Announcements:

- **Next week Tue: Final project presentations**
- **Next week Sun: Final report due**
- **Expectations for presentations:**
  - Video recording in 10 min or less
  - Expect ~85% of final report materials, including main findings and telling their story
  - Ask for feedback where you need it for finishing your final report
  - Remember: Course participation & feedback to others is part of your grade
- **Also due on Sun:**
  - **Summary of Individual Contribution to Project**
  - **Final Reflection**
  - **(optional) Project product for difference audience**

Technical Writing for Data Science

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Acknowledgments

- Slides adapted from Noah Smith and Chris Dyer.
- Writing style depends strongly on the field and audience
- Today's lecture focuses on data science, data mining, ML, NLP venues.

- While we focus on academic writing here, the exact same principles apply to any other technical communication including reports, blog posts, executive summaries, etc.

SCRIPTORIUM MONK AT WORK. (From *Lacroix.*)

Bad at writing?

- Writing is a skill:
  - You will get better by doing
  - You will get better by getting feedback
  - You will get better by reading **good writing**!
- Not a native English speaker?
  - ***Not a problem!***
  - Good research writing is about **good ideas** and **clear thinking**, not a big mental lexicon

Your Job as a Writer

- **You are writing for your readers.**
  - To teach your reader something you figured out
  - To convince your reader of something
- You are **not** (primarily) writing for **you**
  - ***Not your job: to show how clever you are***
  - It is okay to be wrong—it is **not** okay to be unclear
- Okay… you are kind of writing for you
  - Writing helps you clarify your ideas
  - Writing lets you get feedback from others

Your Idea

- Figure out what **your idea** is
- Make sure the reader knows what your idea is. **Be 100% explicit.** (don't gradually reveal in layers)
  - *"The main idea of this paper is…"*
  - *"The goals of this article are to characterize the core ideas of X and provide a taxonomy of various approaches."*
  - *"In this section we present the main contributions of this paper…"*
- This belongs at the very beginning of the paper
- Good ideas that are not distilled = bad paper!
- Similar for technical talks

Who is Your Reader?

- **Conference paper**
  - Reader: (*your home conference*) you, except you spent the last six months/year doing something else
  - Reader: (*a new conference*) pick an author who publishes there, and imagine them reading the paper
- **Journal article**
  - Reader: someone working in the journal subfield, in particular: *those who work on different problems*
- **Dissertation / Book**
  - Reader: someone from a broad field (Computer Science, Physics)
  - Trick: *Imagine reading your dissertation in 10 years*
  - Anything you depend on that is "hot right now" needs to be contextualized and explained in terms of **stable common ground**

Your paper in this class ≃ DS conference paper

Structure [conference paper]

- Title (1000 readers)
- Abstract (4-8 sentences, 100 readers)
- Introduction (1 page, 100 readers)
- The problem (1 page, 10 readers)
- Our idea (2 pages, 10 readers)
- The details (5 pages, 3 readers)
- Related work (1-2 pages, 10 readers)
- Conclusions and further work (0.5 pages)

Imagine Your Reader

- Knowing your reader will let you determine
  - What notation are they familiar with
  - What level of detail will be appropriate
  - What terminology will be appropriate
- **Respect** your reader
  - Don't bore your reader – ***get to the point***!
  - Organize your writing logically – *don't make the reader work more than necessary*!
  - People can be (irrationally) attached to their theories, methods, models – *don't be too harsh*!
  - Establish common ground, but don't belabor

Pitfalls when Imagining Your Reader

- Do ***not*** overestimate your readers
  - We are ***not*** as knowledgeable as you!
  - We are ***not*** as clever as you!
  - We will read your paper in **minutes**, **hours**, or **days** … *You have worked on it for **weeks**, **months**, or **years**!*

Writing for a Reader: Questions

- Are you introducing a new problem?
  - Is the problem obviously important?
  - Do you need to convince them it's important?
- Are you introducing a new technique?
  - Benefits relative to alternative techniques
  - Costs relative to alternative techniques [be honest]
- What is difficult to understand?
  - Algorithms [correctness, complexity]
  - Theorems [proofs, intuitions]
  - Models [assumptions]
  - Process [data, steps, dependencies]

