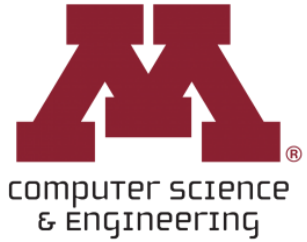
CSCI 5541: Natural Language Processing

Project Guideline

https://jimtmooney.github.io/Courses/S25/hw/Project_Description.pdf

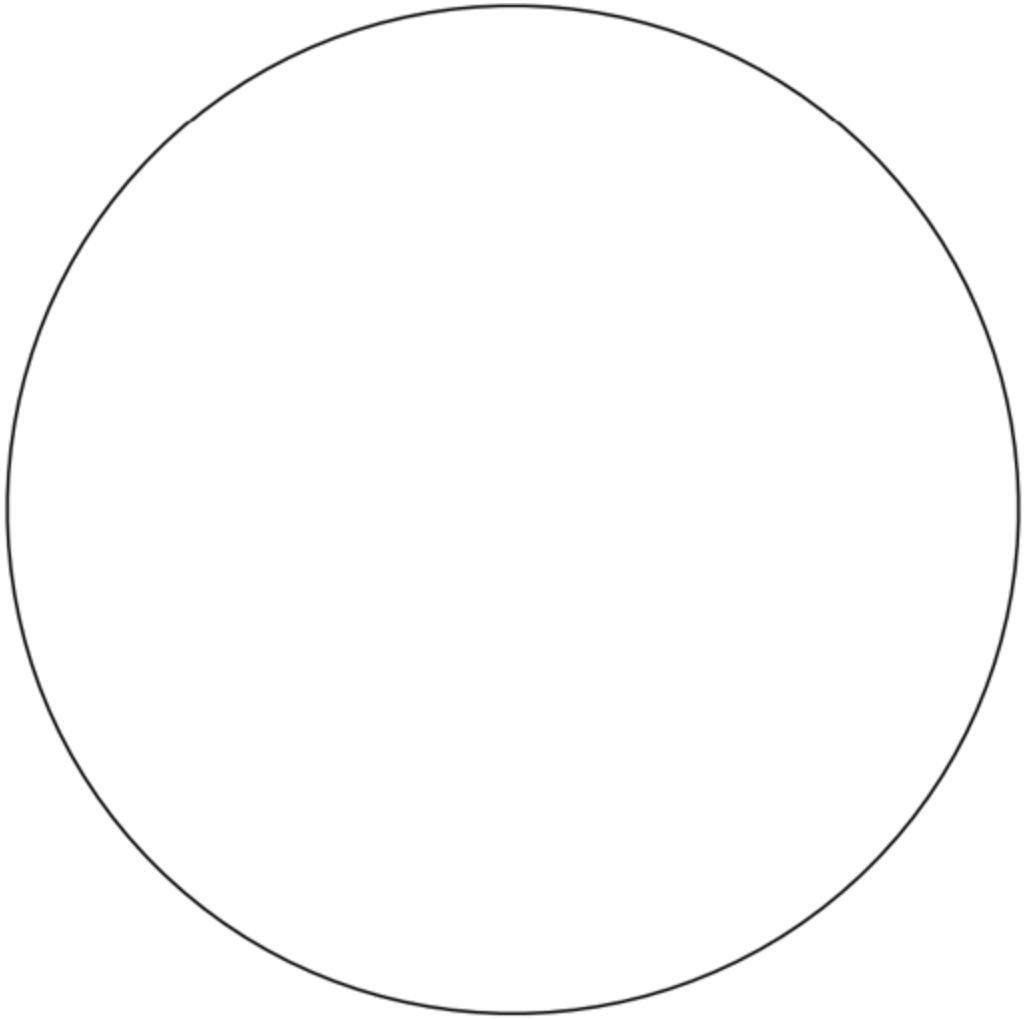


UNIVERSITY OF MINNESOTA
Driven to Discover®

Outline

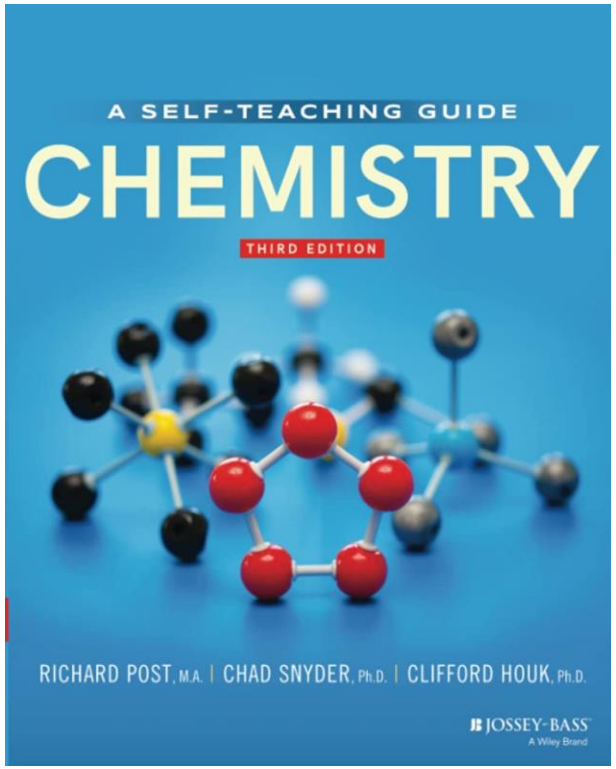
- ❑ Project Goal
- ❑ Project Deliveries and Due
- ❑ Some advices for successful projects
- ❑ Project Types and Topics
- ❑ Past Projects

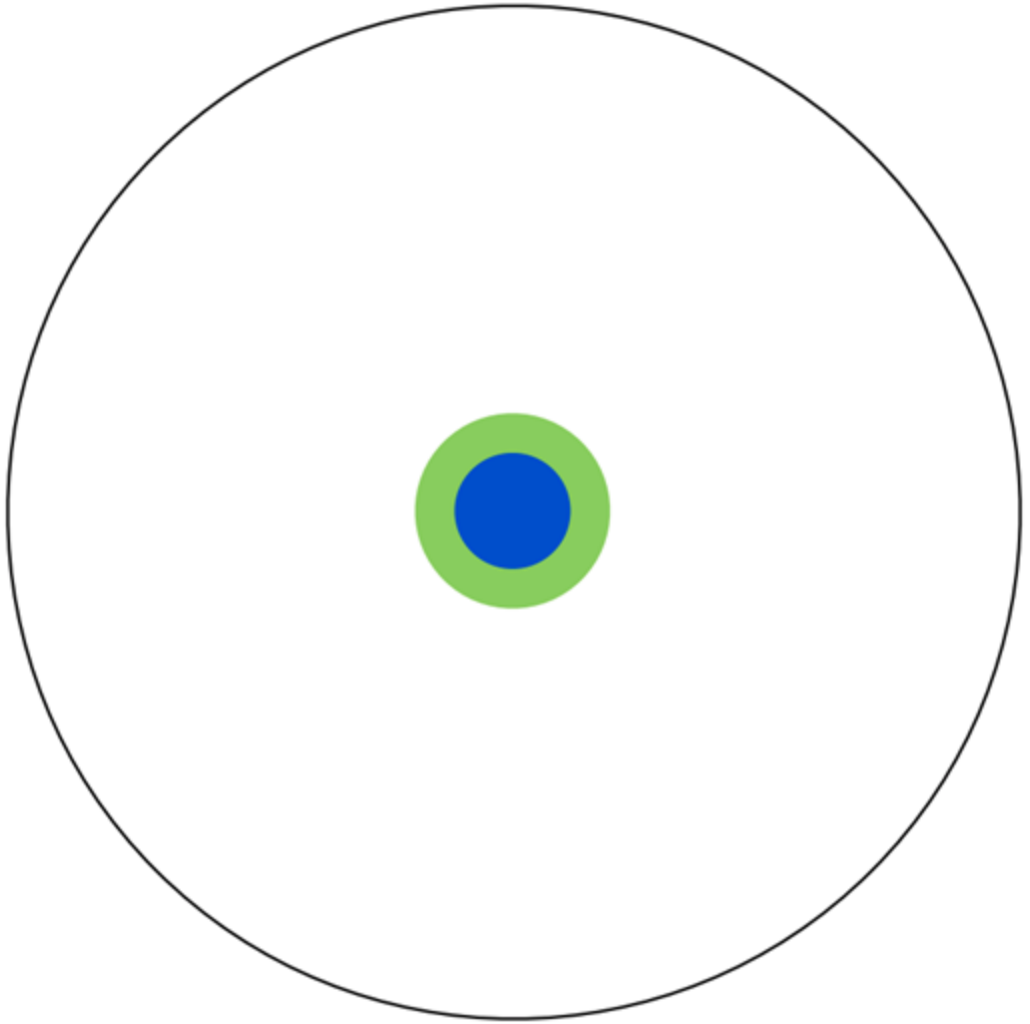
Imagine a circle that contains all of human knowledge:



The illustrated guide to a Ph.D. Matt Might

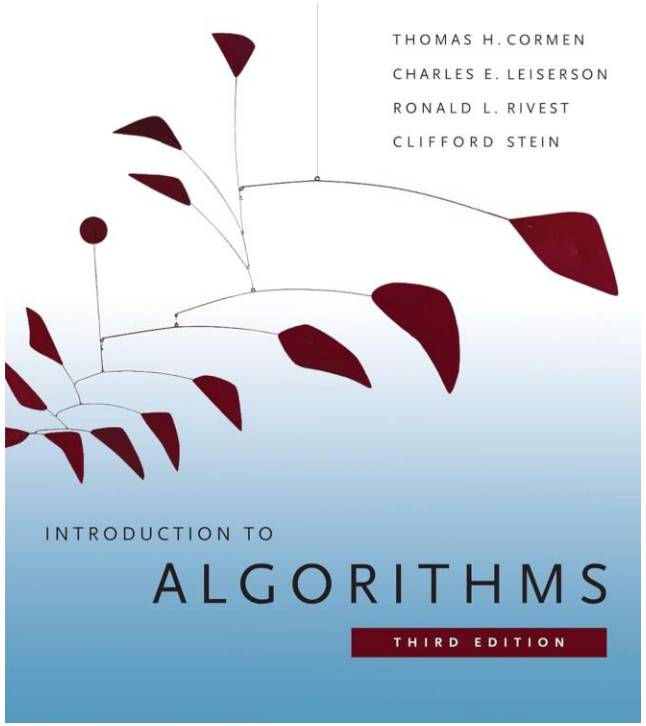
By the time you finish elementary school, you know a little.
By the time you finish high school, you know a bit more:

