The CS166 Research Project

The research project for CS166 will serve as a capstone for the course. It’s a way for you to run wild with a topic, discover something interesting, and share it with everyone.

This handout goes over our expectations for the research project. Some of this is logistics (when are things due?), while some are more big-picture (what’s expected of you?). If there’s anything you’re unclear or unsure about, please let us know – we’re happy to clarify!

Here’s a quick summary of the deadlines for the project components:

- **Proposal due Thursday, April 30th at 2:30PM Pacific.**
- **Checkpoint 1 due Tuesday, May 19th at 2:30PM Pacific.**
- **Checkpoint 2 due Thursday, May 28th at 2:30PM Pacific.**
- **Project due Tuesday, June 9th at 2:30PM Pacific.**
Project Overview
In a team of four, you'll choose a data structure, an algorithm pertaining to a specific data structure, or a theoretical result about data structures that will serve as the focus of your project. You'll spend the remainder of the quarter doing your best to become an expert on the topic. To do that, you'll probably scour a bunch of research papers, blog posts, book chapters, etc. to see what you find.

Once you're up to speed with how the data structure works and how it sits in the larger ecosystem, you'll do two things:

- **Create an explanatory article on your topic.** You and your team will put together a (possibly interactive) article that helps people new to the topic build a deep understanding of how your topic works. You'll also describe your “interesting” component (detailed below) in this explanation.

- **Do something “interesting” with the data structure.** This part of the project, which we refer to as the “interesting” component of the project, is fairly open-ended. You should aim to do something with the data structure that sheds some new light on it in some way. We'll discuss the “interesting” component in more detail later.

There are three steps leading up to the final project deadline.

- A **project proposal** in which you'll tell us who you're working with and list the topics you are most interested in exploring. Once this is submitted, we'll assign project topics and send each team some questions to think about.

- A **first checkpoint** where you'll submit a progress report, answer questions from the course staff, and outline areas that you’d like to learn more about. We’ll then offer feedback and guidance about how to proceed going forward.

- A **second checkpoint** where you'll turn in a more polished draft of your explanatory article and outline the preliminary results from your “interesting” component. We'll then offer more feedback and guidance as you finish things up.

Your project will be evaluated based on the checkpoints, your explanatory article, and your interesting component, all on an S/NC basis.

The rest of this handout details the individual components of the project, as well as expectations for the checkpoints.
The Explanatory Article

Your major deliverable is an explanatory article on your topic. The goal of this article is to guide a reader on a journey from ignorance to understanding. This means that you’ll want to explain your concepts using a mix of high-level intuitions, drawings, animations, interactive demos (if you’d like), and rigorous math.

Think of this as more of an exercise in good storytelling rather than an exercise in writing a journal article. Journal articles are designed to be compact. There’s an expectation that someone reading a paper will sit down with scratch paper and take notes, following along with the reasoning and pausing to verify intermediate results if necessary. In that sense, a journal article can be thought of as “knowledge concentrate” – you need to add some water and stir quite a bit to get something edible out of it.

On the other hand, what we’re asking you to do is to write something that is lucid, clear, well-motivated, and easy to follow. Be conversational. Draw lots of pictures. Provide motivation for the topic and each piece. Give intuitive explanations before going into more difficult math. Create some animations showing how things work, or give the user interactive controls they can use to step through things.

Here are some examples of the sorts of explanatory articles that, in our view, do a great job explaining how things work:

- **Pipes, Forks, & Dups: Understanding Command Execution and Input/Output Data Flow** by Roz Cyrus. This article explains highly technical concepts (I/O and redirection) using a blend of text, code, and pictures. I’m particularly impressed with how the visuals give a sense of “plug and play” that is inherent to how the code works but not directly manifested anywhere in the C source.

- **It’s Triangles All the Way Down** by Diego Hernandez, Griffin Koontz, and Amy Liu. This blog shows an unexpected connection between binary search trees and triangulations of \((n+2)\)-vertex polygons, how they connect with tree rotations, and how they simplify our understanding of deleting from BSTs. This was a past CS166 project and the TAs and I were impressed by how easy this was to follow and how the pictures and animations make everything so easy to follow, and the included interactive demo makes it easy to play with the concepts.

- **Making Maps with Noise Functions** by Amit Patel. This article explores noise functions and how to use them to generate realistic terrains. The interactive slides make this a joy to explore and provide a powerful intuition for how everything fits together.

- **Understanding the Inverse Ackermann Function** by Raimund Seidel. This talk, which was given by a computer scientist who published a new analysis of the union-find data structure, does a great job motivating both the result that he proved and explaining what on earth the Ackermann inverse function actually is. It blends technical rigor with a conversational tone and good use of diagrams.

- **Ukkonen’s Suffix Tree Algorithm in Plain English** by Johannes Goller. This explanation of Ukkonen’s suffix tree construction algorithm proceeds incrementally, showing the algorithm working in different stages and switching back and forth between different representations of the same data.

