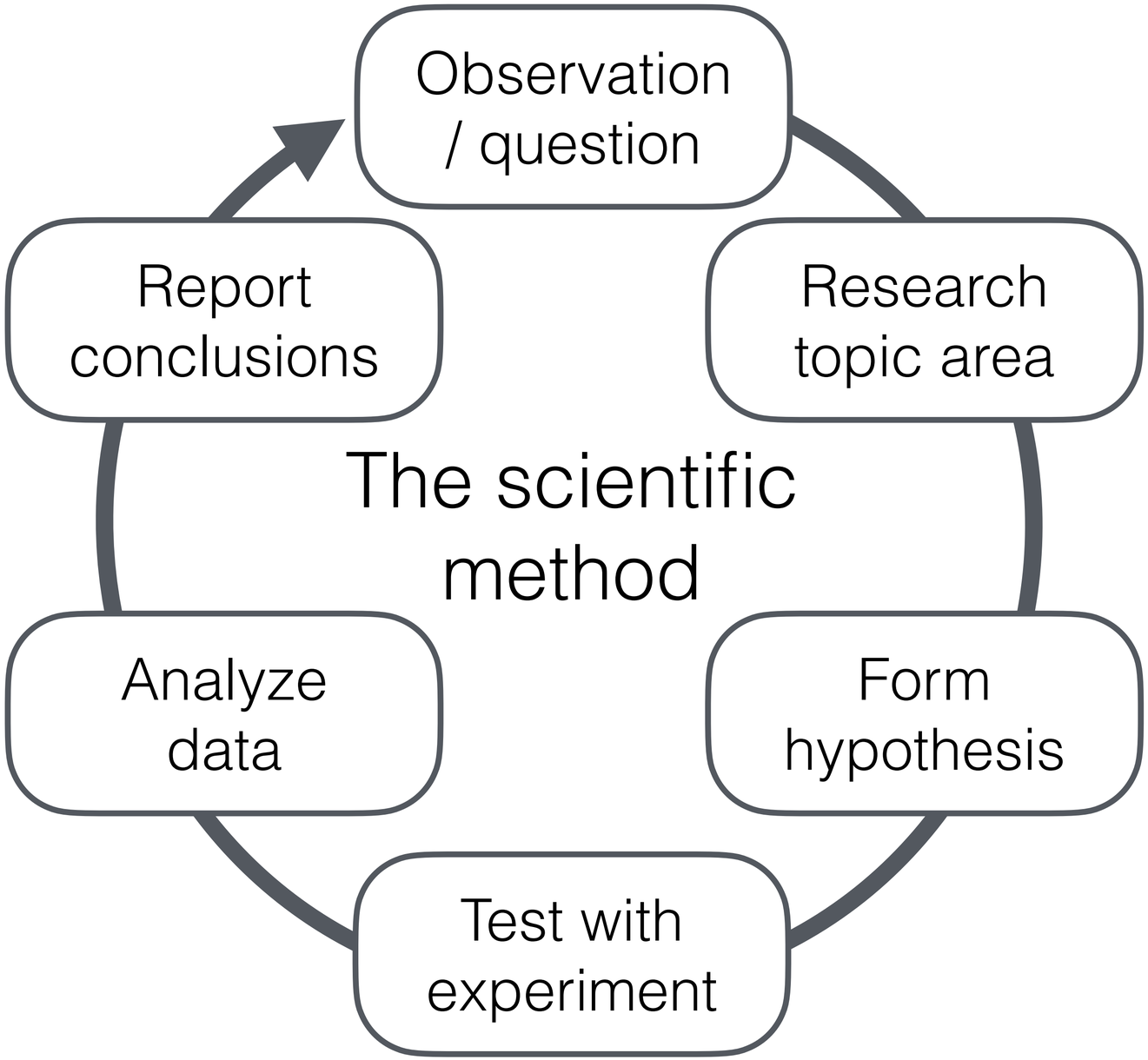
[short break]

Experimental Design

[short break?]

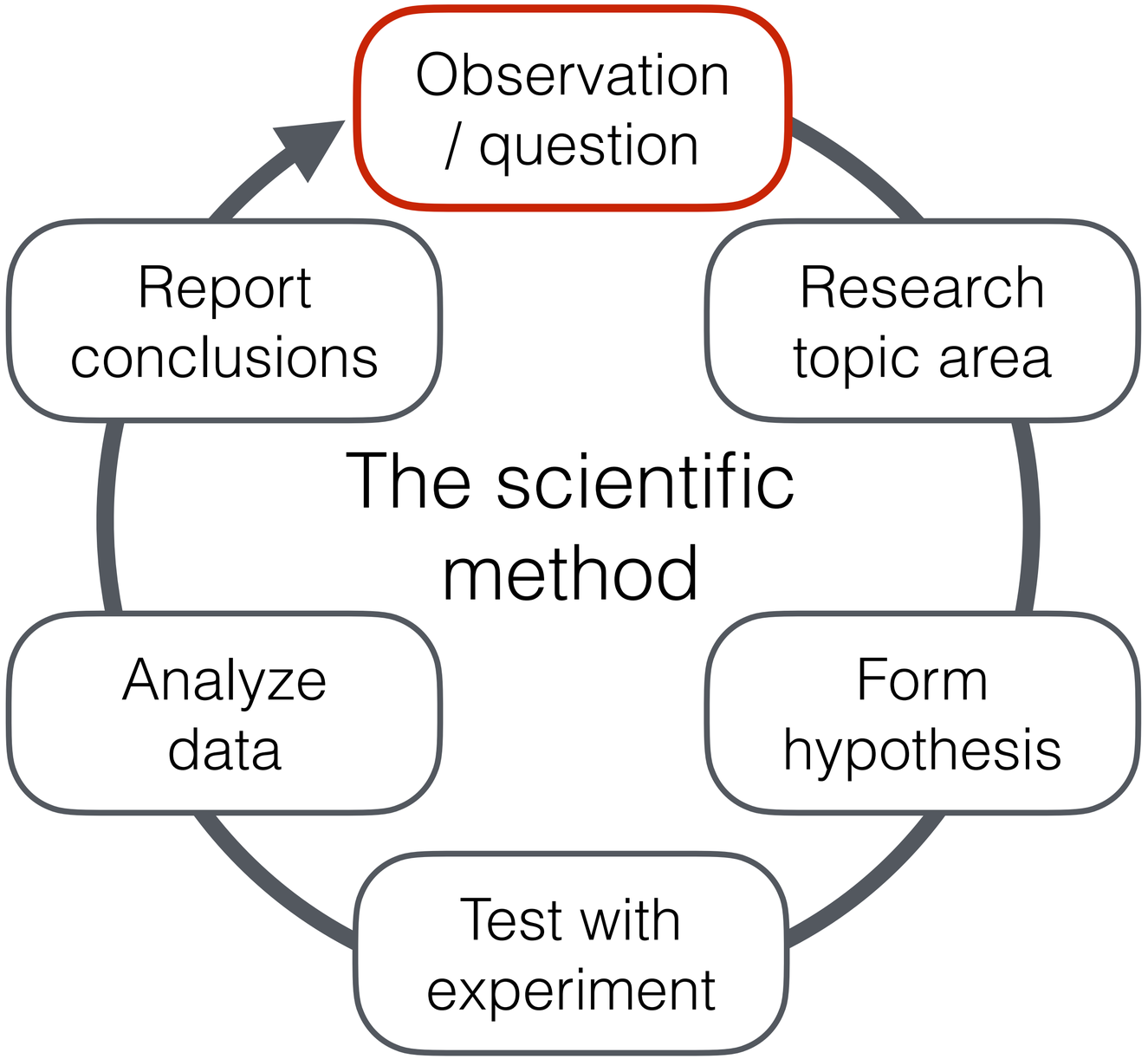
Interpretability of NLP Models

- Demo and in-class exercise



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."