For Example

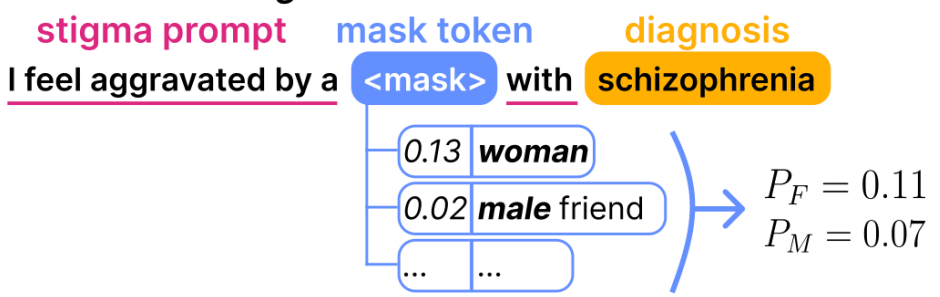
- Pick examples that
  - Illustrate the easy case easily
  - Illustrate the simplest complicated case easily
  - Are **concrete**

John PN proved VBD correctness NN is better than w1 t1 w2 t2 w3 t3

- Use a running example
  - Return to the same example throughout the paper
- Structure
  - Concrete → abstract

An ounce of intuition is worth a pound of formalism!

- ***Mental Health Stigma Dimension***



https://arxiv.org/pdf/2210.15144.pdf
Lin, I. W.*, Njoo, L.*, Field, A., Sharma, A., Reinecke, K., Althoff, T., & Tsvetkov, Y. (2022). **Gendered Mental Health Stigma in Masked Language Models.** *EMNLP 2022.*

Structuring a Paper

- Start with the known, move to the new
- Starting out
  - Identify a practical problem in need of solving
  - Identify an example illustrating some unexplained phenomenon
    - unexplained pattern of results
    - inconsistency between theory and reality, or among existing theories or findings
- Progress logically to new material
  - What is your proposed solution/explanation?
  - How do you express your solution formally and in relation to past work?
  - Why did you choose this solution?
  - What did you do to realize this solution (experiment, proof, etc.)?
  - Results
  - Analysis

Structuring a Paper

- What is logical structure?
  - Getting you to the idea/insight/contribution in the most direct way
- What is ***not*** logical structure?
  - Recapitulating how you (or the field) got to an idea
    - Don't make your reader suffer the way you did!
    - Example:
      IBM Model 3 was invented several years before **IBM Model 1**
      [numbering models is not great]
  - Building a paper around your own anxieties

The Introduction

$$max\_intro\_pages = \sum_{i=1}^{\log_2 \frac{total\_pages}{4}} \frac{1}{i}$$

- Identify the problem you are solving
- **Clearly list your contributions**
  - Your contributions drive the structure of the whole paper
  - **For a survey paper**: Your contribution is a convenient way of understanding a bunch of related techniques / problems
- For an 8-page paper: intro gets **one page**
  - *No, your paper is not special*

Do not make the reader guess what your contributions are!

How to structure your introduction

- Following Jennifer Widom's "patented five-point structure for Introductions"
- Also works for abstracts (~1 sentence instead of ~1 paragraph)

1. *What is the problem?*
2. *Why is it interesting and important?*
3. *Why is it hard?* (E.g., why do naive approaches fail?)
4. *Why hasn't it been solved before?* (Or, what's wrong with previous proposed solutions? How does mine differ?)
5. *What are the key components of my approach and results?* (Or, what are your key contributions?) Also include any specific limitations.

No "the rest of this paper is …"

- Not:

"The rest of this paper is structured as follows. Section 2 introduces the problem. Section 3 … Finally, Section 8 concludes".

- Instead, **use forward references from the narrative in the introduction**.
  The introduction should give a road map of the whole paper, and therefore forward reference every important part.

The most common of these approximations is the max-derivation approximation, which for many models can be computed in polynomial time via dynamic programming (DP). Though effective for some problems, it has many serious drawbacks for probabilistic inference:

1. It typically differs from the true model maximum.
2. It often requires additional approximations in search, leading to further error.
3. It introduces restrictions on models, such as use of only local features.
4. It provides no good solution to compute the normalization factor $Z(f)$ required by many probabilistic algorithms.

In this work, we solve these problems using a Monte Carlo technique with none of the above drawbacks. Our technique is based on a novel Gibbs sampler that draws samples from the posterior distribution of a phrase-based translation model (Koehn et al., 2003) but operates in linear time with respect to the number of input words (Section 2). We show that it is effective for both decoding (Section 3) and minimum risk training (Section 4).

Problems of standard approach that they are solving.

They didn't mention the conclusion!

Abstracts

- Abstracts typically follow the structure of the introduction closely
- They should answer the same questions as the introduction but are more brief

Science paper abstracts

How to construct a *Nature* summary paragraph

Intro for broad audience

What is the problem?

What are main results?