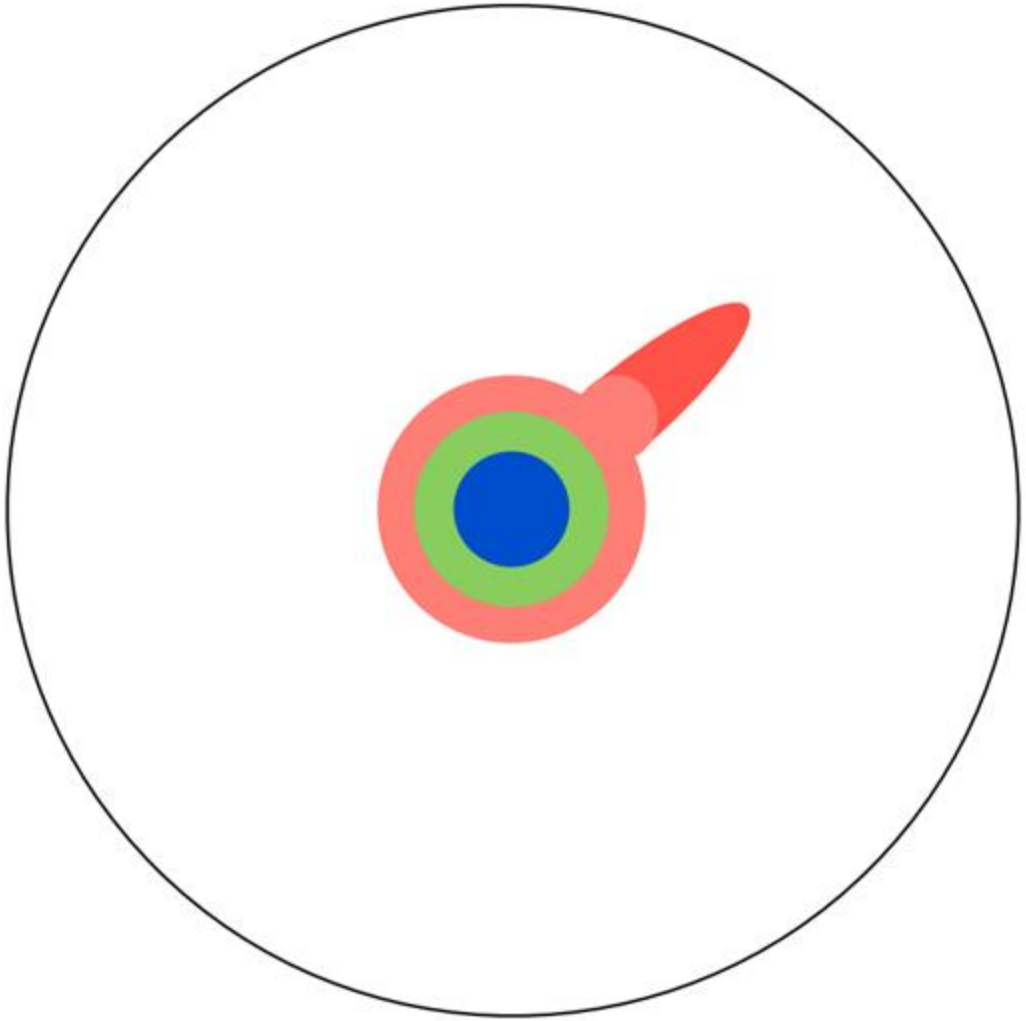


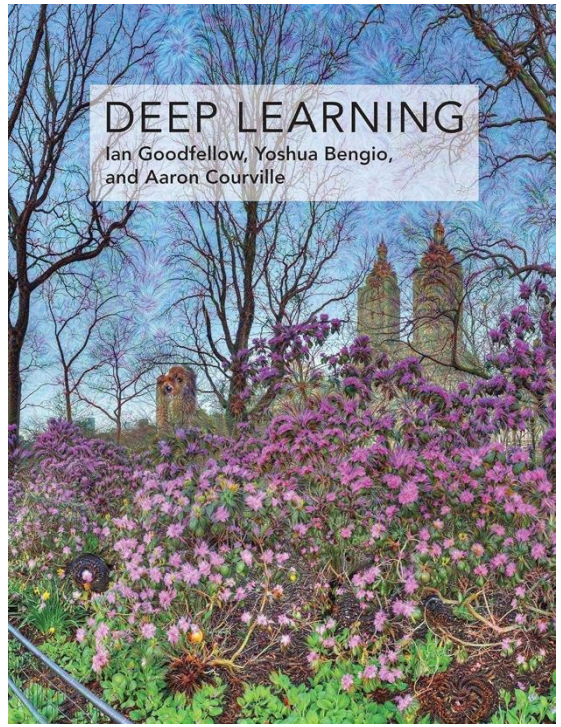


The illustrated guide to a Ph.D. Matt Might

With a bachelor's degree, you gain a specialty:
A master's degree deepens that specialty:

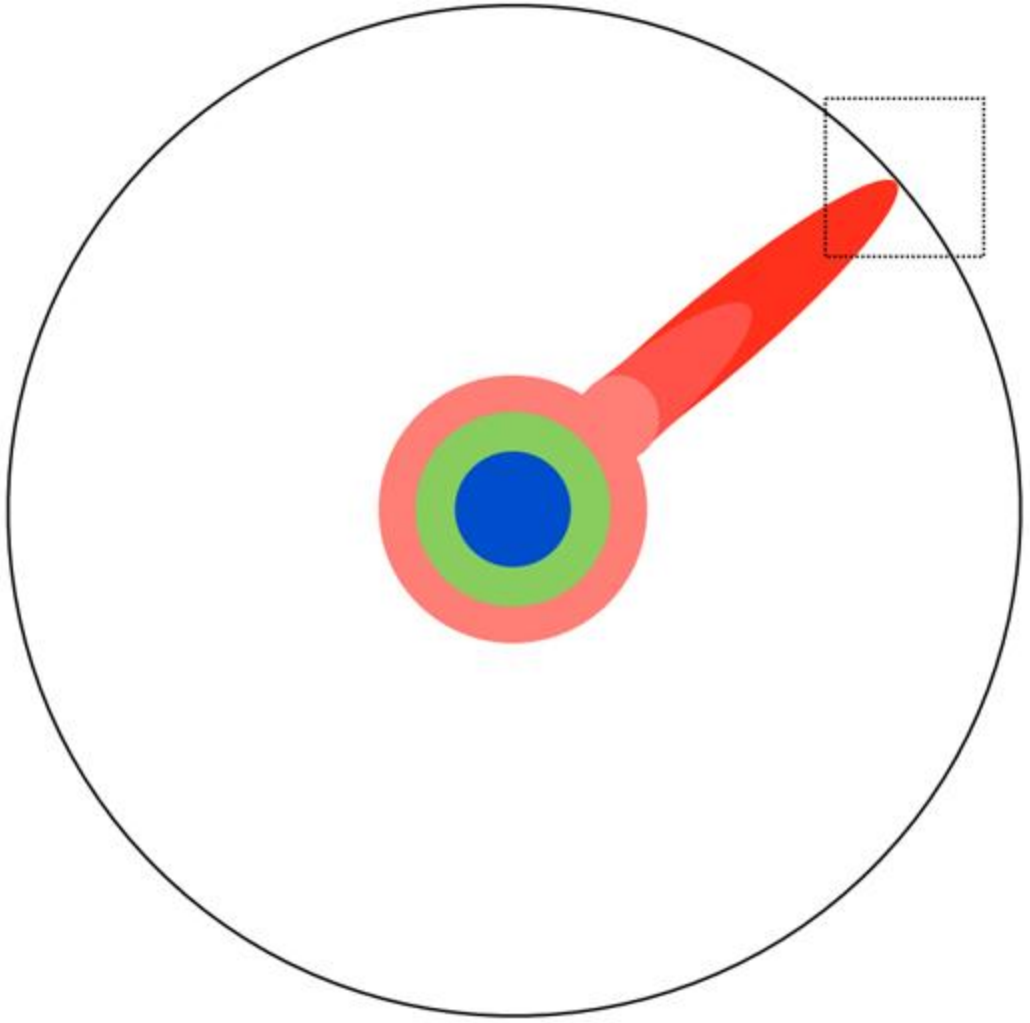






The illustrated guide to a Ph.D. Matt Might

Reading research papers takes you to the edge of human knowledge:
Once you're at the boundary, you focus:



Attention Is All You Need

Ashish Vaswani[*] Google Brain avaswani@google.com
Noam Shazeer[*] Google Brain noam@google.com
Niki Parmar[*] Google Research nikip@google.com
Jakob Uszkoreit[*] Google Research usz@google.com
Llion Jones[*] Google Research llion@google.com
Aidan N. Gomez[*] [†] University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser[*] Google Brain lukaszkaiser@google.com
Illia Polosukhin[*] [‡] illia.polosukhin@gmail.com

Abstract

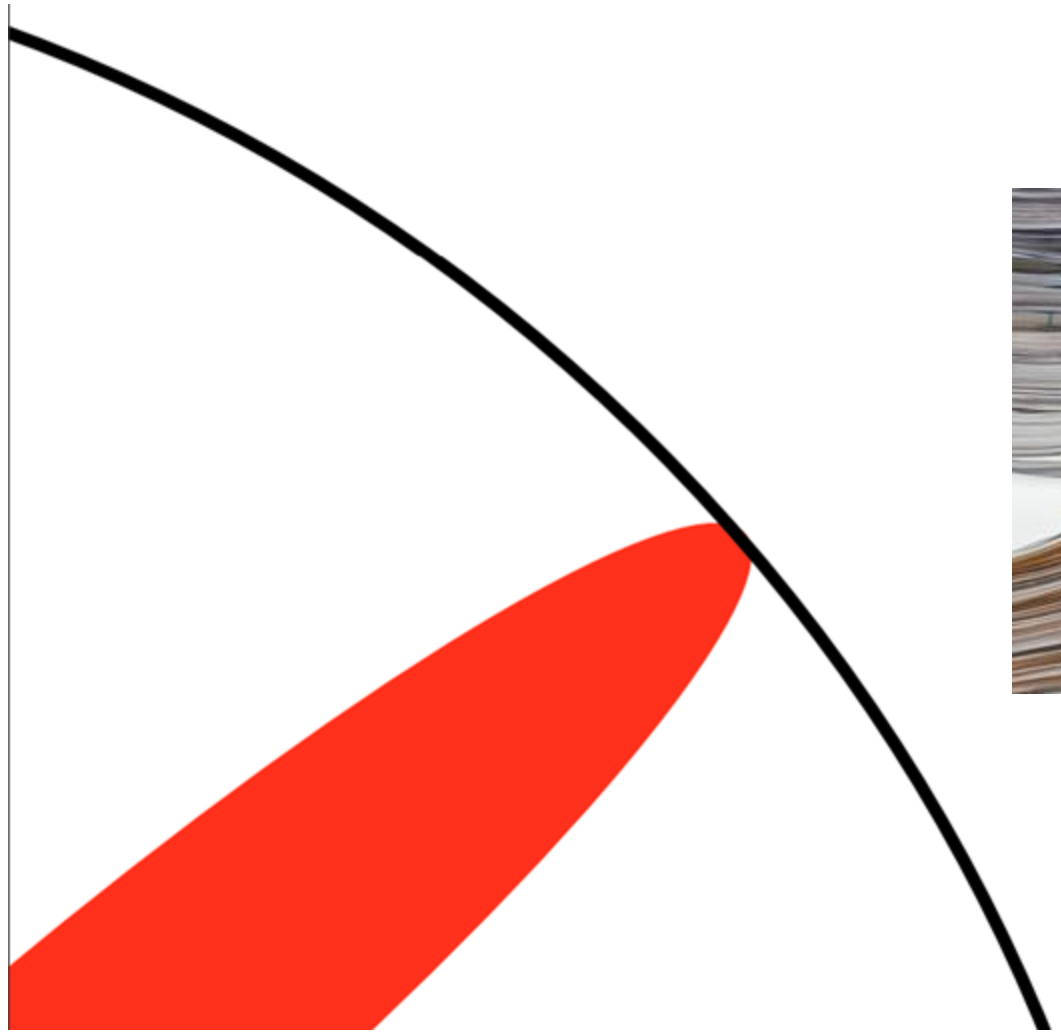
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

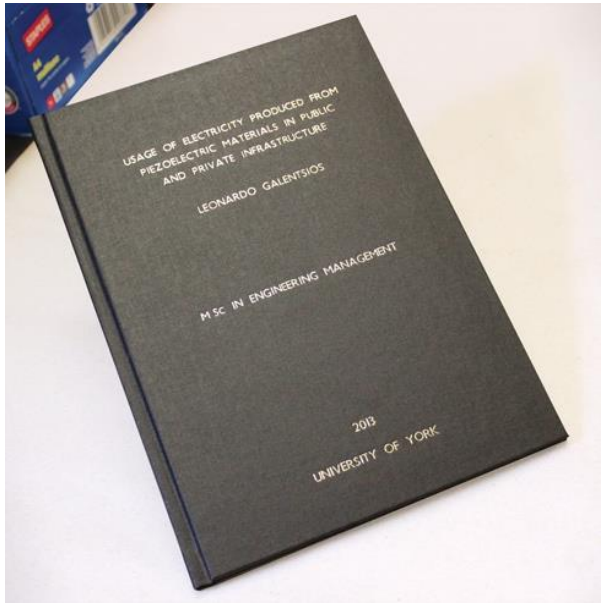
The illustrated guide to a Ph.D. Matt Might