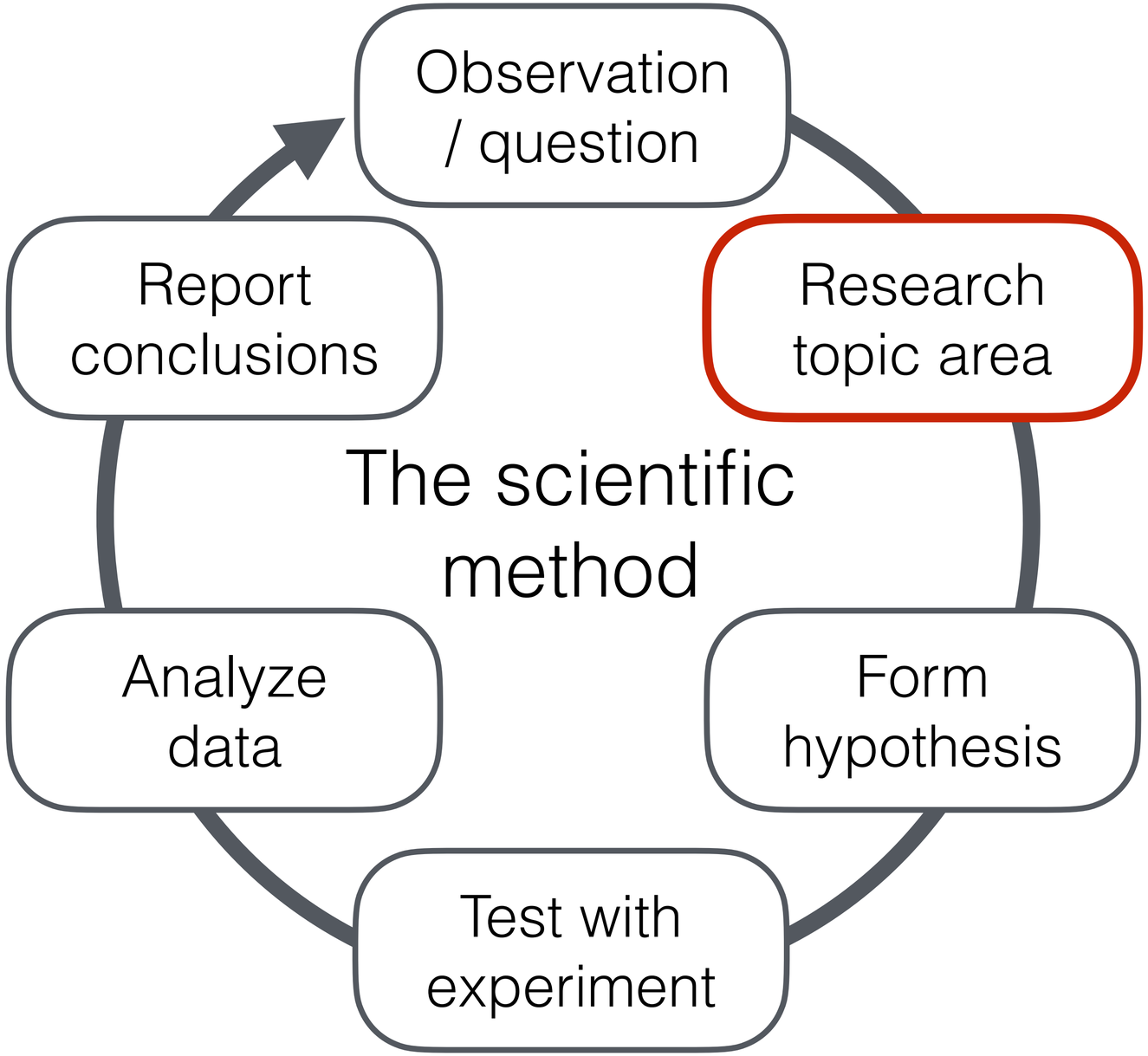
Examples of Curiosity-Driven Research

- 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

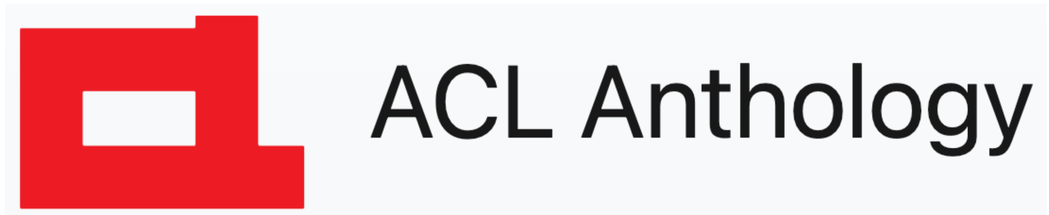
Identifying Good Research Directions



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://aclanthology.org/

https://scholar.google.com/

ACL Anthology

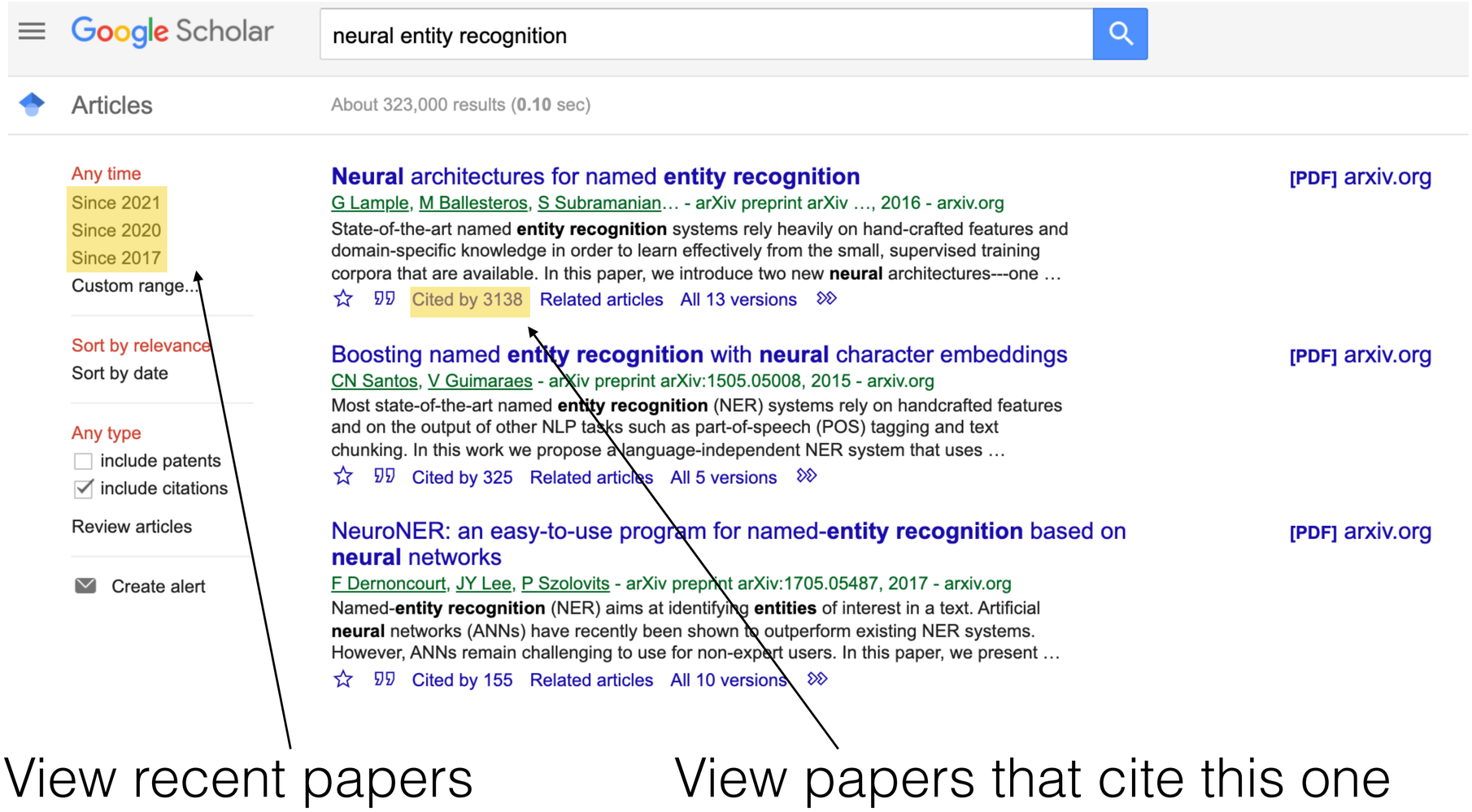
- Covers many prestigious venues in NLP
- Start with past 3-5 years of several top venues (e.g. ACL, EMNLP, NAACL, TACL)

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*.

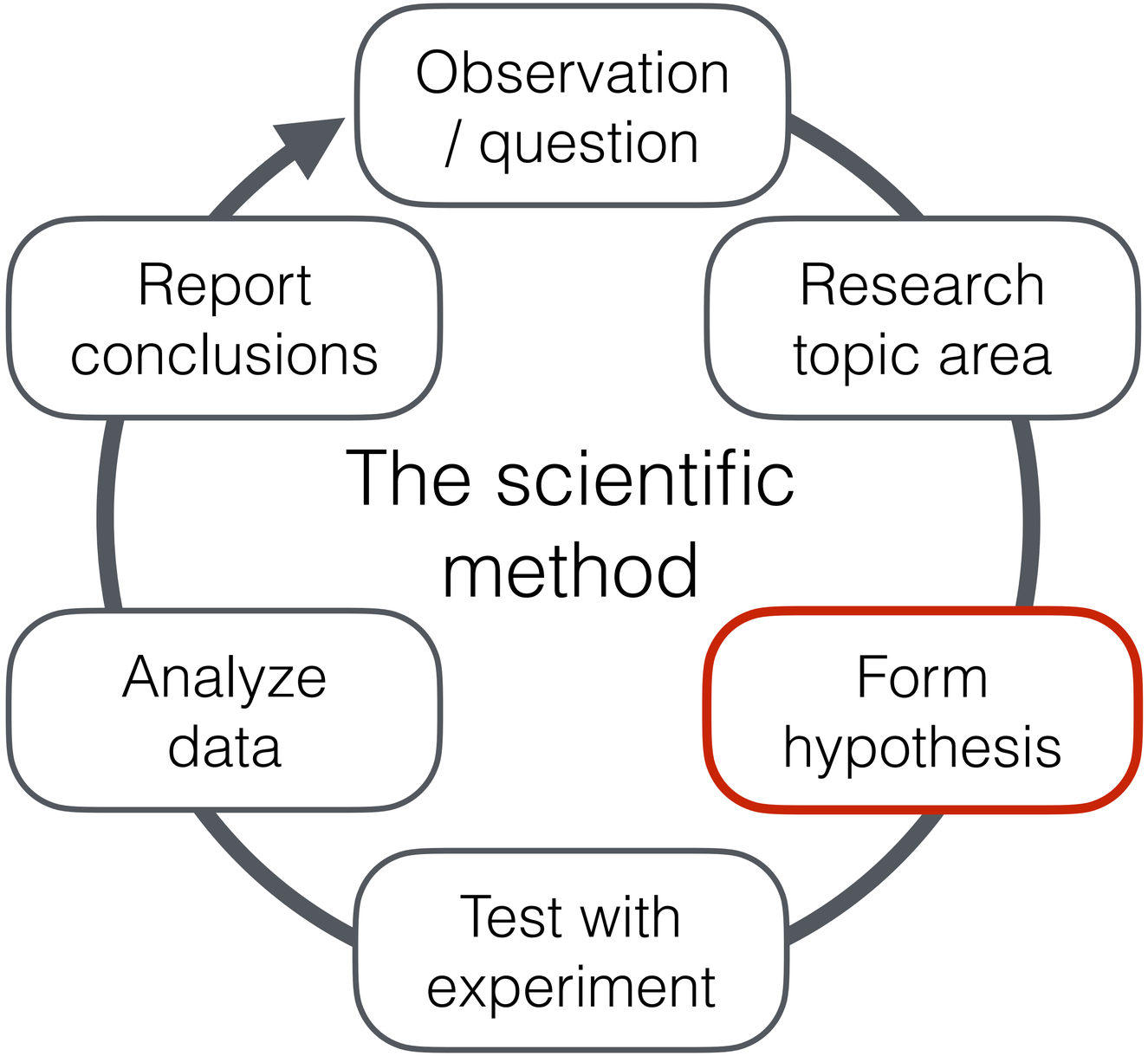
Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. In *Advances in NIPS*, 5014–5026.

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.

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) — 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.