- **The Evolution of Trust** by Nicky Case. This interactive exploration of the iterated prisoner’s dilemma gives an intuitive feel for the game and the different strategies involved, and in doing so surveys the literature in a deep but approachable way. Is this an “article” in the conventional sense? Nope! But would we be okay with you submitting something like this? Oh yes, definitely!

- **The Essence of Calculus, Chapter 1** by Grant Sanderson. This video on integrals masterfully uses visualizations to motivate basic concepts of calculus in a way that both foreshadows how more advanced topics work and makes the whole idea of an integral intuitive and worthwhile. (And yes, if you’d like to submit a video explanation, go right ahead! But we’re still going to expect it to be rigorous. 😊)
To get to the point where you’re able to produce something lucid and clear like the works cited above, you will need to get very comfortable with your topic. If you’re reading primary sources (e.g. the original papers on your topic), you’ll likely need to do some reverse-engineering and unpacking of the topics you’re exploring. If you’re reading secondary sources (blog posts, slide decks, or course materials), some of this may be done for you, but you’ll still need to find your own way to tell the story. Here is our advice about how to interrogate a topic and better understand how it works:

• **Draw pictures!** This may seem like a simple piece of advice, but so many complicated concepts from CS theory become so much easier once you know the right picture to draw. This is especially helpful when looking at complicated data structures, mathematical analyses, etc. And don’t just draw one picture. Can you draw the same concept in multiple ways? Some data structures make so much more sense when you see them in action.

• **Switch from operational to mechanical definitions.** You’ve seen the benefits of having both mechanical and operational definitions for different data structures. Often, when you’re first reading up on a data structure, you’ll only see one of those two definitions. Finding the other can be incredibly helpful in seeing where things come from. For example, I personally found the rules for B-trees to be totally inscrutable until I realized they’re what happens if you obey the two simple rules of “add things to leaf nodes” and “split big nodes and kick keys higher up.”

• **Perturb things.** Some design decisions in data structures are totally arbitrary. Others are completely necessary for things to work out the way they do. Think the choice of $b = \frac{1}{4} \log_2 n$ in Fischer-Heun: there’s a balance between two competing forces ($b$ must be small so that there aren’t too many block types, but large so that the sparse table can be built quickly). If you don’t see why something is there, change it and think about what would happen. You can learn a lot about a data structure by seeing how things break!

• **Break things into layers.** Some data structures are difficult to understand because there are many moving parts linked together. In many cases, what you’re looking at is really several independent ideas that are all stacked on top of one another. For example, Fibonacci heaps are way easier to understand if you recognize that there’s a basic idea (use trees with lazy deletion) combined with something messier (oh yikes, we need a lot of pointers and linked lists to make this work quickly).

• **Look for isometries.** Imagine you had stumbled upon the pseudocode for heapsort. The code for heapsort is purely done through a series of array operations based on multiplying and dividing by two. If you just see that code, it’ll be difficult to figure out what on earth the algorithm is doing and why it works. But if you realize that there array is a representation of another data structure (binary heaps), then the algorithm’s correctness becomes “obvious.”

• **Read voraciously.** If you’re hitting a dead end, see what other authors have had to say about your topic. Read the original paper, see what it’s citing, and see what cites it. Look for lecture slides or book chapters on the topic. Chances are that seeing multiple perspectives might give you a better foothold to master the topic yourself.

• **Code things up.** Sometimes, things will only click when you code up a data structure and go through the process of debugging it. You might realize that something isn’t working as intended not because you have a bug in your code, but because you didn’t fully understand how the data structure works. (A note: there are some data structures that are so complex that coding them up would be a nightmarish experience. Don’t feel that you have to do this!)

• **Narrow your scope.** Some topics are just denser and more complicated than others, and if you’re overwhelmed by the scope of what’s in front of you, feel free to narrow your focus! Some of the best projects we’ve seen just took one aspect of one part of one data structure and ran with it. This doesn’t mean “pick the easiest part and just do that.” Rather, it means “find something about the topic you’re excited about, and go wild exploring just that part.”
Your explanatory article will be evaluated on the following metrics:

- **Clarity:** How clear is your exposition on the topic? Does it flow naturally in a way that guides the reader through a difficult idea, or does it force the reader to keep pausing to refer back to concepts that are described earlier? Does it contain spelling, grammatically, or formatting errors, or is it well-proofread? Remember that your goal isn’t to just brain-dump everything you’ve learned; it’s to explain how something works and why it’s worth learning about in the first place.

- **Intuition:** How well does your writeup communicate the intuition behind your topic? For example, if there’s some key technical lemma necessary for things to work properly, did you just present the lemma “as-is,” or did you help the reader contextualize what it says and understand why it’s true? If there’s some unusual choice of a constant, can you account for why that constant is chosen that way? And if there’s something that’s surprising, can you provide an explanation as to why, in retrospect, it might not be as surprising as initially perceived?

- **Rigor:** This is a theory class, after all, and so we will be looking at how rigorous you are in your writeup. You don’t need to prove everything, but you shouldn’t hand-wave away every result. Pick a few interesting bits here and there and really dive deep into them! This is something that, hopefully, you’ll already have done when reading up on the topic in the first place.