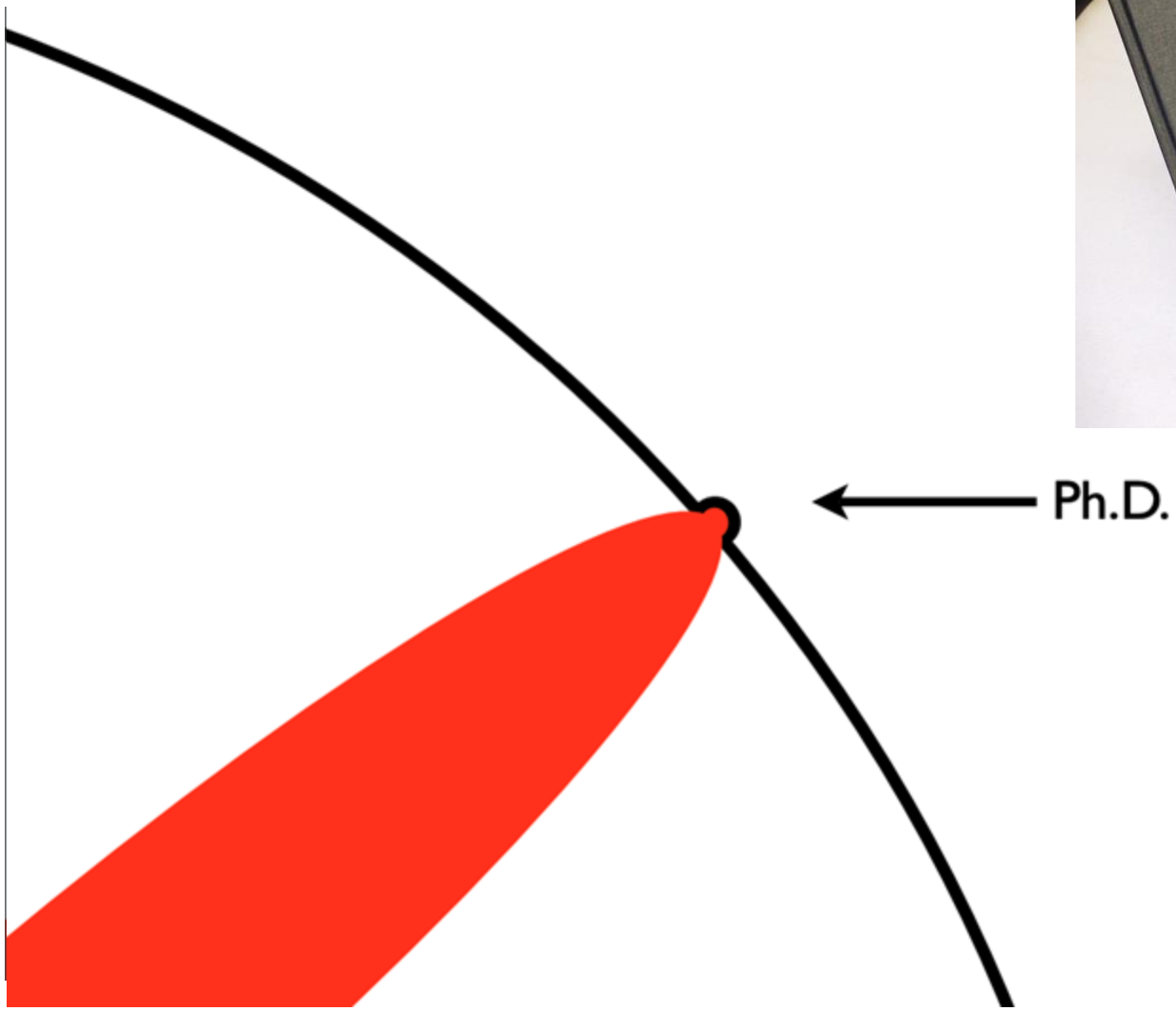
You push at the boundary for a few years:
Until one day, the boundary gives way:



The illustrated guide to a Ph.D. Matt Might

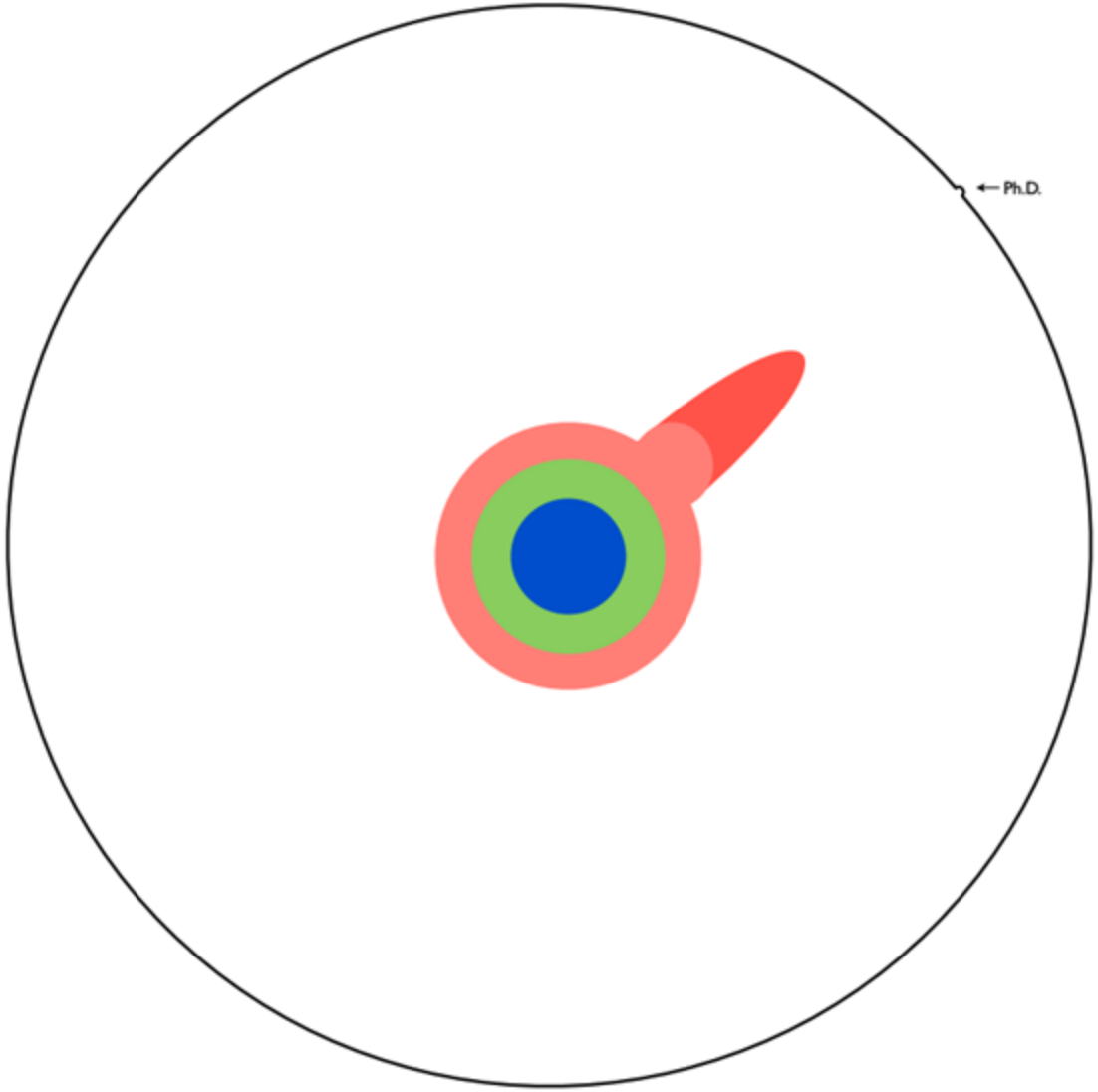
You push at the boundary for a few years:
Until one day, the boundary gives way:
And, that dent you've made is called a **Ph.D.:**