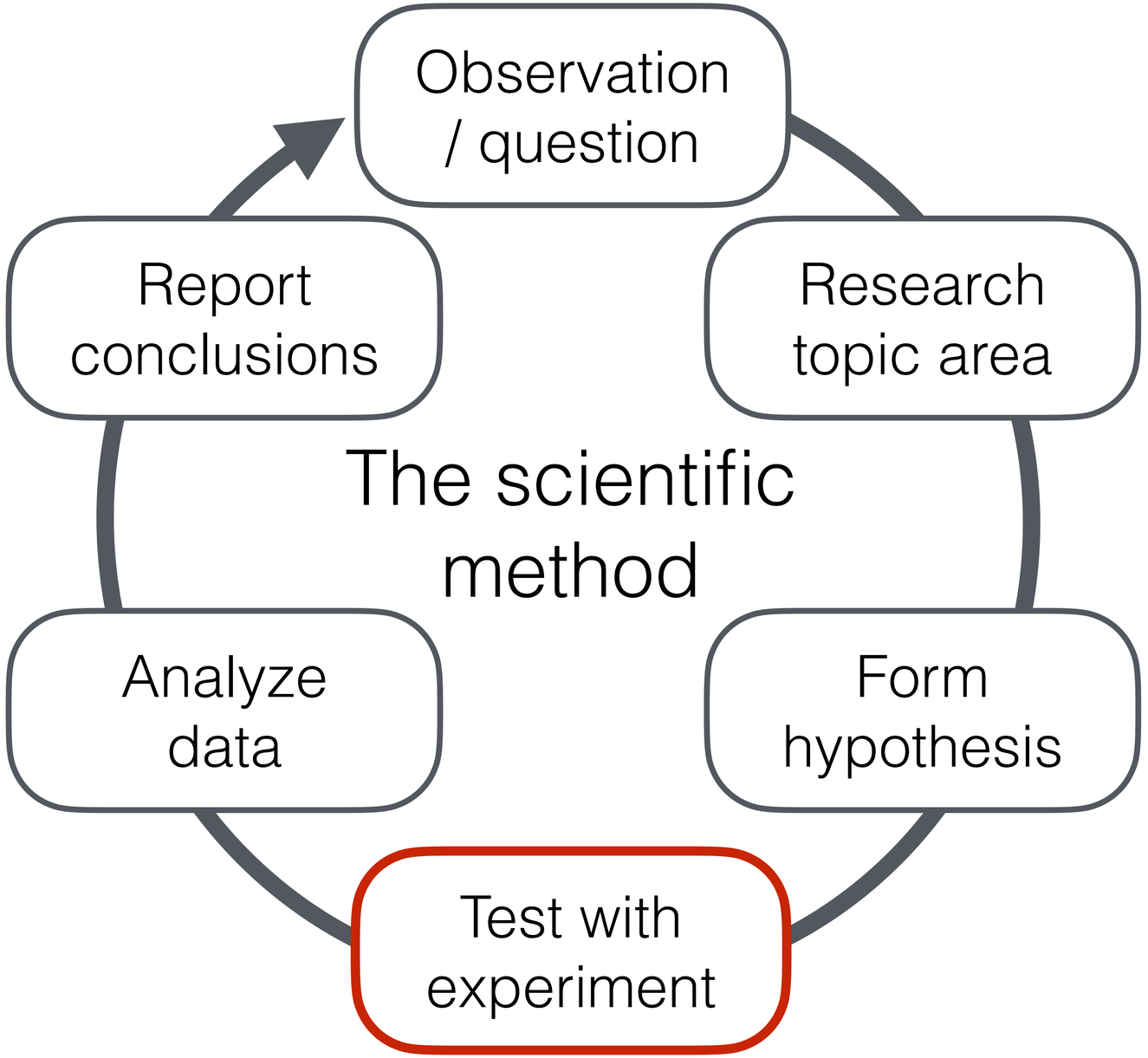
Probably will help?

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

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

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:

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!

Assessing Annotation Quality

- **Human Performance (Accuracy/BLEU/ROUGE):** Double-annotate some data, measure metrics
- **Kappa Statistic** (Carletta 1996):

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

p_o: Observed agreement

p_e: Expected agreement

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

Obtaining Training Data

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

Running Experiments

Workflow Automation

- Modularize each step of experiment into directory in -> directory out
- Name directories by parameters
  `transformer-layer8-node512-dropout0.5-labelsmooth0.02`
- Don't re-run directories that are already done
- More sophisticated: duct-tape (https://github.com/CoderPat/ducttape)

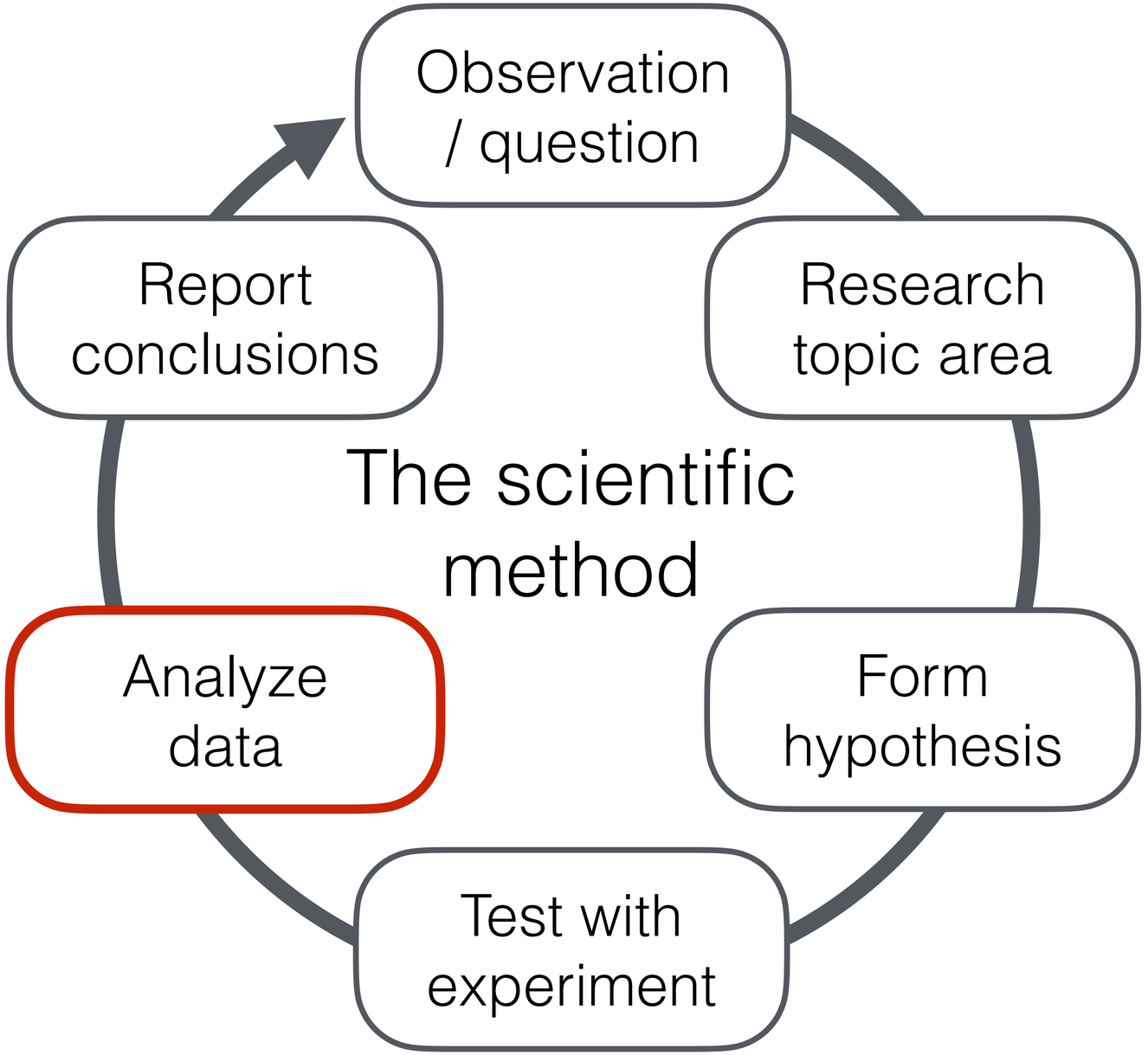
Evaluation

- See previous lectures!
- **Train on train, tune on dev, eval on test**
- **Types of metrics**
  - Accuracy
  - Precision/Recall/F-measure
  - NLG metrics
  - Extrinsic evaluation
- **Statistical significance**

Result Reporting

- **Plan results section** in advance!
  - Identifies unjustified experimental claims
  - Allows for planning in the "best case scenario"
- **Result generation scripts**:
  - Generate paper LaTeX directly from log files
  - Efficient, and minimizes errors
  - Also allows you to pre-emptively plan experiments

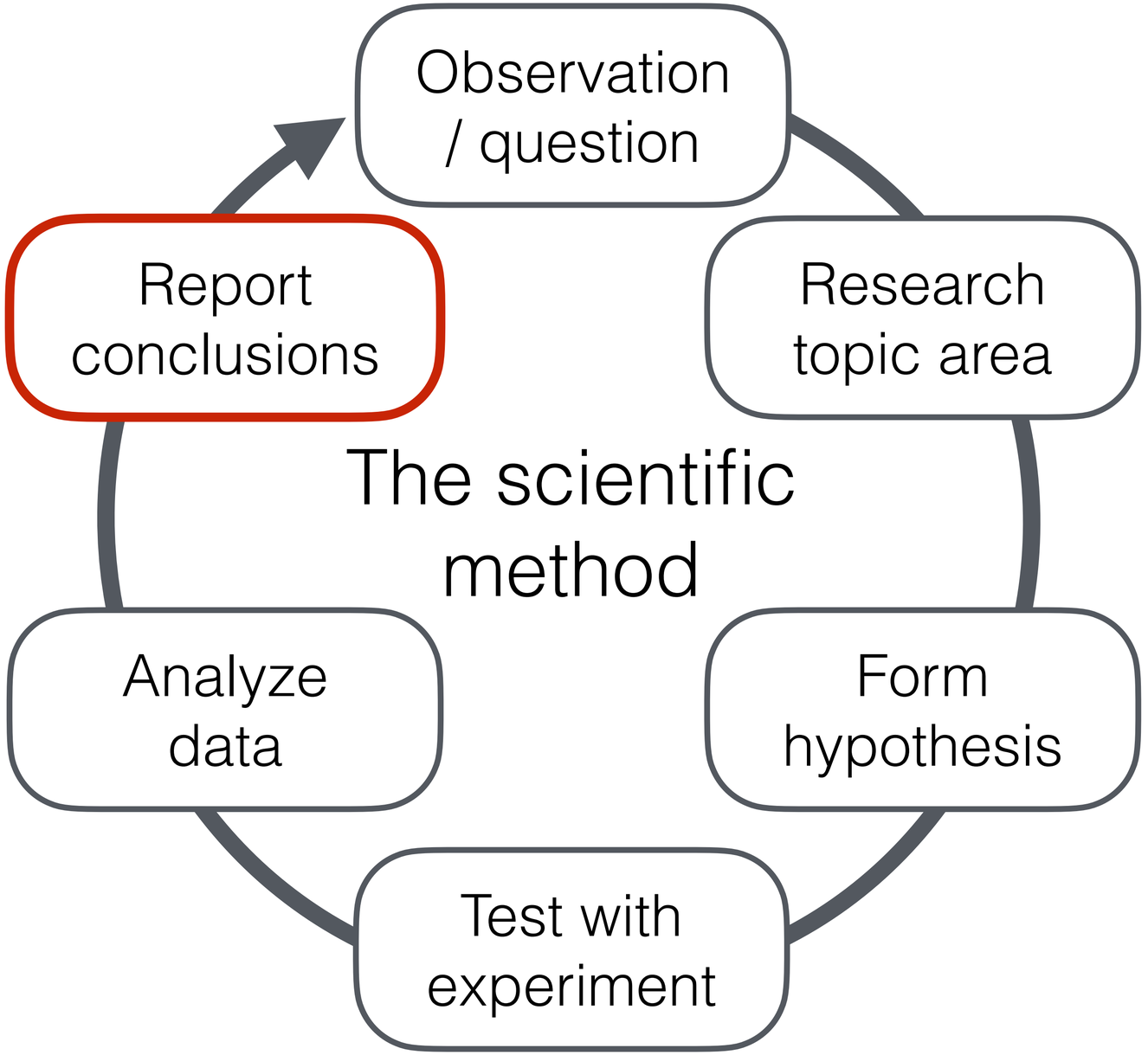
Analyzing Data



Data Analysis

- Look at the data, of course!
- Quantitative analysis
- Qualitative analysis
- Model explanations
- Look at the ‘interpretation” section

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/

Interpretability

NLP Beyond Accuracy

- Is accuracy the only thing we should care about?
- Model training:
  - Where comes the data? Is it private?
  - Is the model easy to train? (e.g., sensitivity to hyper-parameters, time consumption)
  - How much compute power is needed? Is it affordable for general population? Energy consideration (e.g., carbon footprint)
- Model after deployment:
  - Performance in practice: Interpretable behaviors? Robust? Easily attacked?
  - Ethical consideration: any unethical behaviors (bias, toxic language, privacy exposure etc.)? easy to control? accessible to diverse populations (the “democratization” of tech)?