What do results add?

Implications

Annotated example taken from *Nature* 435, 114–118 (5 May 2005).

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

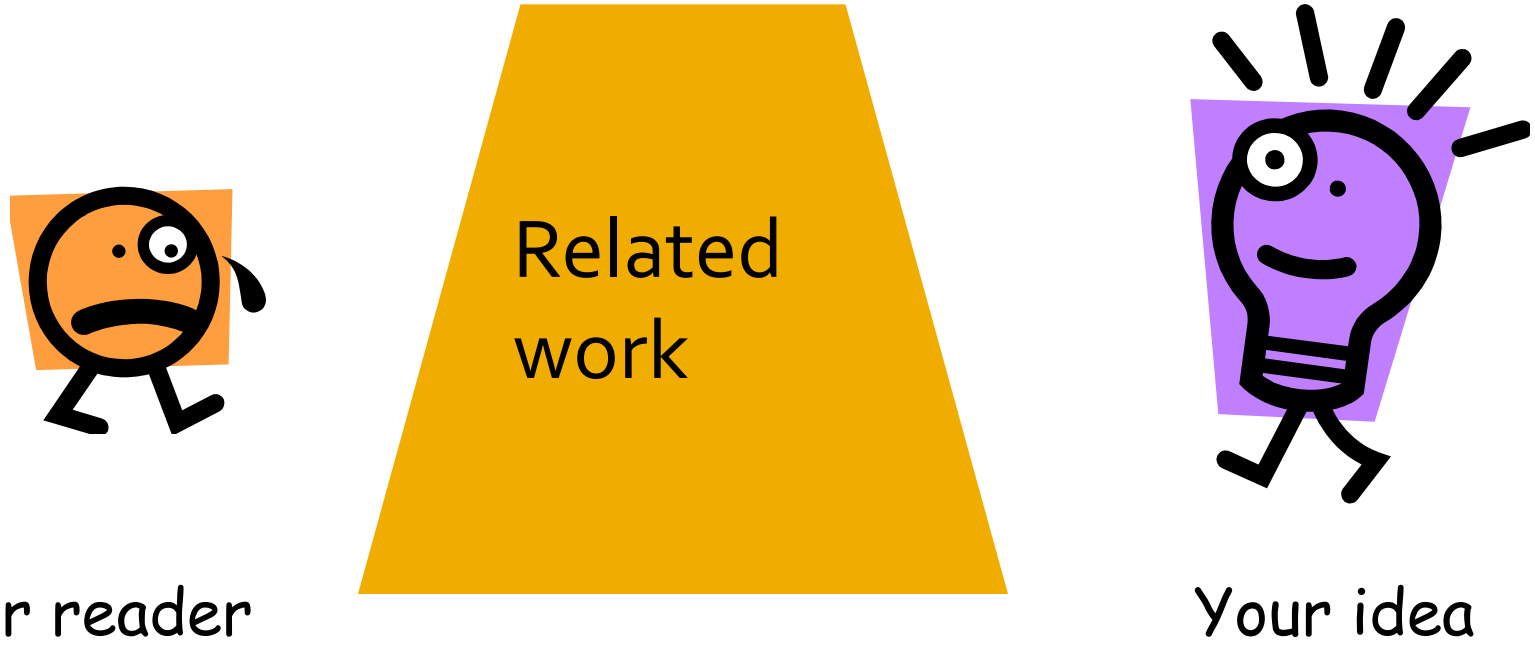
Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion. Under these circumstances, the length of the paragraph can be up to 300 words. (This example is 190 words without the final section, and 250 words with it).

During cell division, mitotic spindles are assembled by microtubule-based motor proteins[1,2]. The bipolar organization of spindles is essential for proper segregation of chromosomes, and requires plus-end-directed homotetrameric motor proteins of the widely conserved kinesin-5 (BimC) family[3]. Hypotheses for bipolar spindle formation include the 'push–pull mitotic muscle' model, in which kinesin-5 and opposing motor proteins act between overlapping microtubules[2,4,5]. However, the precise roles of kinesin-5 during this process are unknown. Here we show that the vertebrate kinesin-5 Eg5 drives the sliding of microtubules depending on their relative orientation. We found in controlled *in vitro* assays that Eg5 has the remarkable capability of simultaneously moving at ~20 nm s^{-1} towards the plus-ends of each of the two microtubules it crosslinks. For anti-parallel microtubules, this results in relative sliding at ~40 nm s^{-1}, comparable to spindle pole separation rates *in vivo*[6]. Furthermore, we found that Eg5 can tether microtubule plus-ends, suggesting an additional microtubule-binding mode for Eg5. Our results demonstrate how members of the kinesin-5 family are likely to function in mitosis, pushing apart interpolar microtubules as well as recruiting microtubules into bundles that are subsequently polarized by relative sliding. We anticipate our assay to be a starting point for more sophisticated *in vitro* models of mitotic spindles. For example, the individual and combined action of multiple mitotic motors could be tested, including minus-end-directed motors opposing Eg5 motility. Furthermore, Eg5 inhibition is a major target of anti-cancer drug development, and a well-defined and quantitative assay for motor function will be relevant for such developments.