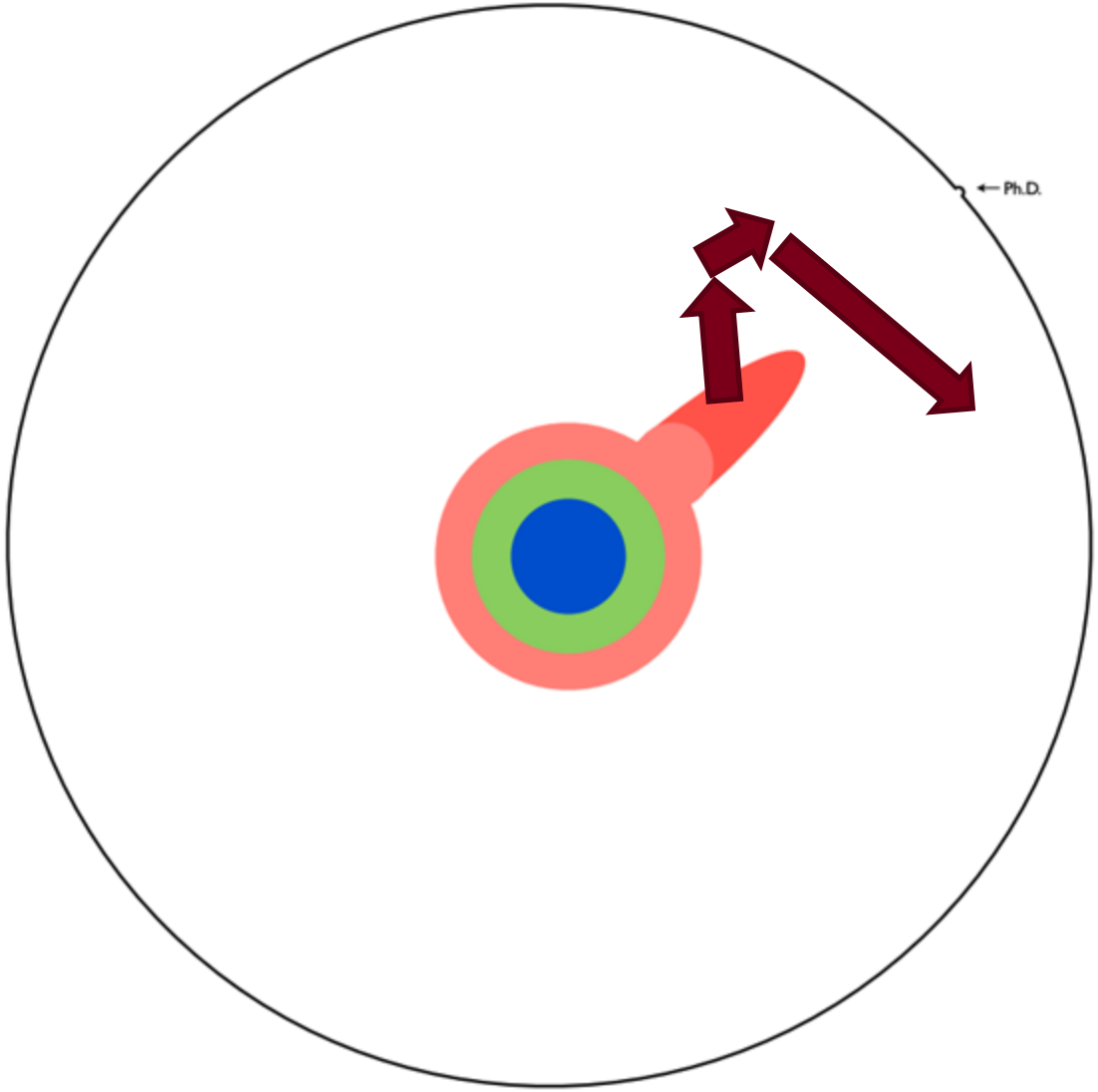




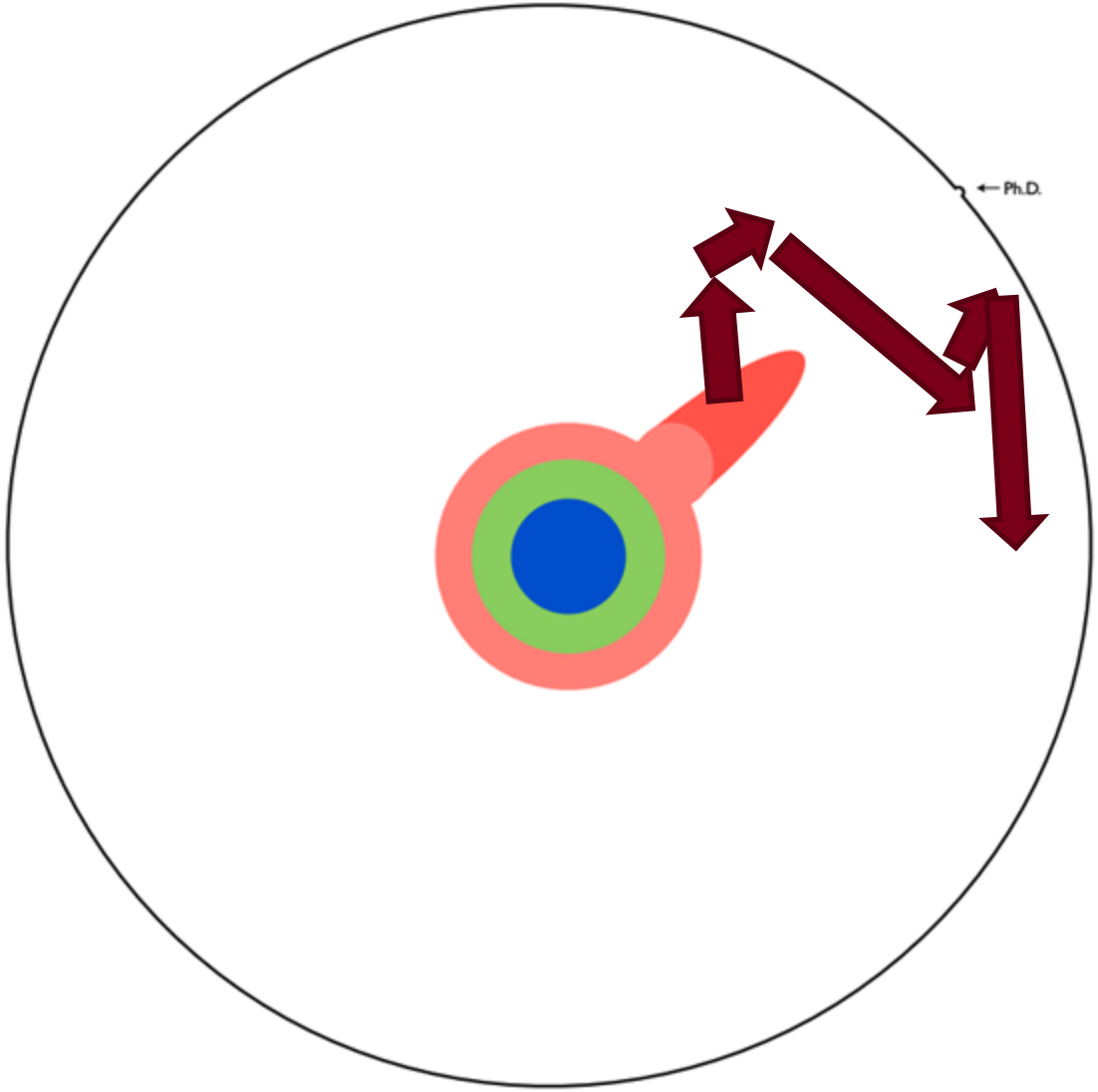
The illustrated guide to a Ph.D. Matt Might

Where you are now!

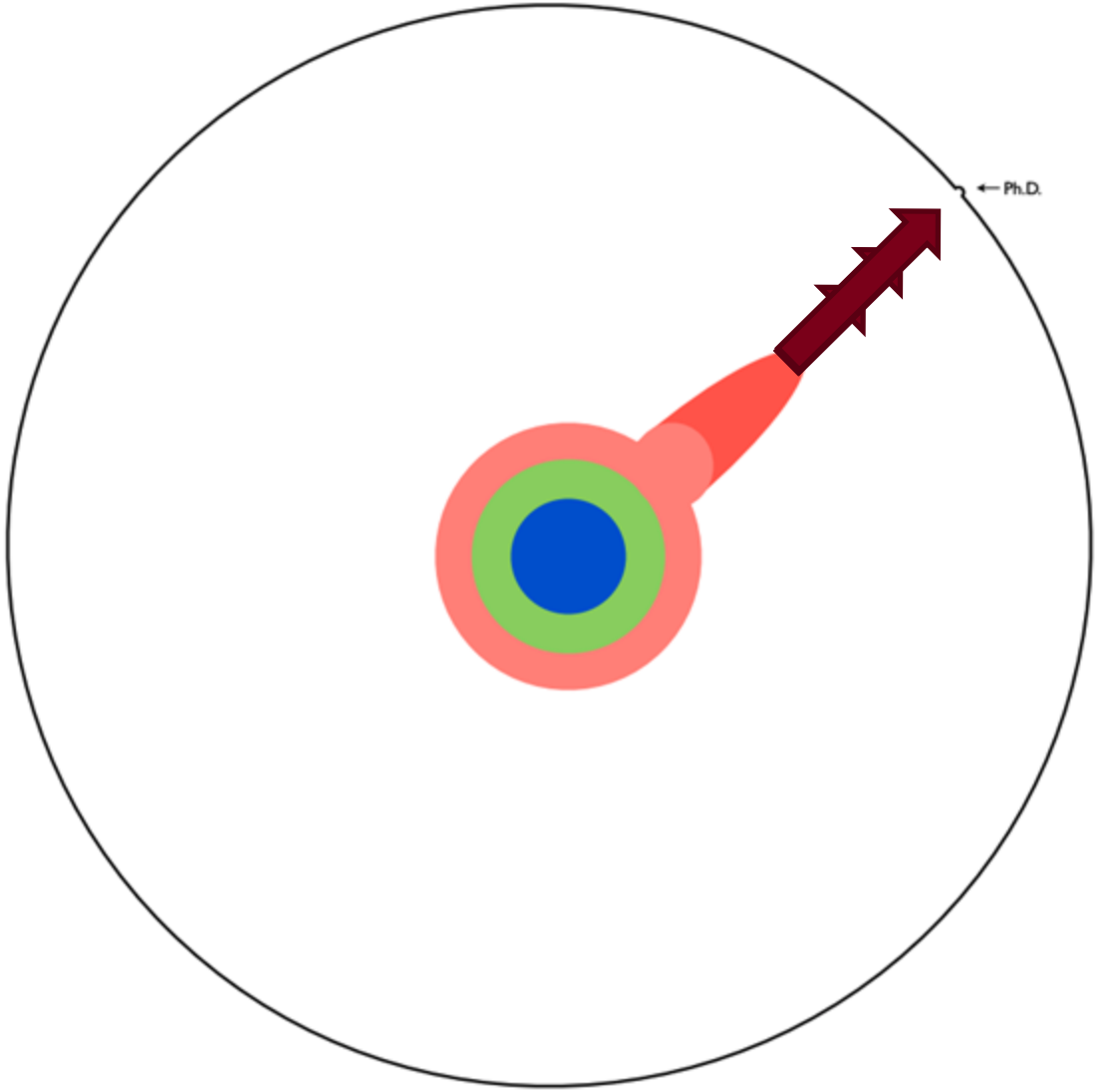



← Ph.D.

What we can help you on!



Your project goal: experience the NLP research

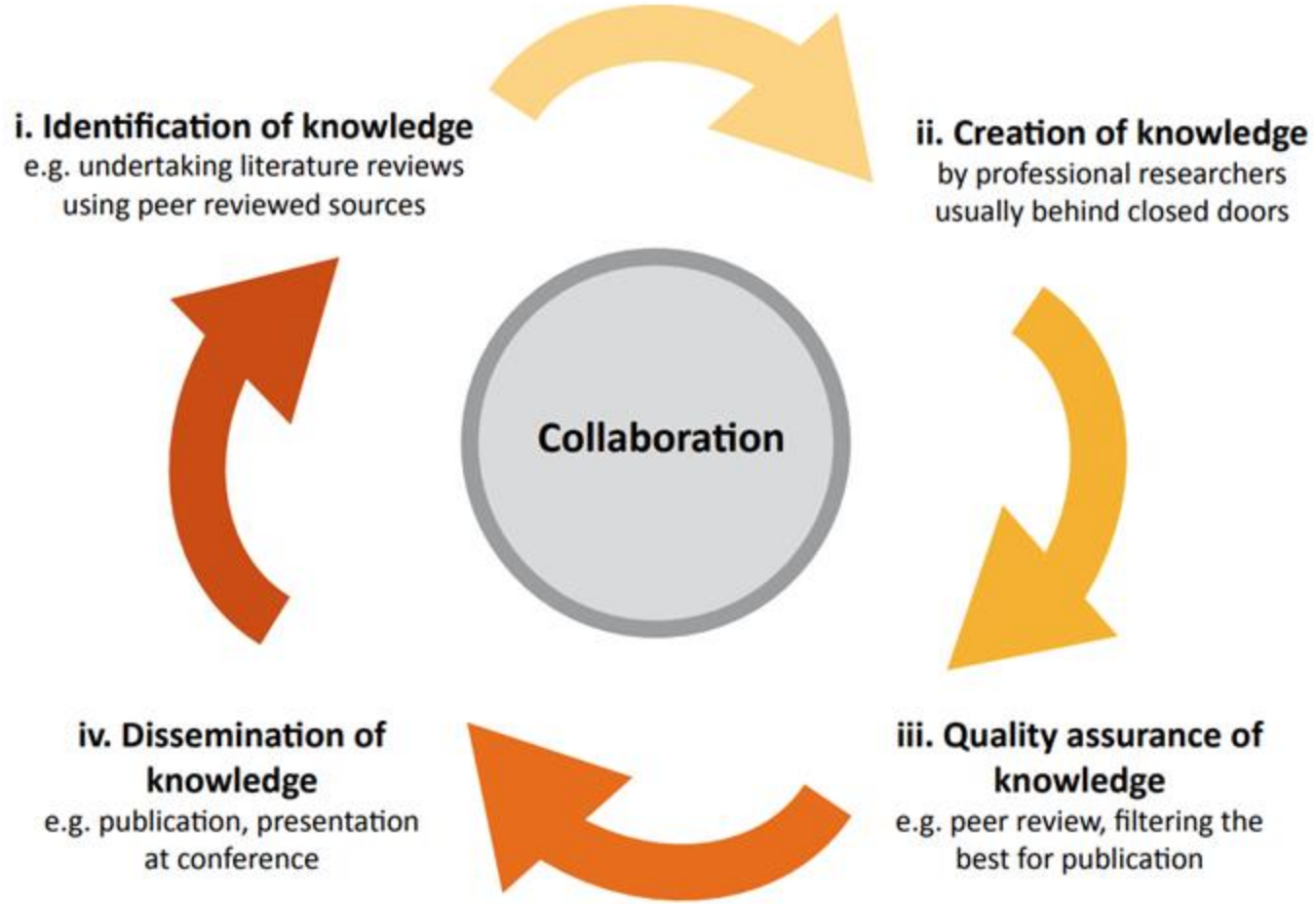


Project Goal (30%)

- Experience a full pipeline of NLP research
  - Proposal, research, presentation, feedback, etc
- Good time to interact with other researchers in NLP
  - Your team members, instructors, and mentors
- You can make your project as an *extension of your homework* but there should be novel extension and research contribution.