NLP Beyond Accuracy

- Is accuracy the only thing we should care about?
- Model training:
  - Where comes the data? Is it private?
  - Is the model easy to train? (e.g., sensitivity to hyper-parameters, time consumption)
  - How much compute power is needed? Is it affordable for general population? Energy consideration (e.g., carbon footprint)
- Model after deployment:
  - Performance in practice: **Interpretable behaviors**? Robust? Easily attacked?
  - Ethical consideration: any unethical behaviors (bias, toxic language, privacy exposure etc.)? easy to control? accessible to diverse populations (the “democratization” of tech)?

Explaining NLP models

- “Explaining” an NLP model: What does it mean? Why is it an important problem? How to “explain”?

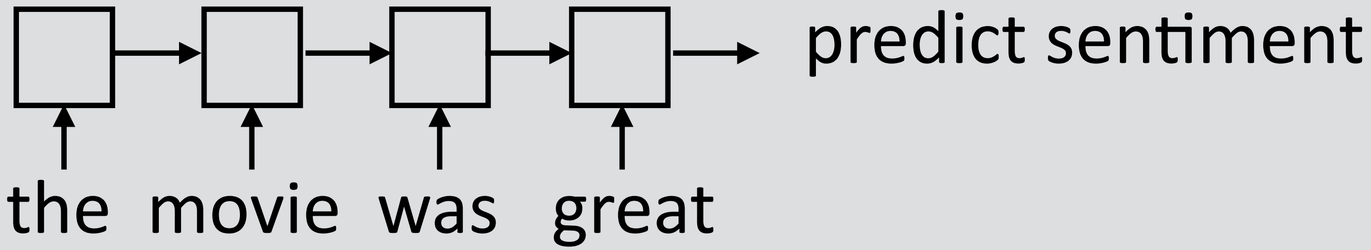
Explaining NLP models

- “Explaining” an NLP model: What does it mean? Why is it an important problem? How to “explain”?
- Local explanations:
  - Erasure techniques
  - Gradient-based methods
- Text-based explanations
- Evaluating explanations

Interpreting Neural Networks

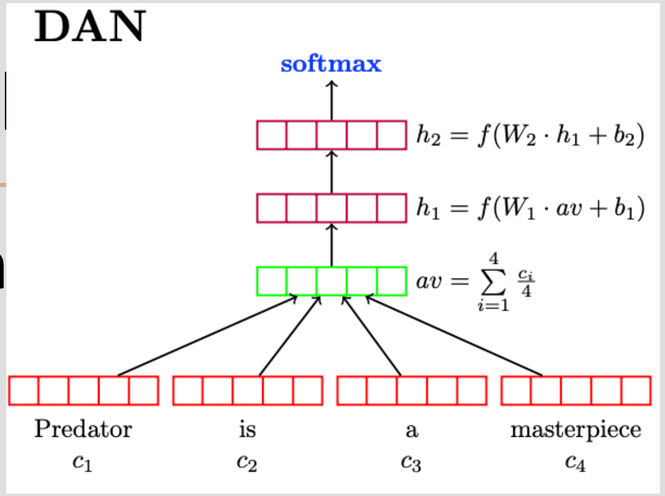
Interpreting Neural Networks

- Neural models have complex behavior. How can we understand them?
- Sentiment w/LSTMs



- Looking at individual neurons usually doesn't tell us much
- Sentiment w/ BERT: there are hundreds of attention computations… which ones actually mean something?

Interpreting Neural Netwo



- Neural models have complex behavior. How can we un
- Sentiment w/DANs:

	DAN	Ground Truth
this movie was not good	negative	negative
this movie was good	positive	positive
this movie was bad	negative	negative
the movie was not bad	negative	positive

  - Left side: predictions the model makes on individual words
  - Tells us how these words combine
- **How do we know why a neural network model made the prediction it made?**

Iyyer et al. (2015)

Why explanations?

- **Trust:** if we see that models are behaving in human-like ways and making human-like mistakes, we might be more likely to trust them and deploy them
- **Causality:** if our classifier predicts class *y* because of input feature *x*, does that tell us that *x* causes *y*? Not necessarily, but it might be helpful to know
- **Informativeness:** more information may be useful (e.g., predicting a disease diagnosis isn't that useful without knowing more about the patient's situation)
- **Fairness:** ensure that predictions are non-discriminatory

Lipton (2016)

Why explanations?

- Some models are naturally **transparent**: we can understand why they do what they do (e.g., a decision tree with <10 nodes)

- Explanations of more complex models

  - **Local explanations:** highlight what led to this classification decision. (Counterfactual: if these features were different, the model would've predicted a different class) — focus of this lecture

  - **Text explanations:** describe the model's behavior in language

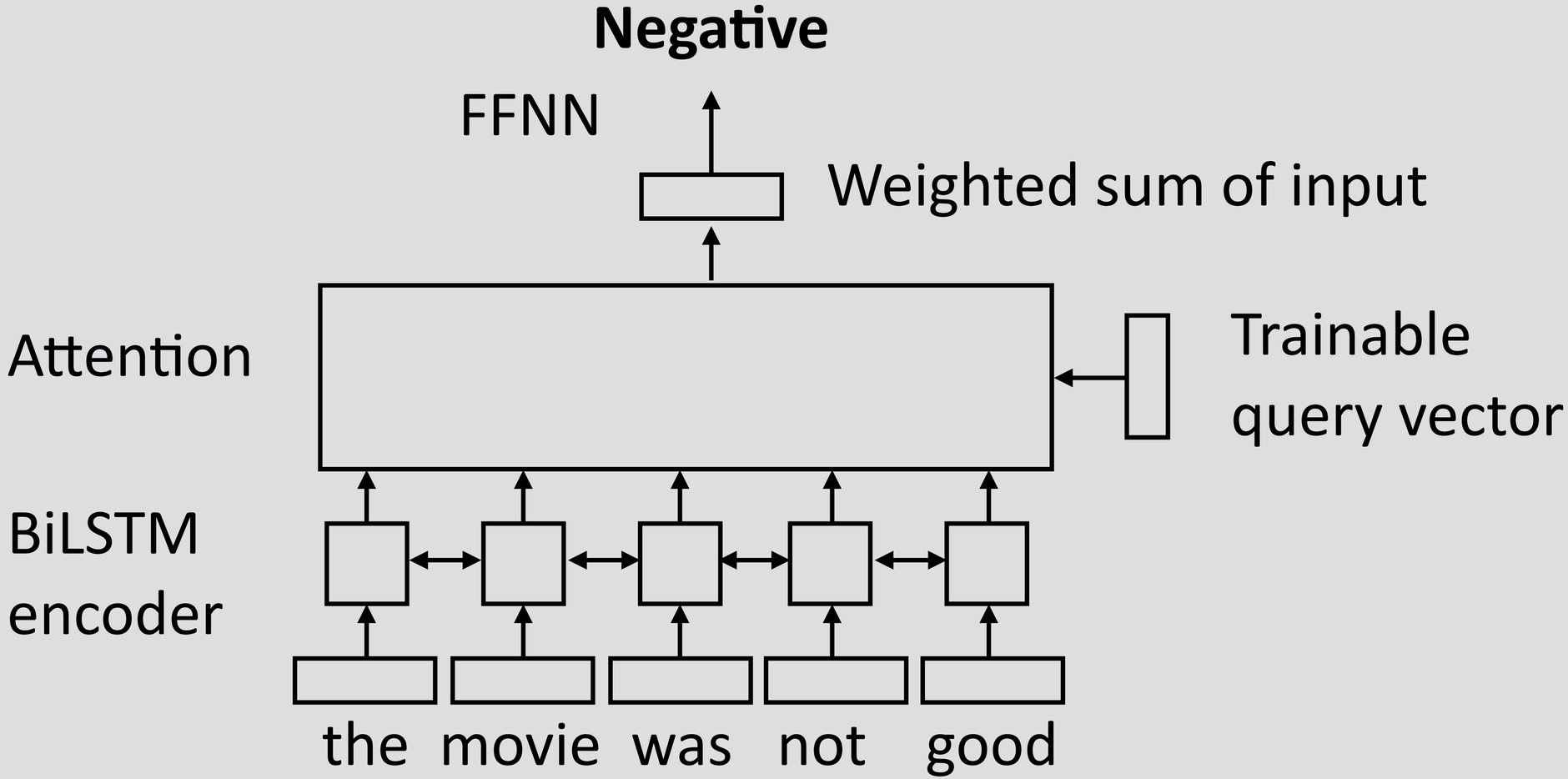
  - **Model probing:** auxiliary tasks, challenge sets, adversarial examples to understand more about how our model works

Lipton (2016); Belinkov and Glass (2018)

Local Explanations

(which parts of the input were responsible for the model's prediction on this particular data point?)