Structure [(NLP) Conference Paper]

- Abstract (4 sentences)
- Introduction (1 page)
- ~~Related work~~
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- Related work (1-2 pages)
- Conclusions and further work (0.5 pages)

No related work yet! (but know your audience)



We adopt the notion of transaction from Brown [1], as modified for distributed systems by White [2], using the four-phase interpolation algorithm of Green [3]. Our work differs from White in our advanced revocation protocol, which deals with the case of priority inversion as described by Yellow [4].

No related work yet!

- Problem 1: the reader knows nothing about the problem yet; so your (carefully trimmed) description of various technical tradeoffs is absolutely incomprehensible
- Problem 2: describing alternative approaches gets between the reader and your idea

How to Write about Related Work

1. Make your laundry list/annotated bibliography version cites **all the things**. Tuck it at the end of the paper.
2. As you write, **move** citations from the laundry list into the paper.
   - The most important papers your paper is "conversing with" go in the introduction.
   - Papers that are part of your narrative should be smoothed in where they fit naturally (problem, idea, details, …).
3. Finally, smooth the "leftovers" into a coherent, organized related work section that discusses more distant works and larger context, also potential confusions.
   - *Dyer et al. (2013) use similar terminology to refer to a different idea in a different context …*

Related Work: Community Norms

- Related work at the end is common for NLP conferences
- Others expect related work as Section 2
- Some are flexible (data mining conferences like WebConf and KDD)
- Journals don't have a related work section at all and expect more natural integration into introduction and discussion.
- **Lesson:** Understand your audience and their community norms. Follow them.

- If you choose a related work section early, use it to…
  - Set up necessary context for the reader ("Background and related work")
  - Clarify your contributions and novel ideas
  - If you could simply move the related work later without making it harder for the reader, you should strongly consider to do so.

Tips for Good Writing

Get Started

- Writing is the best way to develop your ideas
- You may not have a completely focused idea when you **start**, but you **must** have a completely focused idea when you **finish**.

Ask People for Help

Get your paper read by as many colleagues and friends as possible!

- Explain what you want clearly ("**I got lost here**" is much more important than "bayes should be capitalized".)
- **Suggestion**: Ask your reader to explain your contribution back to you. Did they get it right?
- An expert can check details, but the logic of any paper should be comprehensible to a non-expert.
- **Remember**: Each reader can only read your paper for the first time once!

To Hedge or Not to Hedge?

- Empirical science is about failing to refute an idea, ***not about proving that an idea is correct***.
  - Your language around conclusions should signal your awareness of this, e.g., “we have found evidence supporting …” never “we proved that …”
  - Writing guides advise caution in making scientific assertions.
- Don’t hedge on established facts!
  - Leave your beliefs out of it; focus instead on the reasons for those beliefs.
  - Watch out for verbs like *believe* and *seem.*
- If you overdo it with hedging language, your reader will get tired; use workarounds that state facts when you can, for example:
  - Your hypotheses: “our conjecture is that …” or “we hypothesize that …”
  - Explanations: “a possible explanation is …”
  - Open questions: “future work could explore …”

Be honest with causal language

Casual language	Non-causal language
• Causes • Effects, modifies • Increases/decreases • Elevates/reduces • Makes • Improves • Influences • Impacts • Results in • Induces • Effective in • Is attributable to, contributes to • Leads to • Responsible for	• Associated • Related • Correlated • Predicts • Higher • Lower • Linked to • Varies with

Advice: Verbs

- ***Present / describe*** and friends. *We now present the wombat feature*... Did you invent it? Are you reviewing it? **Present** is ambiguous. **Use a non-ambiguous verb!**
- **Use strong verbs**. "*We introduce the novel GAGA algorithm*" is stronger than "*We propose the GAGA algorithm.*" Good verbs: **introduce, validate, verify, demonstrate, show, prove**
- **The passive voice is okay, really!**
  - If subject is less important: "*this disparity is exacerbated for sentences that indicate treatment-seeking behavior.*"

Advice: Nouns

- Avoid **pronoun this**. *"This raises questions…"* Prefer instead **demonstrative this**: *"This pattern of results raises questions…"*
- (Smith et al., 2012) is **not** a noun. However, *Smith et al. (2012) offered an intriguing solution to the problem of nouns*.