Academic Research Cycle

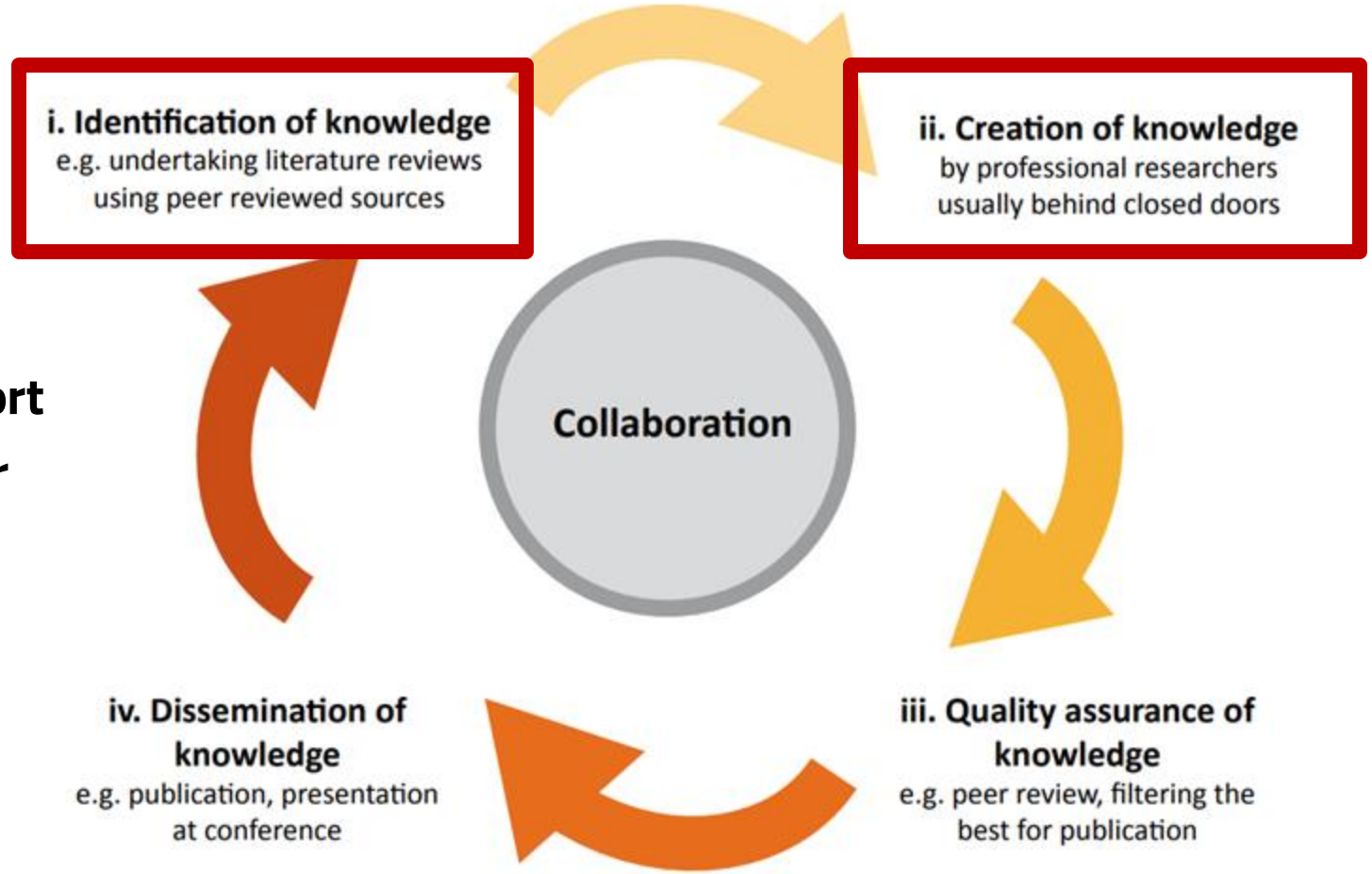
Figure 1: The academic research cycle



Academic Research Cycle

- **Proposal pitch/report**
- **Midterm office hour participation**

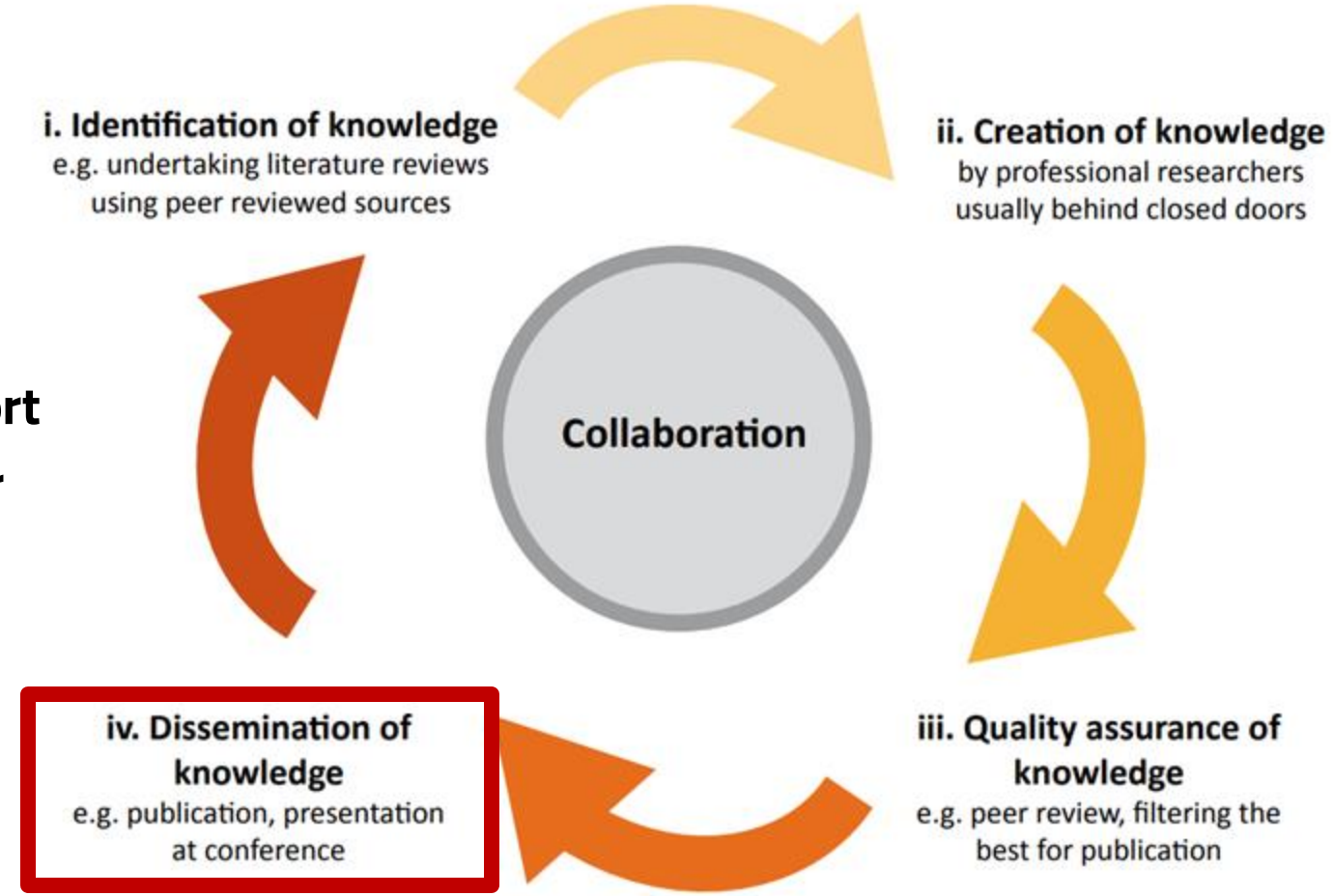
Figure 1: The academic research cycle



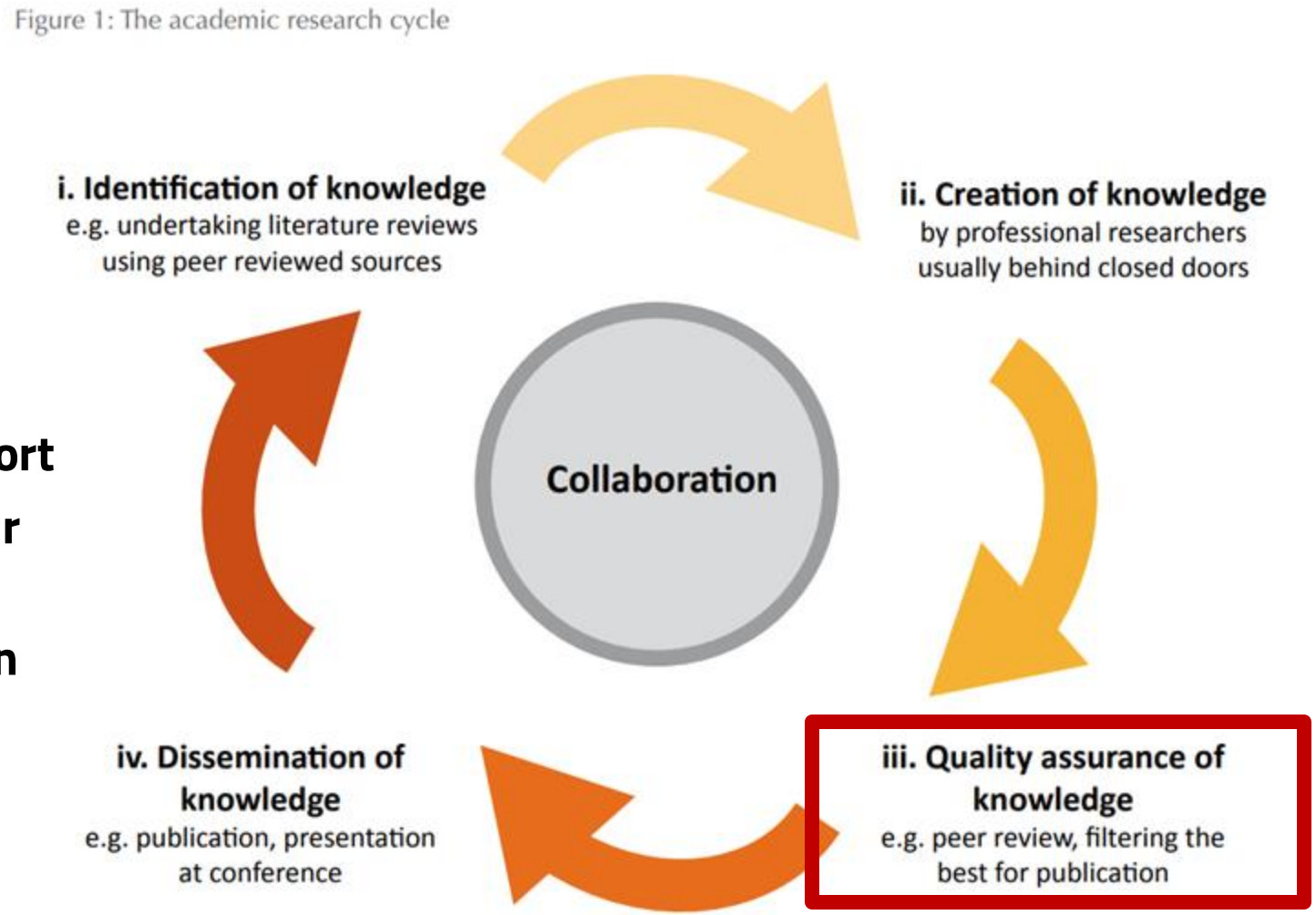
Academic Research Cycle

- **Proposal pitch/report**
- **Midterm office hour participation**
- **Poster presentation**