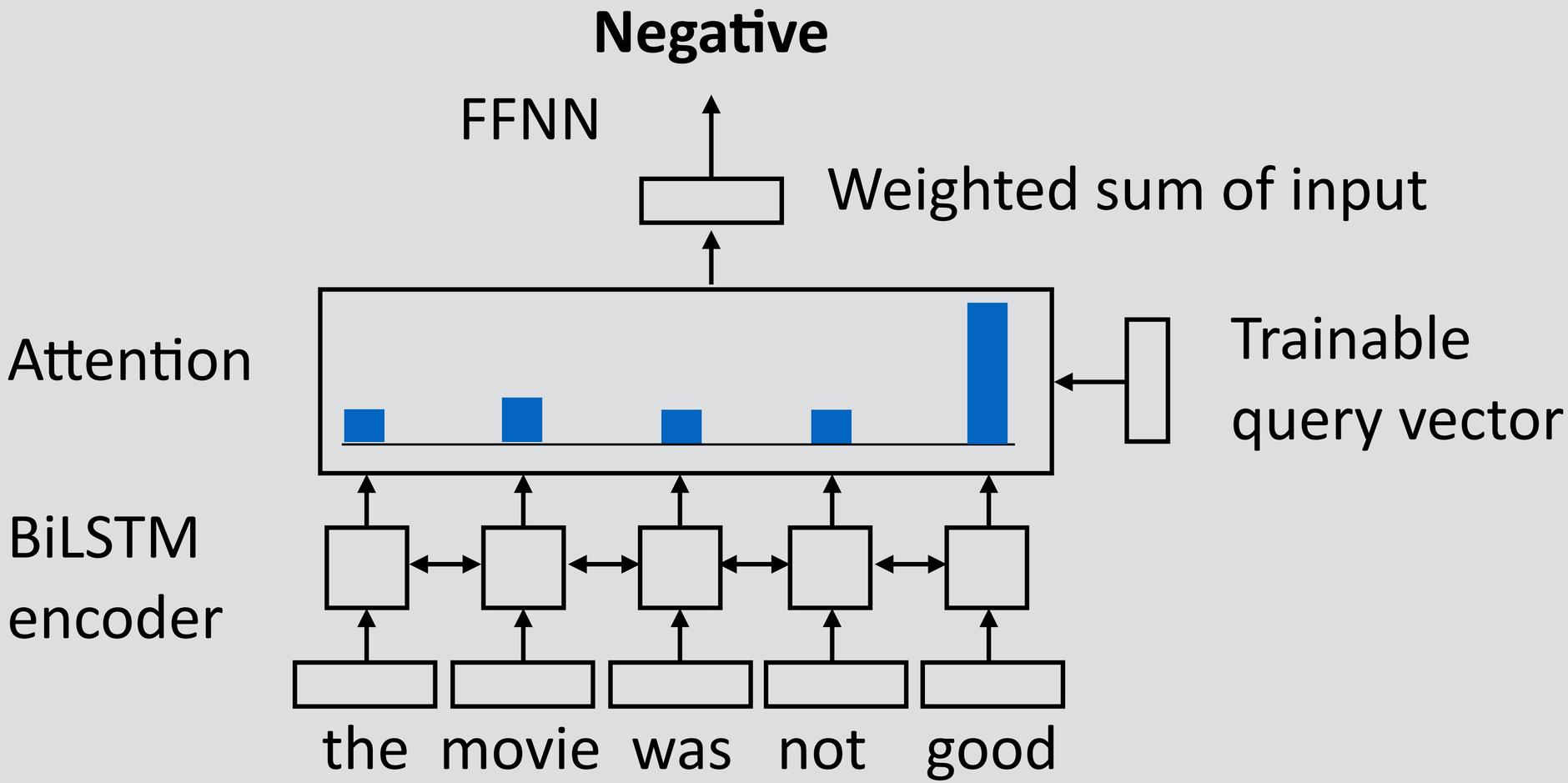
Sentiment Analysis with Attention



- Similar to a DAN model, but (1) extra BiLSTM layer; (2) attention layer instead of just a sum

Jain and Wallace (2019)

Attention Analysis



- Attention places most mass on *good* — did the model ignore *not*?
- What if we removed *not* from the input?

Jain and Wallace (2019)

Local Explanations

- An explanation could help us answer counterfactual questions: if the input were ***x***' instead of ***x***, what would the output be?

	Model
that movie was not great , in fact it was terrible !	—
that movie was not ____ , in fact it was terrible !	—
that movie was ____ great , in fact it was ____ !	+

- Attention can't necessarily help us answer this!

Will get to it shortly!

Erasure Method

- Delete each word one by and one and see how prediction prob changes

that movie was not great , in fact it was terrible ! — prob = 0.97

___ movie was not great , in fact it was terrible ! — prob = 0.97

that ____ was not great , in fact it was terrible ! — prob = 0.98

that movie ____not great, in fact it was terrible ! — prob = 0.97

that movie was ___ great, in fact it was terrible ! — prob = 0.8

that movie was not ____, in fact it was terrible ! — prob = 0.99

Erasure Method

- Output: highlights of the input based on how strongly each word affects the output

  that movie was not great , in fact it was terrible !

  - *not* contributed to predicting the negative class (removing it made it less negative), *great* contributed to predicting the positive class (removing it made it more negative)

- Will this work well?
  - Inputs are now unnatural, model may behave in “weird” ways
  - Saturation: if there are two features that each contribute to negative predictions, removing each one individually may not do much

LIME

- Locally-interpretable, model-agnostic explanations (LIME)

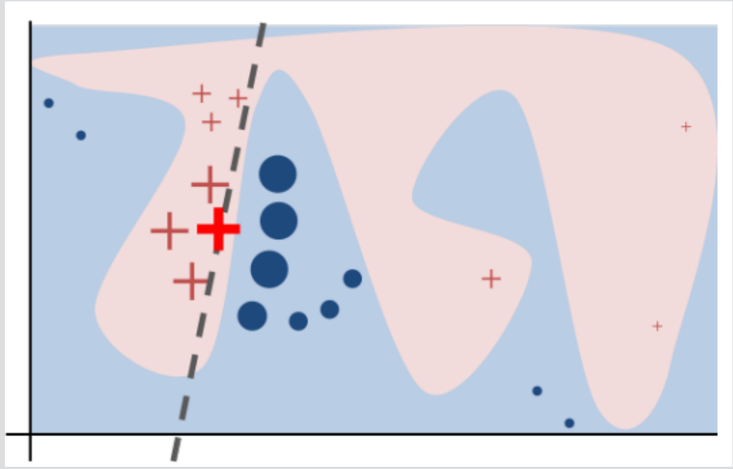
- Similar to erasure method, but we're going to delete collections of things at once
  - Can lead to more realistic input (although people often just delete words with it)
  - More scalable to complex settings

Ribeiro et al. (2016)

LIME

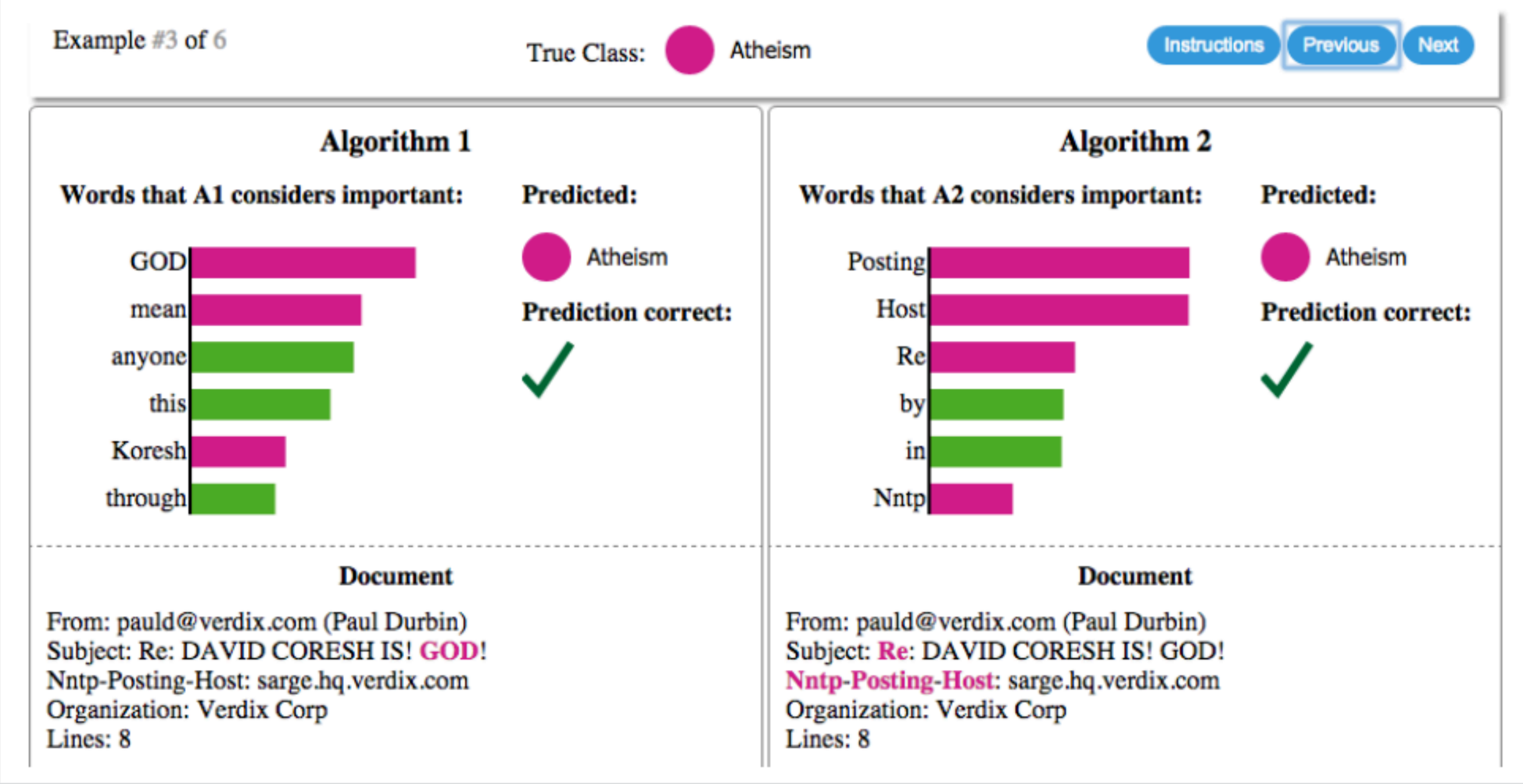
Key ideas:

- Assume a target model: *x* -> *f(x)*
- Model-agnostic explainer (a linear classifier): *x'* -> *g(x')*
  - Where *x'* is the "interpretable" feature representation of the instance of *x*
  - Learning *g* such that *g* replicates *f*, then interpreting *g* (e.g., checking its weights on features) is explaining *f*
- *Local* explanation
  - More feasible than *global* explanation