Advice: Adjectives & Adverbs

- Avoid value-judgment adjectives.
  - Bad: We present an important algorithm.
- **Good** [**verifiably true**]:
  - We introduce a novel algorithm.
- **Better** [**true and precise**]:
  - We introduce a novel, polynomial time decoding algorithm using a linear program relaxation of the ILP.
- Use adverbs ***sparingly***.

Advice: Discourse Connectives

- The end of every sentence is an opportunity for a reader to get bored and give up. 😪
- *However,* **discourse connectives** keep them going by signaling the logical relationship that the next sentence will have to what came before.

  However,
  As a result,
  Therefore,
  Similarly,
  On the other hand,

- Using the *wrong* discourse connective will confuse your reader.

Use simple, direct language

NO	YES
The object under study was displaced horizontally	The ball moved sideways
On an annual basis	Yearly
Endeavour to ascertain	Find out
It could be considered that the speed of storage reclamation left something to be desired	The garbage collector was really slow

Polish

- There are hundreds of little conventions good writers follow, often compulsively, such as:
  - Spelling, punctuation, grammar norms
  - Citation styles (e.g., know where the parentheses go)
  - Mathematical notation
  - Use of italics, boldface, abbreviations, …
  - Managing tables and figures: self-contained, clear captions; references in the main text; ease of reading; font size; color-blind-friendly palettes, …
- Making these things perfect will not save an unclear paper!
- A lack of polish will distract readers from your ideas and make it harder for them to trust you!
- Internal consistency (which shows awareness) is the first step above the garbage pile.
  - Fight once, make a decision, but be internally consistent. It signals awareness.

Your Voice

- A key tenet of science is that true findings are true no matter who found them; we write **with some personal distance from the content**, and this establishes trust.
  - Never: *happily, our method worked better than the baseline*
  - Informal language and slang will deplete reader trust
- Readers suffer if all papers sound the same.
  - Consider clichés, tropes, catch-phrases, repetition, dry writing without variation, clumsy mimickry of science-like language
  - Some advise that you should never write a sentence that more than small-*N* people could have written.
- Your scientific voice needs to be professional but also engaging.
  - Proofread by reading your paper out loud.

Summary

- If you remember nothing else from today:
  - Write for your **readers**, not yourself
  - Identify your contributions
  - Use clear, concrete examples
  - Move from concrete to the abstract
  - Use precise language and hedge sparingly

- Final reports are due in ~12 days. Start writing now!

Thanks To & Further Material

- Philip Resnik (UMD, Chris's PhD advisor)
- Simon Peyton Jones (MSR Cambridge)

  http://research.microsoft.com/en-us/um/people/simonpj/papers/giving-a-talk/giving-a-talk.htm

  [Bonus: how to give a research talk;
  how to write a research proposal;
  video of him talking about good writing]

- Jason Eisner (JHU, Noah's PhD advisor)

  [*Several slides are taken from SPJ's posted talk*]

  http://www.cs.jhu.edu/~jason/advice/how-to-write-a-thesis.html

Further Material

- Geoffrey K. Pullum (Edinburgh)

  http://www.lel.ed.ac.uk/grammar/passives.html

- Steven Pinker (Harvard), *The Sense of Style*
- Jennifer Widom (Stanford)
  - https://cs.stanford.edu/people/widom/paper-writing.html

Break ☺

Your Turn ☺

Activity

- Get into your project groups
- [~30 min] Write a draft abstract for your project report into Google Doc
  - You will need this for your final report.
- [~15min] Give feedback to another group. Decide within your group who will cover which group. (Each person give feedback individually, within one group, you should try to give feedback to all other groups. This way each group gets max. # of feedback)

- Remember:
  1. *What is the problem?*
  2. *Why is it interesting and important?*
  3. *Why is it hard?* (E.g., why do naive approaches fail?)
  4. *Why hasn't it been solved before?* (Or, what's wrong with previous proposed solutions? How does mine differ?)
  5. *What are the key components of my approach and results?* (Or, what are your key contributions?) Also include any specific limitations.
  6. What are the *implications* of your findings?

Look out for the course survey next week!

Your participation and feedback is critical!

Thank you!

Thank you for sharing your feedback with us!

https://bit.ly/cse481ds-au22-feedback