Figure 1: The academic research cycle



Academic Research Cycle



- **Proposal pitch/report**
- **Midterm office hour participation**
- **Poster presentation**
- **Final report**

Research Pipeline

- Motivation and problem formulation
- Data annotation or understanding of existing dataset
- Model development and replication of baseline models
- Experiment and error analysis (be critical and suspicious!)
- Discussion on limitations and ethical consideration
- Conclusion and future work

Project Evaluation

- HWs are generously graded but the **projects are not**! Therefore, students should consider the potential contribution of the projects rather than trying to play it safe. Playing it safe won't give them full marks.
- Three important rubrics:
  - **Novelty**: Compared to the state-of-the-art methods/systems/datasets, how novel is your approach? Is your work publishable?
  - **Significance**: How strong is your result? Is your finding still holding if different setups or prompting tricks?
  - **Clarity**: How clear and easy-to-follow is your report? Do you have well organized presentation of your results and problem definition?
  - https://dykang.github.io/classes/csci5541/F24/rubrick.html

Project Deliveries and Due

- ❑ Team formation and brainstorming (1 point each, due: Feb 6, Feb 18)
- ❑ Proposal pitch (3 points, due: Feb 25 and 27)
- ❑ Proposal report (5 points, due: Mar 2)
- ❑ Midterm office hour participation (5 points, due: Apr 1)
- ❑ Poster presentation (5 points, due: Apr 29 and May 1)
- ❑ Final report (10 points, due: May 8)

Project Information

Throughout the semester-long project, we aim to give you a taste of the full pipeline of NLP research, including problem formulation, literature surveys, data annotations, model replication, experiments, and analysis, as well as paper writing and presentation. Additionally, you will learn how to collaborate with your teammates and make regular progress on your research project. The mentors will be assigned to each team after you submit your team formation and brainstorming ideas, so you will have the opportunity to collaborate and discuss with other NLP researchers including DK and TAs. Please carefully read the following document that outlines instructions for your class projects, including the types of contributions, timeline and dues, types of project topics, and evaluation criteria.

Note: Please note that homeworks are generously graded but the projects are not. Therefore, students should consider the potential contribution of the projects rather than trying to play it safe. Playing it safe wont give them full marks.

1 Project Deliverables and Due Dates

Your project takes up 30% of your class grade. Every group member (maximum of 4 people) should submit their report, link to code (or a zipped code), and presentation slides/poster/webpages on Canvas before the deadline. Below is the list of your deliverables by due dates and link to Canvas submission:

- §1.1 Team formation (1 point, due: Sep 19) (canvas)
- §1.2 Project Brainstorming (1 point, due: Oct 1) (canvas)
- §1.3 In-class proposal pitch (3 points, Oct 8 and 10) (Slide deck for Group A and Group B)
- §1.4 Proposal report (5 points, due: Oct 13) (canvas)
- §1.5 Midterm office hour participation (5 points, due: Oct 31) (canvas)
- §1.6 Poster presentation (5 points, due: Dec 3) (canvas)
- §1.7 Final report (10 points, due: Dec 12) (canvas)

The late days and penalty will be applied to all team members for project deliverables. For each deliverable, please carefully read the specific instructions and the evaluation criteria below.

https://jimtmooney.github.io/Courses/S25/hw/Project_Description.pdf

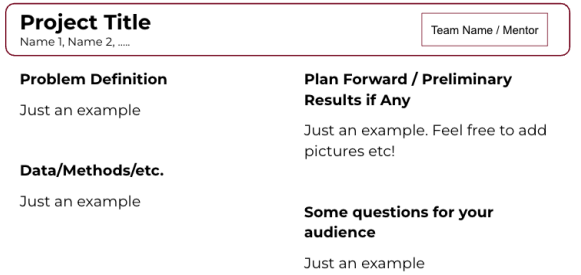
Team Formation and Brainstorming (2 points)

- ❑ Submit your team name and members to Canvas
- ❑ Submit a list of project ideas, titles, and plans (i.e., a few sentences) to Canvas.
- ❑ You will be assigned a project mentor with feedback within one week of submitting your ideas

```
Rubric (1 point) for team submission :
(0.5 point) Team name
(0.5 point) Member names
```

```
Rubric (1 point) Brainstorming ideas:
(0.5 point) Potential project titles and ideas
(0.5 point) Clear description of the ideas and execution plan
```

Proposal Pitch (3 points)



- Before submitting the proposal, your team needs to give a 3-minutes presentation of your proposal idea
  - Every member of your team should present in person or virtually for UNITE/remote students
- You need to follow the example template and create a slide for your own project in the slide deck, including the following:
  - What problem you are solving, what datasets/models you intend to use, what next steps to take, and questions about your project
- Your proposal should clearly address the comments and feedback

```
Rubric (3 points) for Proposal Pitch:
(1 point) Clear formulation and definition of your problem
(1 point) Specific execution plan (e.g., datasets, models, systems)
(1 point) Preliminary results if possible and questions for audience
```

Proposal Report (5 points)

- Maximum 3 pages report excluding references
- Upload your PDF report to Canvas using the class LaTex template
- Feedback on your proposal will be ready within two weeks after your submission

```
Rubric (5 points) for Proposal Report:
(0.5 point) Title, team name, members, and role assignment
(1 point) Clear Motivation and Problem definition
(1 point) In-depth Literature survey (at least three relevant and latest
↪  papers)
(2 point) ``Novel'' proposed idea and your execution plan (novelty: compare to
↪  the state of the art methods/systems/datasets, how novel is your
↪  approach?)
(0.5 point) Plan to address feedback from the pitch presentation
```

Midterm office hour participation (5 points)

- Schedule an office hour meeting with your assigned mentor (15 to 20 minutes) and discuss your intermediate results and progress with your project website
- The mentor expects you to give an update on your progress, ask questions, and consult with your plan until the final presentation.
- After the meeting, you should summarize what intermediate progress you made and what feedback and discussion you had with your mentor and submit it to Canvas.

```
Rubric (5 points) for Midterm Office Hour Participation:
(1 point) Additional development of your ideas after the proposal
(1 point) Submission of the project webpage
(2 points) Preliminary results and comparison to the baseline performance
↪  (e.g., experimental results, findings, visualization)
(1 point) Plan to address the mentor's feedback and plan until the end of the
↪  semester
```

Midterm office hour participation (5 points)

- Create a project webpage and show them during the discussion with your mentor.
  - Example template: https://github.com/minnesotanlp/csci5541-project-template
  - Example project website: https://csci5541-umn.github.io/
- You have to submit the updated website for the final submission

Midterm office hour participation (5 points)

Your Title Here

Fall 2024 CSCI 5541 NLP: Class Project - University of Minnesota

Team Name

Member 1 Member 2 Member 3 Member 4

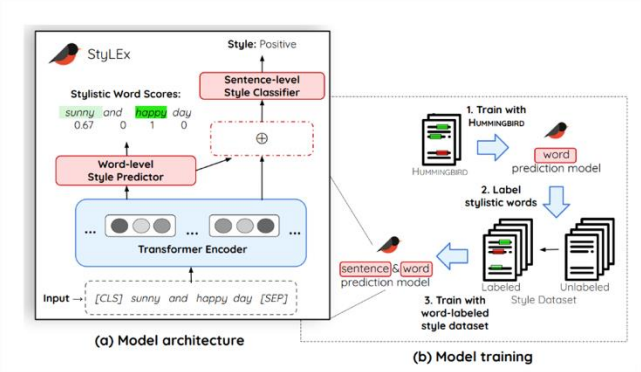
Final Report Code Model Weights

Abstract

One or two sentences on the motivation behind the problem you are solving. One or two sentences describing the approach you took. One or two sentences on the main result you obtained.

Teaser Figure

A figure that conveys the main idea behind the project or the main application being addressed. This figure is from StyLEx.



Any subsection

If you need to explain more about your figure

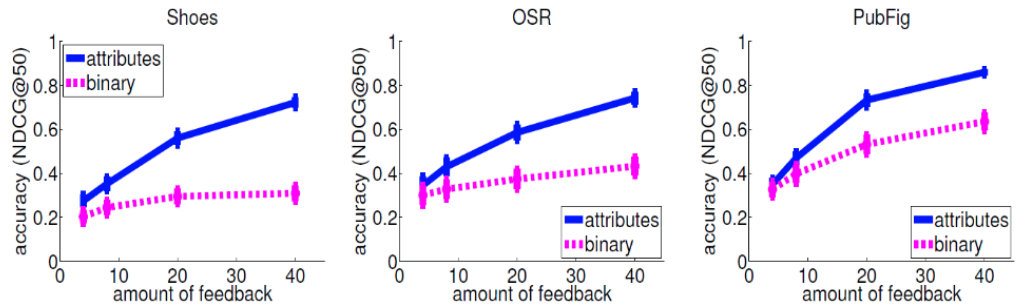
Results

How did you measure success? What experiments were used? What were the results, both quantitative and qualitative? Did you succeed? Did you fail? Why?

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Experiment	1	2	3
Sentence	Example 1	Example 2	Example 3
Errors	error A, error B, error C	error C	error B

Table 1. This is Table 1's caption



Conclusion and Future Work

How easily are your results able to be reproduced by others? Did your dataset or annotation affect other people's choice of research or development projects to undertake? Does your work have potential harm or risk to our society? What kinds? If so, how can you address them? What limitations does your model have? How can you extend your work for future research?

Poster presentation (5 points)

- Everyone on your team should present their work at your assigned poster session.
- Upload your poster PDF to Canvas before your presentation
- Evaluation:
  - Instructors will use the same rubric used in the final report except for the completeness of your work.
  - Every peer group is assigned a random team on their session day to review based on a rubric provided by instructors. Based on that, the team winning best poster will be given extra credit.

```
Motivation
Literature survey
Problem definition
Proposed ideas
Contribution
Experimental results and comparison with baselines
Main findings
Limitation and discussion
Plan for the final report.
```

Poster Sessions

- Print your “32x40” poster
- Location: Shepherd 164 (aka Drone Lab)
- Time: Apr 24 (Group A), Apr 26 (Group B)

- Printing instructions are provided at this link; you can request it using the form (details on how to fill out initial fields on next slide).
  - Keep in mind, they guarantee posters submitted 2 **business** days in advance, but **do not** work on the weekends.

CSE- Poster Printing Request Form

Request Details

Select your department *

Computer Science (CS)

Choose a printer *

Pick a printer that is large enough for your poster and prints on the material you want. One dimension of your poster must be less than or equal to the number indicated in the option.

Semigloss - 42"

Poster dimensions in inches *

Provide the size of the poster in inches. Examples: 72" x 42", 42" x 48", 20" x 36"

32"x40"

Advisor Approval *

The advisor approving this request. If you are the advisor, you can select your own information here.