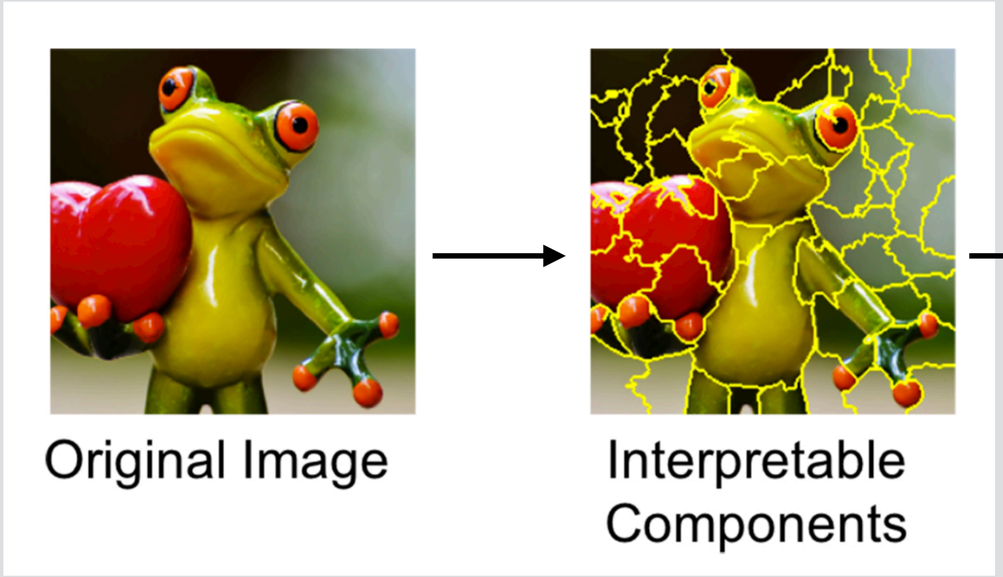


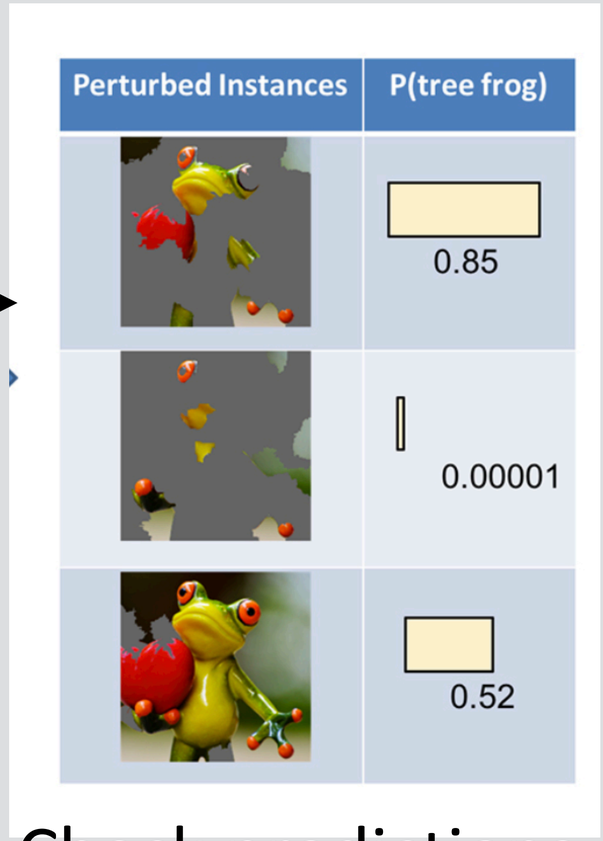
LIME

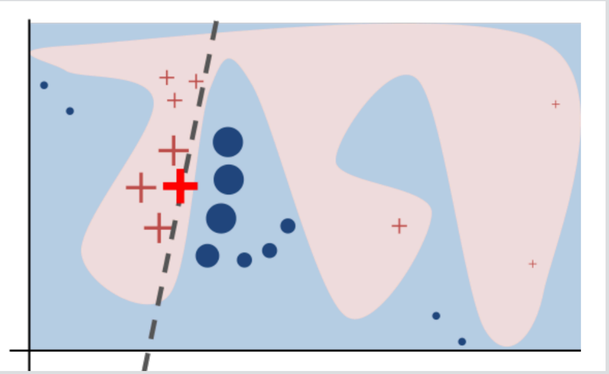
Example: predicting if a document is about “Christianity” or “Atheism”



LIME



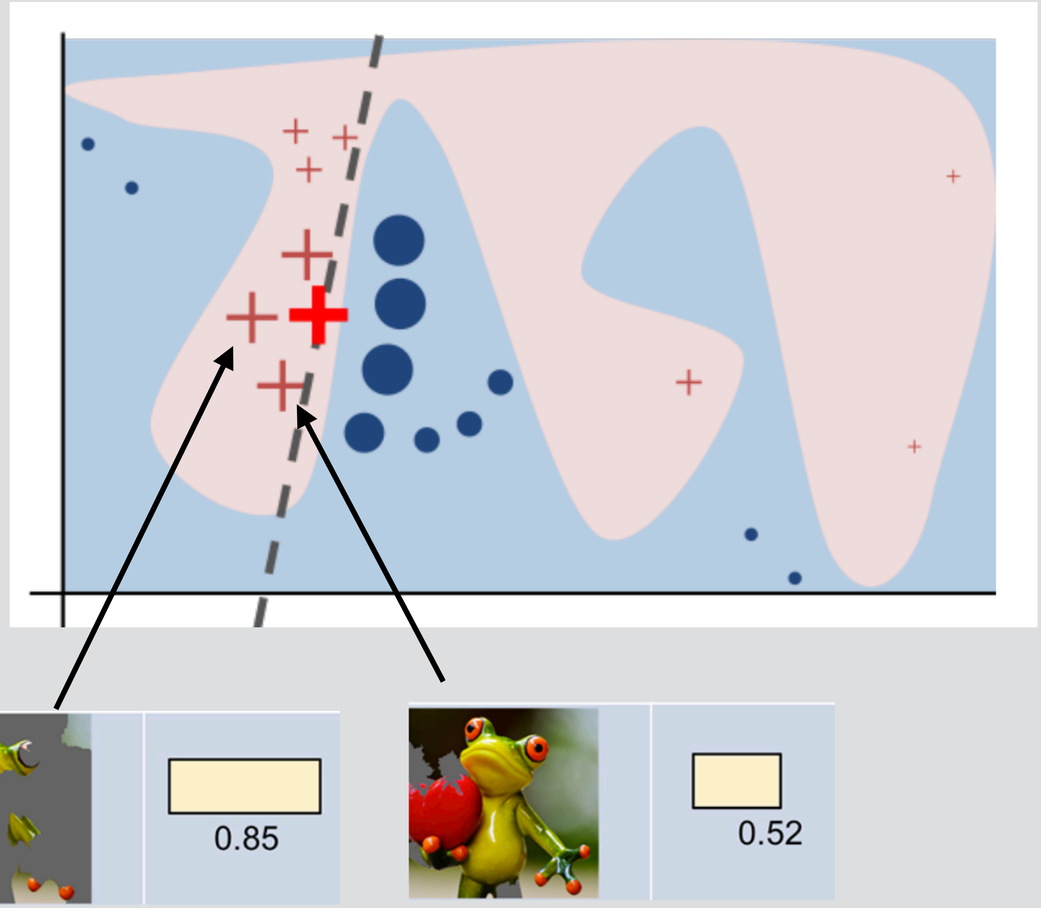




- Break input into components (for text: could use words, phrases, sentences, …)
- Check predictions on subsets of those
- Now we have model predictions on perturbed examples

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

LIME (cont'd)



- This is what the model is doing on perturbed examples of the input
- Now we train a classifier to predict **the model's behavior** based on **what subset of the input it sees**
- The weights of that classifier tell us which interpretable components of the input are important

LIME (cont'd)

- This secondary classifier's **weights** now give us highlights on the input

The movie is mediocre, maybe even bad.	**Negative** 99.8%
The movie is mediocre, maybe even ~~bad~~.	**Negative** 98.0%
The movie is ~~mediocre~~, maybe even bad.	**Negative** 98.7%
The movie is ~~mediocre~~, maybe even ~~bad~~.	**Positive** 63.4%
The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~.	**Positive** 74.5%
The ~~movie~~ is mediocre, maybe even ~~bad~~.	**Negative** 97.9%

The movie is mediocre, maybe even bad.

Wallace, Gardner, Singh
Interpretability Tutorial at EMNLP 2020

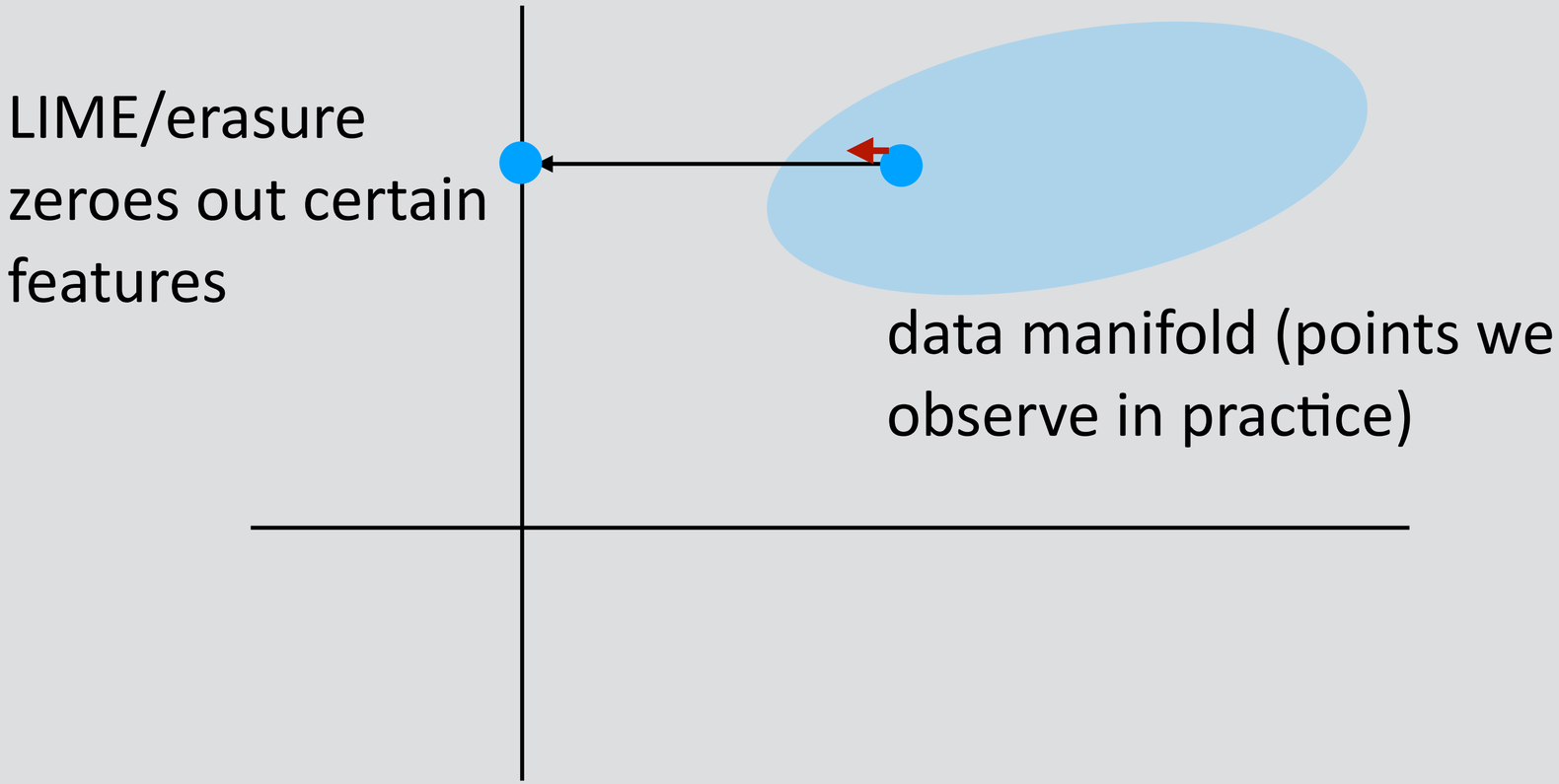
Problems with LIME

- Lots of moving parts here: what perturbations to use? what model to train? etc.
- Expensive to call the model all these times
- Linear assumption about interactions may not be reliable

Gradient-based Methods

Problems with LIME

- Problem: fully removing pieces of the input may cause it to be very unnatural



- Alternative approach: look at what this perturbation does locally right around the data point using gradients

Gradient-based Methods

score = weights * features
(or an NN)

Learning a model

Compute derivative of score with respect to weights: how can changing weights improve score of correct class?