Dongyeop Kang

Final report (10 points)

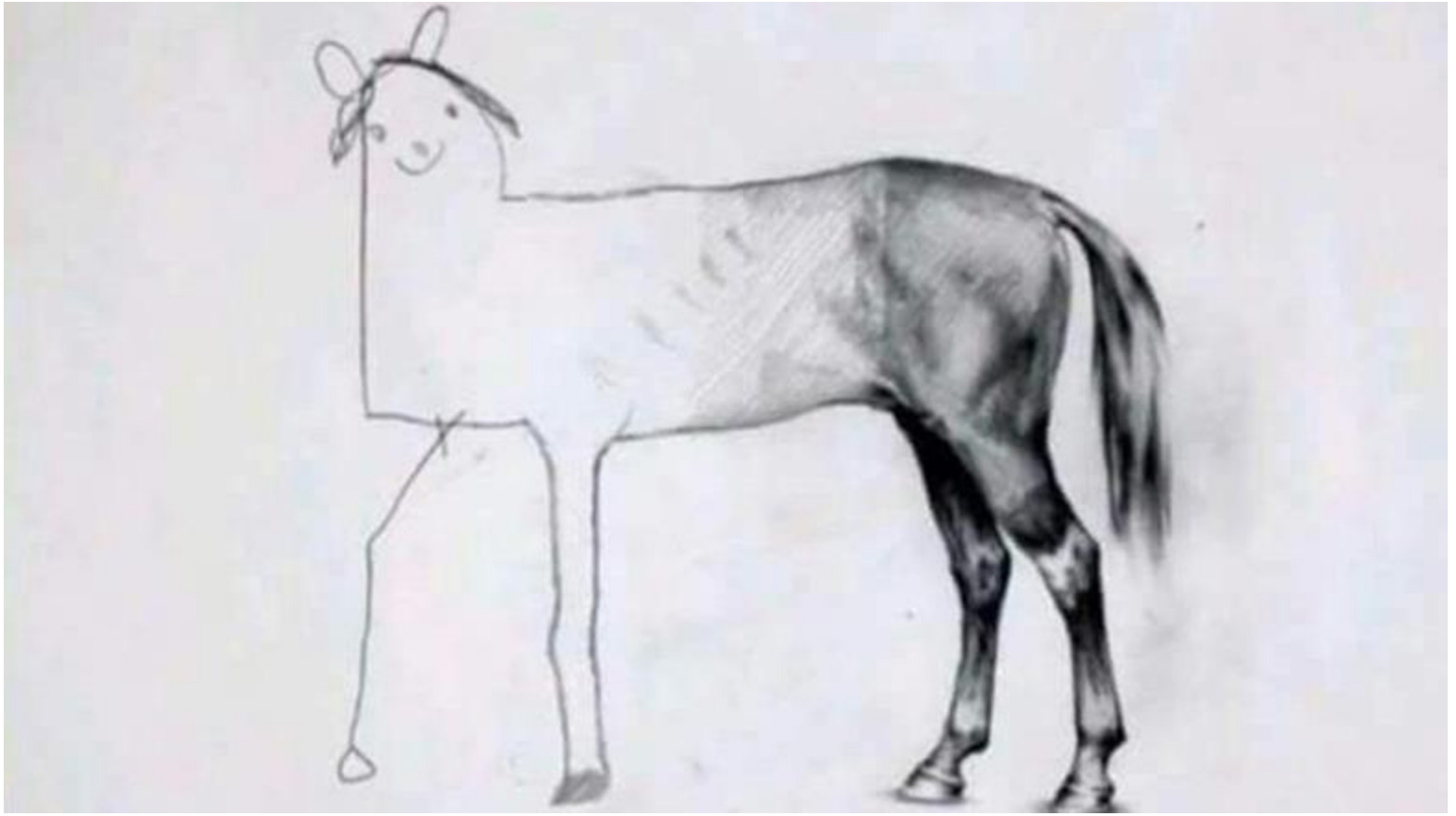
- Upload your PDF report, website, and code on Canvas
  - Maximum 8 pages with unlimited reference and appendix
  - Website with updated results
  - Zipped code or link to your github
- Rubric for the final evaluation
  - https://jimtmooney.github.io/Courses/S25/rubrick.html
  - 100 points will be normalized to 10 points in grading
  - This is a relative evaluation

```
Rubrik (100 points) for Final Report

Below are three general evaluation criteria:
(10 points) Novelty: Compared to the state-of-the-art
↪  methods/systems/datasets, how novel is your approach? Is your work
↪  publishable?
(10 points) Significance: How storng is your result? Is your finding still
↪  holding if different setups or prompting tricks?
(10 points) Clarity: How clear and easy-to-follow is your report? Do you have
↪  well organized presetnation of your results and problem definition?
Introduction / Background / Motivation:

Introduction / Background / Motivation:
(5 points) What problem do you try to solve? Describe your objectives cleraly
↪  without using any technical jargon.
(5 points) How is it done today by other researchers? What are the limitations
↪  and challenges of current practice?
(5 points) Who might be interested in your work? What kinds of impact can you
↪  make?

Approach:
(5 points) What did you do exactly? How did you solve the problem? Why did you
↪  think it would be successful? What is your hypothesis?
(5 points) What challenges did you anticipate and/or encounter during the
↪  development of your approach? Did the very first thing you tried work?
(5 points) What is scientific novel of your approach to address the
↪  challenges?
Experiments / Results / Error Analysis:

Experiments / Results / Error Analysis:
(5 points) How did you measure success? What research questions do you want to
↪  validate? What evaluation metrics and experiments were used? What were the
↪  results, both quantitative and qualitative? Did you succeed? Did you fail?
(5 points) No matter you succeed or fail, why? Which data points are
↪  incorrectly predicted by yours but previous models can't, or vice versa.
(5 points) Are there still some failure cases? Why can't your approach address
↪  them? Any potential solutions?
(5 points) Are the ideas/probelm/results presented with appropriate
↪  illustration?
Additional points:

Discussion points:
(5 points) Replicability: How easily are your results able to be reproduced by
↪  others?
(5 points) Datasets: Did your dataset or annotation affect other people's
↪  choice of research or development projects to undertake?
(5 points) Ethics: Does your work have potential harm or risk to our society?
↪  What kinds? If so, how can you address them?
(5 points) Discussion: What limitations does your model have? How can you
↪  extend your work for future research?
```

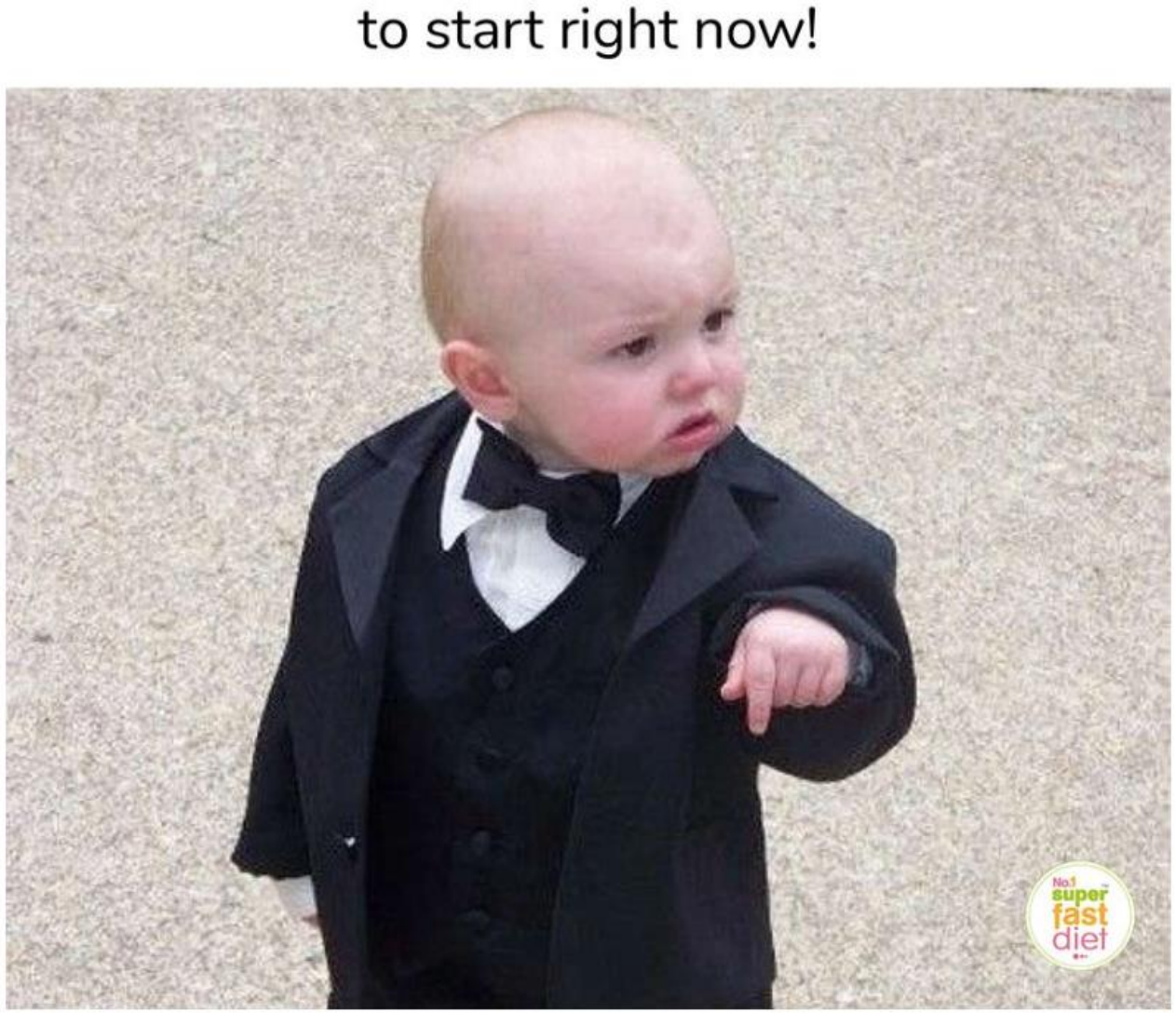
Some advices for successful projects

Don’t be ambitious

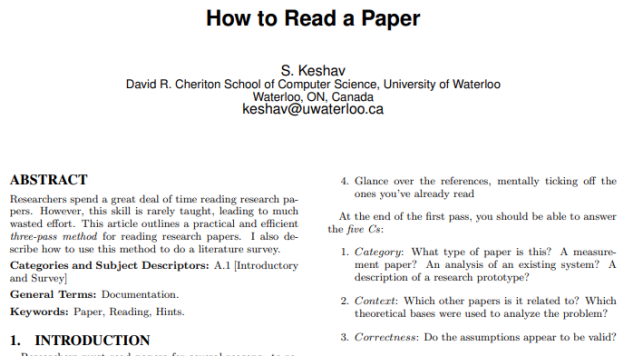
Don’t be ambitious



Start RIGHT NOW!



Literature survey



How to Read a Paper

S. Keshav
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, Canada
keshav@uwaterloo.ca

ABSTRACT

Researchers spend a great deal of time reading research papers. However, this skill is rarely taught, leading to much wasted effort. This article outlines a practical and efficient *three-pass method* for reading research papers. I also describe how to use this method to do a literature survey.

Categories and Subject Descriptors: A.1 [Introductory and Survey]

General Terms: Documentation.

Keywords: Paper, Reading, Hints.

1. INTRODUCTION

4. Glance over the references, mentally ticking off the ones you've already read

At the end of the first pass, you should be able to answer the *five Cs*:

1. *Category*: What type of paper is this? A measurement paper? An analysis of an existing system? A description of a research prototype?
2. *Context*: Which other papers is it related to? Which theoretical bases were used to analyze the problem?
3. *Correctness*: Do the assumptions appear to be valid?

- Do a thorough literature search
  - Google Scholar, ACL anthology (https://aclanthology.org/), arXiv (https://arxiv.org/archive/cs), OpenReview (https://openreview.net/), etc
  - If you find a similar/relevant paper, check out the other papers that recently cited it.
  - Examine gaps in more recent works – fill these gaps with your work
- Check out papers-with-code, github, project pages, etc
  - Re-use existing code on github or authors' sites.
  - Check out latest benchmark results in PapersWithCode leaderboard
- Tips for reading papers:
  - Do not read from the beginning to the end in order
  - Tables and figures with captions provide useful information at first glance
- Make a clear distinction of how your approach is different from prior work

Set Clear Project Novelty

- ❑ Novel dataset collection
- ❑ Interactive demonstration of an algorithm or system
- ❑ Research (significant findings and validation)
- ❑ SOTA beating
- ❑ ..?

Model Development

- ❑ Replicate and evaluate your **baseline** first
  - The following two baselines **MUST be included in your report**, if your paper's contribution is to propose a better model
    - ✓ Existing fine-tuned models or pre-trained models
    - ✓ ChatGPT, GPT4, and other LLMs
- ❑ Use Git(Hub) to version control your project
- ❑ Check out Huggingface's data and model cards
- ❑ Use Wandb and tensorboard for tracking your training
- ❑ Demonstrate your algorithm/model using Gradio or Streamlit

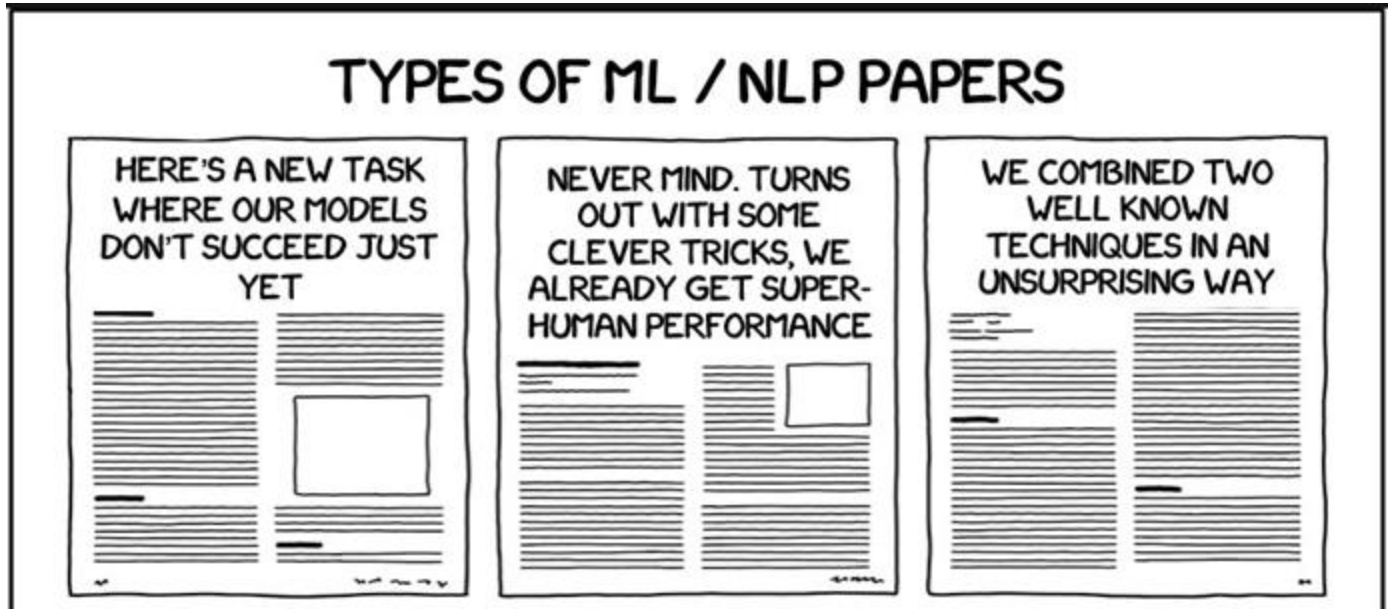
The DOs (for successful projects)

- Clearly divide work between members
- Start **EARLY** and work **REGULARY** every week rather than rush at the end
- Set up workflow – download data, verify data, set up base code on github, communicate via Slack, sharing results on Google spreadsheet, etc
- Have a clear, well-defined hypothesis to be tested
- Conclusions and results should provide some insights and takeaways
- Meaningful tables and plots to display the key results
  - ++ nice visualizations or interactive demos
  - ++ novel/impressive engineering feat
  - ++ good empirical results in both qualitative and quantitative ways.

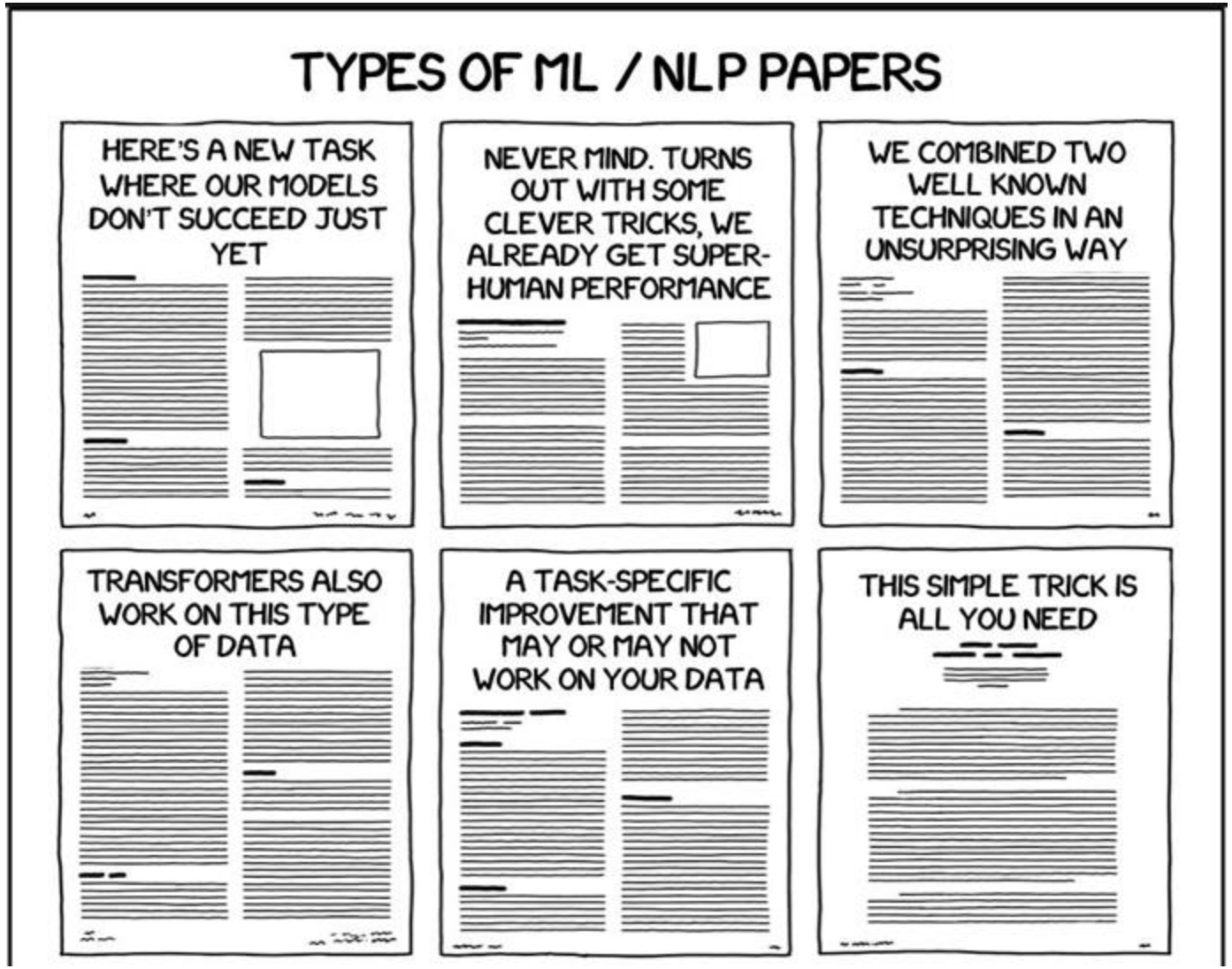
The Don'ts

- Data not available or hard to get access to, which stalls progress
- All experiments run with prepackaged source – no extra code written for model/data processing
- Team starts **LATE** – only draft of code up before dues
- Just ran model once or twice on the data and reported results (not much hyper-parameter tuning and statistical significance test)
- A few standard graphs (loss curves, accuracy) without any analysis
- Results/Conclusion don't say much besides that it didn't work
  - Negative results are fine, but only with in-depth analysis and justification

Project Types and Topics



https://twitter.com/seb_ruder/status/1387886948438708224



https://twitter.com/seb_ruder/status/1387886948438708224

Different types of contributions

The following are possible types of contributions you could make along with example papers:

- Critical analysis of existing model/dataset, e.g., [NRS+18, KL18, RKR20]
- New benchmark results and findings judged suitable for acceptance to an NLP or ML workshop,
- Collection of your own dataset on new tasks, (complex social) problems [EZM+21] or adversarial datasets [PWGK21] that can fool the existing systems,
- An in-depth literature survey on emerging topics [FGW+21, ZKK23],
- Interactive demonstration (e.g., Chrome Extension, Flask) [DKR+22, KMWK23] or visualization of existing systems [WTW+19],
- Applying NLP tools to your own domain of research (e.g., psychology [Kos23, Ull23], law [CHMS23], education, robotics [ABB+22]),
- New open-source repository or dataset with a high impact on the community
- Others (consult your mentors as soon as possible if you wish to do other types of projects).

https://jimtmooney.github.io/Courses/S25/hw/Project_Description.pdf

Trendy Topics in COLM CFP

https://colmweb.org/cfp.html

- All about **alignment**: fine-tuning, instruction-tuning, reinforcement learning (with human feedback), prompt tuning, and in-context alignment
- All about **data**: pre-training data, alignment data, and synthetic data --- via manual or algorithmic analysis, curation, and generation
- All about **evaluation**: benchmarks, simulation environments, scalable oversight, evaluation protocols and metrics, human and/or machine evaluation
- All about **societal implications**: bias, equity, misuse, jobs, climate change, and beyond
- All about **safety**: security, privacy, misinformation, adversarial attacks and defenses
- **Science of LMs**: scaling laws, fundamental limitations, emergent capabilities, demystification, interpretability, complexity, training dynamics, grokking, learning theory for LMs
- **Compute efficient LMs**: distillation, compression, quantization, sample efficient methods, memory efficient methods
- **Engineering for large LMs**: distributed training and inference on different hardware setups, training dynamics, optimization instability

Trendy Topics in COLM CFP (Cont'd)

https://colmweb.org/cfp.html

- **Learning algorithms** for LMs: learning, *un*learning, meta learning, model mixing methods, continual learning
- **Inference algorithms** for LMs: decoding algorithms, reasoning algorithms, search algorithms, planning algorithms
- **Reasoning:** Use of LLMs with Reinforcement Learning with Verifiable Rewards (RLVR) such as those from Deepseek-R1
- LMs for **everyone**: multi-linguality, low-resource languages, vernacular languages, multiculturalism, value pluralism
- LMs and **the world**: factuality, retrieval-augmented LMs, knowledge models, commonsense reasoning, theory of mind, social norms, pragmatics, and world models
- LMs and **embodiment**: perception, action, robotics, and multimodality
- LMs and **interactions**: conversation, interactive learning, and multi-agents learning
- LMs with **tools and code**: integration with tools and APIs, LM-driven software engineering
- LMs on **diverse modalities and novel applications**: visual LMs, code LMs, math LMs, and so forth, with extra encouragements for less studied modalities or applications such as chemistry, medicine, education, database and beyond

What to do now?

- Brainstorming
  - Each member produces ideas
  - Refine and filter out ideas
    - ✓ Data availability
    - ✓ Has the same idea been done before (with possibly existing github code)? Do lit survey
    - ✓ ..
  - Replicate a baseline model using HuggingFace model
  - Consult with your mentors
  - Examine Previous examples on course webpage