Gradient-based Explanations

Compute derivative of score with respect to ***features***: how can changing ***features*** improve score of correct class?

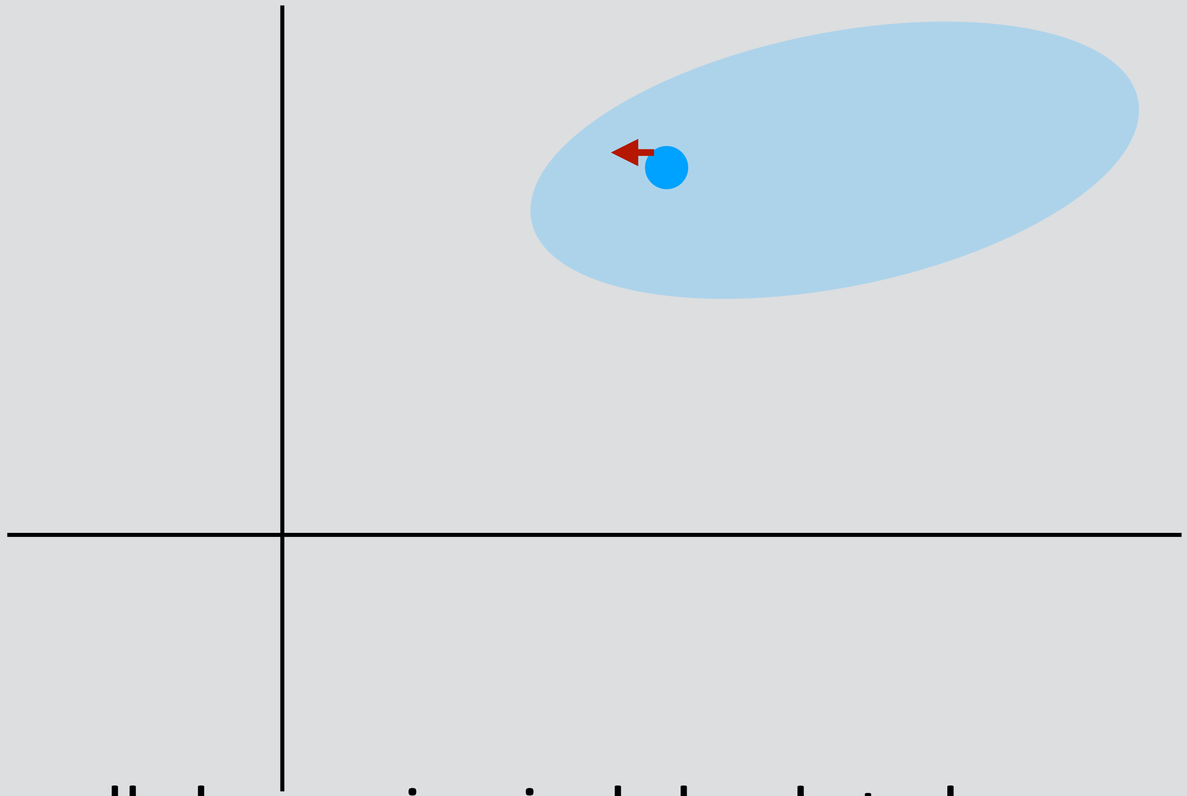
Gradient-based Methods

- Originally used for images

S_c = score of class c

I_0 = current image

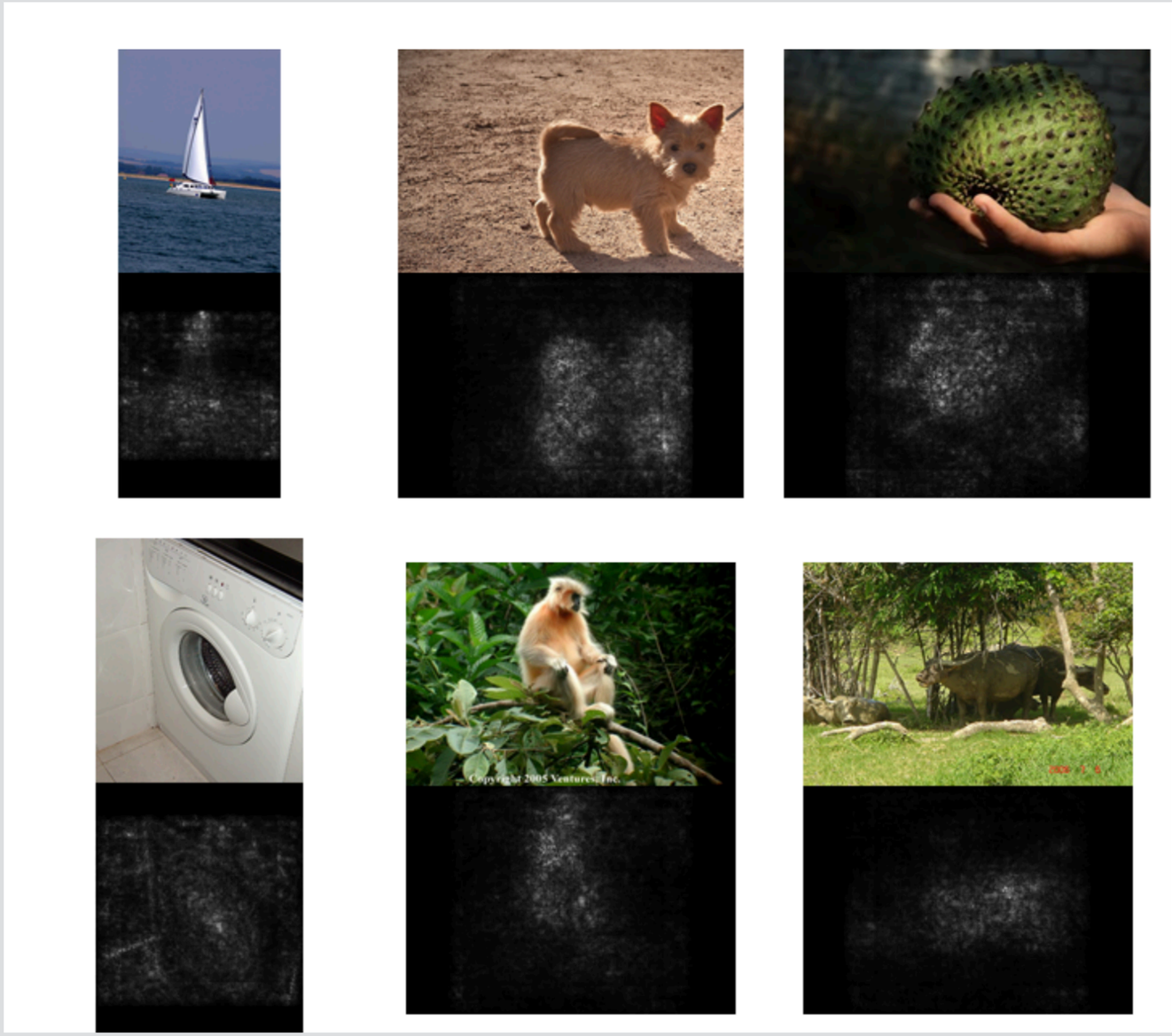
$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0}$$



- Higher gradient magnitude = small change in pixels leads to large change in prediction
- For words: “pixels” are coordinates of each word’s vector, sum these up to get the importance of that word

Simonyan et al. (2013)

Gradient-based Methods



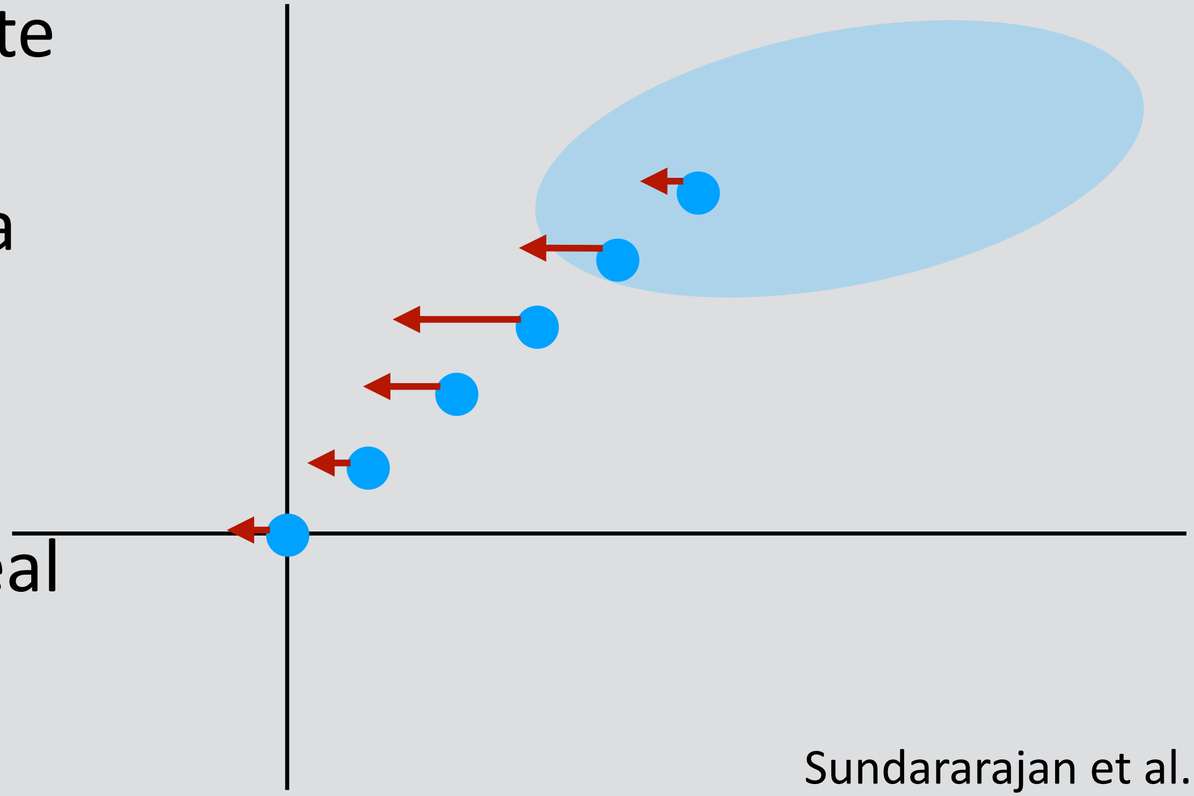
Simonyan et al. (2013)

Integrated Gradients

- Suppose you have prediction = A OR B for features A and B. Changing either feature doesn't change the prediction, but changing both would. Gradient-based method says neither is important

- Integrated gradients: compute gradients along a path from the origin to the current data point, aggregate these to learn feature importance

- Intermediate points can reveal new info about features



Sundararajan et al. (2017)

Integrated Gradients

$$\text{IntegratedGrads}_i^{approx}(x) ::= (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

Scale by total distance

Compute gradient at the *k*th point along the way w.r.t. the *i*th feature

Average over the m steps

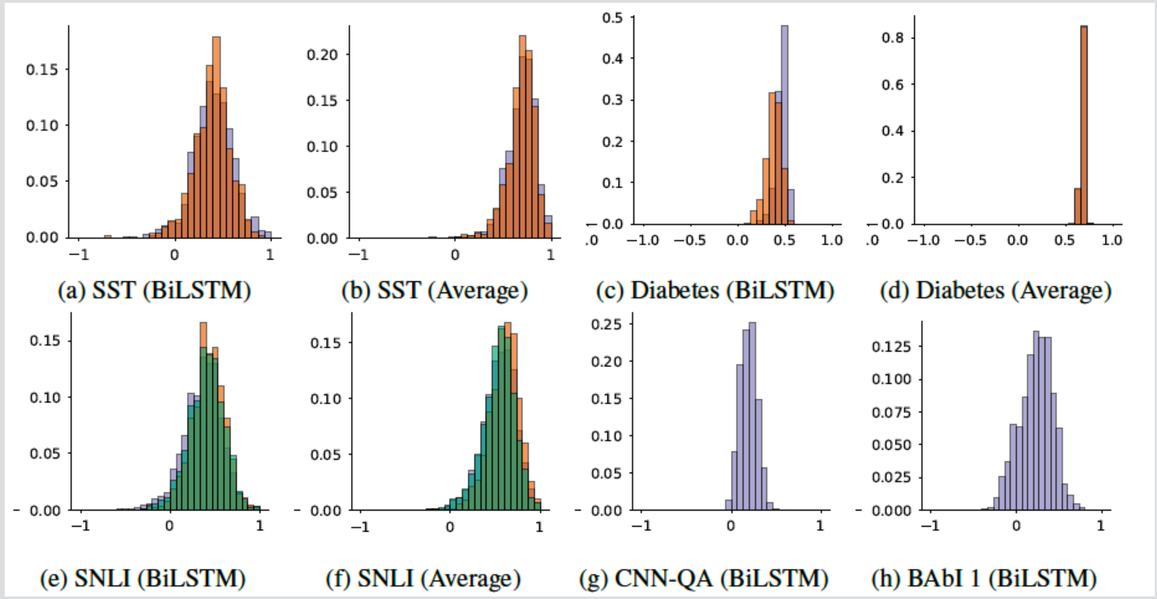
x'_i = "baseline" — all PAD or MASK tokens (MASK usually works better)

- Can be expensive: requires calling forward() and backward() at *m* steps along the way

Sundararajan et al. (2017)

Is Attention Good Explanation?

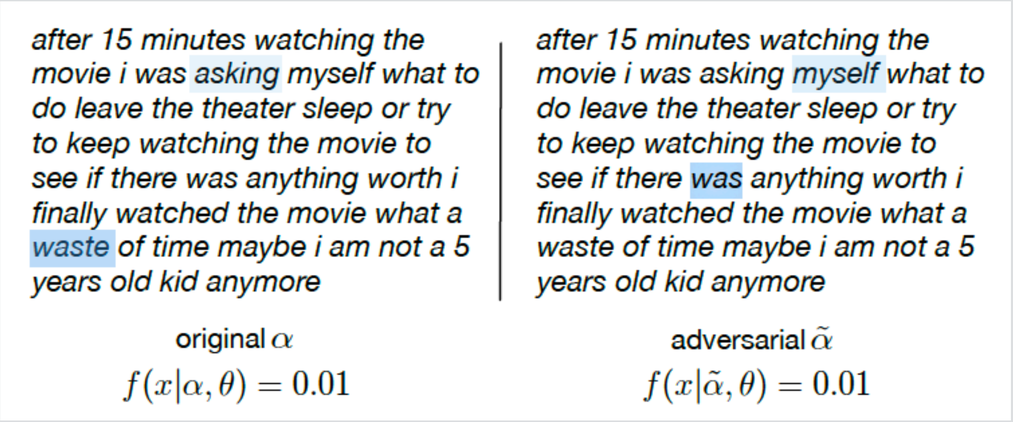
- Why? Because, ideally, we expect that a model learns to attend to important parts of the input when it makes decisions
- Jain and Wallace (2019): Attention is *not* Explanation
  - Experimental setting: binary text classification, QA, NLI
  - Observation 1: poor correlation of attention with gradient/leave-one-out importance measures



Is Attention Good Explanation?

- Why? Because, ideally, we expect that a model learns to attend to important parts of the input when it makes decisions
- Jain and Wallace (2019): Attention is *not* Explanation
  - Experimental setting: binary text classification, QA, NLI
  - Observation 1: poor correlation of attention with gradient/leave-one-out importance measures
  - Observation 2: permuted or adversarially searched attention distributions may not change the prediction significantly

Are the conclusions convincing?



Is Attention Good Explanation?

- Why? Because, ideally, we expect that a model learns to attend to important parts of the input when it makes decisions
- Wiegreffe and Pinter (2019): Attention is ***not*** *not* Explanation
  - Same experimental setting as Jain and Wallace (2019)
  - Major claim: the attention does encode useful information that the model relies to make decisions
  - Flaws in Jain and Wallace's experiments:
    - Some datasets are too easy, so attention is not used at all
    - Should get rid of the inherent model variance (as a "baseline")
    - Should test the assumption on simpler models (e.g., FFNN rather than LSTM)

Other Explanation Techniques

- A highly recommended book: Interpretable Machine Learning by Christoph Molnar (online)
- Local explanations
  - LIME, Gradient-based methods
  - DeepLIFT: modeling the reference/baseline; importance with directions (pos vs. neg). Trying to address issues with existing methods.
  - Shapley value: from the game theory, typically with sampling tricks.
  - (Kernel) SHAP: a linear LIME with different weighting kernel and loss function, which theoretically recovers the Shapley value.

Text Explanations

Explanations of Bird Classification

Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

- An explanation should be relevant to both the class and the image
- Are these features *really* what the model used?

Hendricks et al. (2016)

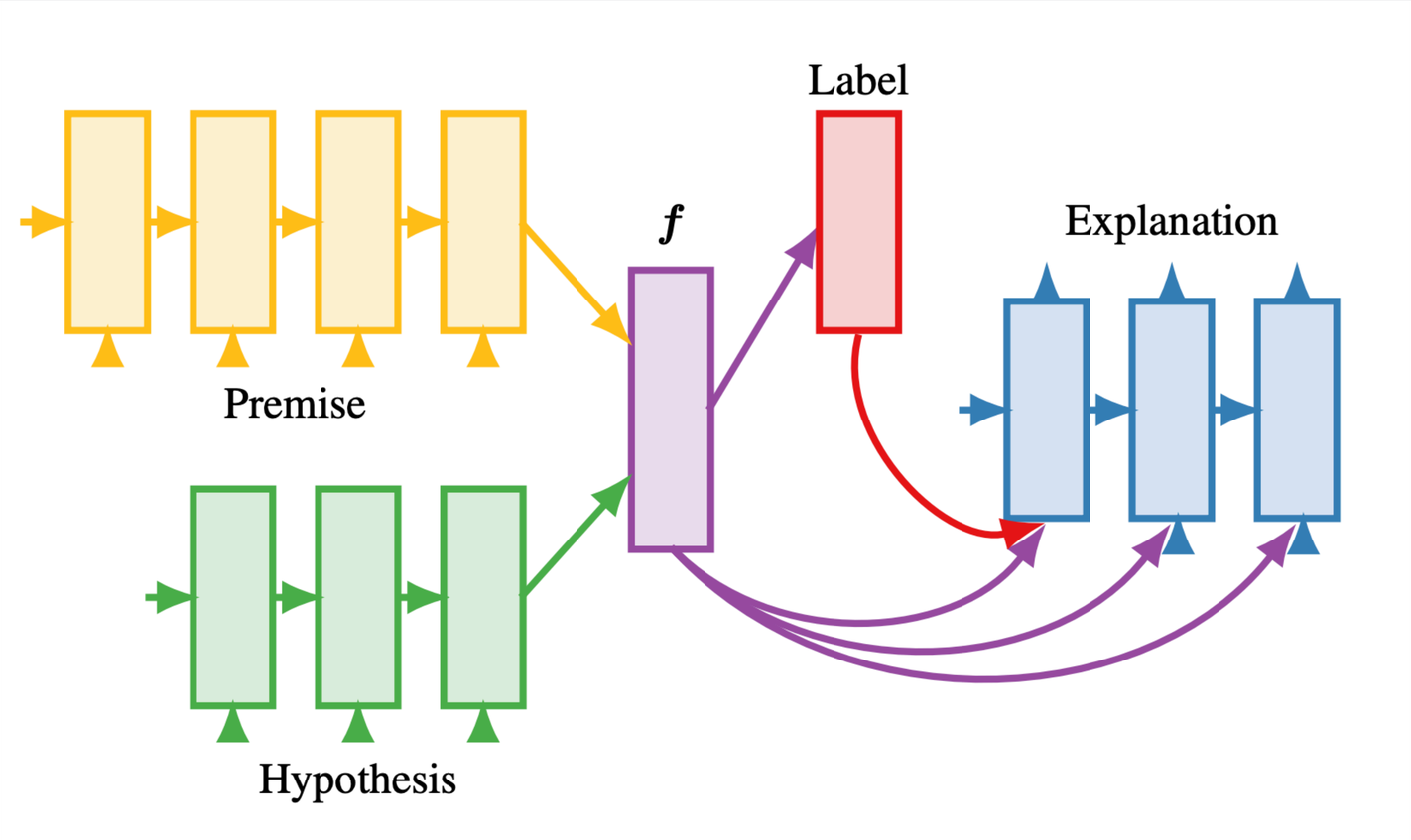
Explanations of NLI

Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

- How do we use this information? If we produce a network to predict it, does that make it an actual explanation of what's happening?

Camburu et al. (2019)

Explanations of NLI



- Information from *f* is fed into the explanation LSTM, but **no constraint that this must be used**. Different coordinates from *f* could predict label and explanations
- But does not seem to help the NLI task. In fact, low inter-agreement among annotators as well.

Evaluating Explanations

Faithfulness vs. Plausibility

- Suppose our model is a bag-of-words model with the following:

  the = -1, movie = -1, good = +3, bad =0

  the movie was good prediction score=+1

  the movie was bad prediction score=-2

- Suppose explanation returned by LIME is:

  the movie was good

  the movie was bad

- Is this a “correct" explanation?

Faithfulness vs. Plausibility

- *Plausible* explanation: matches what a human would do

  the movie was good the movie was bad

  - Maybe useful to explain a task to a human, but it's not what the model is really doing!

- *Faithful* explanation: actually reflects the behavior of the model

  the movie was good the movie was bad

  - We usually prefer faithful explanations; non-faithful explanations are actually deceiving us about what our models are doing!
  - Rudin: *Stop Explaining Black Box Models for High-Stakes Decisions and Use Interpretable Models Instead*

AllenNLP Demo

- https://demo.allennlp.org/sentiment-analysis/glove-sentiment-analysis