A major point that we need to address: your writeup should be your own words and your own presentation. Your writeup should not be a paraphrase of your sources. That is, if we were to line up the sources you’ve cited and the paper you wrote, we shouldn’t feel as though we’re reading the same thing twice. You should be aiming for synthesis rather than recall.

Regardless of what you do, you must cite your sources in your writeup, and not just if you’re taking a passage from somewhere else. Remember, it’s plagiarism to take another person’s work and present it as your own! This applies even if you aren’t literally copying text or figures. If your writeup essentially follows along with someone else’s explanation, you need to make that clear through appropriate citations.

You’re welcome to share drafts of your writeup with other CS166 students to get feedback, and in fact you’re encouraged to do so. Just make a note of it somewhere in your writeup.
The “Interesting” Component
The other major aspect of your project is the “interesting” component, where you’ll take your topic and do something “interesting” with it. There is no separate deliverable for your “interesting” component. Instead, you’ll describe it and report your findings in your explanatory article.

You have wide latitude in choosing your “interesting” component. Here’s a sampling of some of the stronger “interesting” components we’ve seen in the past. This list is meant to help you brainstorm ideas and is not exhaustive. Every year we’ve seen something new that totally blew us away!

- **Apply the data structure to solve a problem, then analyze and interpret the results.**
  - Using persistent B-trees (a variation on B-trees where all previous versions of the data structure are accessible after any update) to represent a series of snapshots of webpages, then analyzing in what cases they work well for compression.
  - Implementing an optimized version of the Burrows-Wheeler transform using processor intrinsics as part of a data compression pipeline, then analyzing the efficiency gains.

- **Implement the data structure, compare its performance against benchmarks or against theoretically-predicted behavior, and analyze the results.**
  - Exploring the potential function for splay trees on different workflows, providing empirical evidence for a published upper bound being tight.
  - Optimizing a learned index structure (a hybrid of a machine learning model and a Bloom filter) and running experiments to determine what makes it tick and what its limits are.
  - Developing an image compression algorithm based on splay trees and Hilbert curves, then comparing that compression algorithm against standard lossless compression schemes.

- **Prove a theoretical result about a data structure and use that proof to improve understanding.**
  - Proving that Robin Hood hash tables (hash tables where elements are pushed around so that few elements are in bad places) satisfy a key structural invariant, then using that invariant to simplify the implementations of the key hash table operations.
  - Determining that the costs of accesses in a $k$-d tree (a multidimensional search tree) are related to parabolas in metric spaces, and using this to explain empirical performance data.

- **Generalize a data structure, or explore why such a generalization isn’t typically used.**
  - Exploring how rewrite rules used to simplify majority/inverter graphs (used to represent logic circuits in synthesis software) can be generalized from 3-input gates to 5-input gates, and why doing so loses some of the elegance of the 3-input case.
  - Investigating how cuckoo hash table performance degradation thresholds change as the number of tables and slots change.

As a note, regardless of what you choose to do,

> prioritize depth over breadth.

In other words, we would much rather see a deep dive into a narrow area than a broad but shallow survey. Some of the best “interesting” components we’ve seen have looked at just one small part of a data structure, but in a way that totally changed our conception of how that data structure behaved.
Your “interesting” component will be evaluated on three metrics:

• **Creativity:** How original is your idea? Are you breaking new ground, or are you doing something fairly standard? Simply coding up a data structure is likely a low-creativity endeavor, though if you use that implementation in a novel way, that could be high-creativity.

A good way to come up with a high-creativity project is to ask questions like “why isn’t this data structure used in this place I would have expected to have seen it?” or “why don’t we do this some other way?” These questions typically lead in promising directions, as answering them requires you to dive deeper into an area you might not have explored.

• **Difficulty:** How challenging was your project? If you arrived at a novel result or produced a new perspective on a topic, that’s almost certainly high-difficulty (if not, someone else would have already done it!). If you took off-the-shelf implementations of different data structures and ran them against each other, it’s a little bit less impressive.

“Difficulty” doesn’t mean that your final project is necessarily complicated. It just needs to be non-trivial. Doing a lot of work to arrive at a simple result will not result in you getting a low difficulty score. Picking something you know is going to be easy from the get-go likely will.

• **Insight:** How much did we learn from your “interesting” component? If you’ve exposed a hidden structure inside of a well-known data structure, or demonstrated convincingly why one particular algorithm runs quickly, or fundamentally shifted our perception about how to think about problem-solving, that’s high-insight. If you confirmed that a slow data structure is slow without diving into why it’s slow, that’s low-insight.

Want to come up with a high-insight final project? Before you decide what to do, make a list of questions you haven’t answered to your own level of satisfaction about your topic. Then ask – what could we do to answer this question? Projects chosen this way typically end up with some very cool insights at the end.
**Deliverables: The Project Proposal**

To ensure that we don't end up with too many people covering the same topics in their final projects, your first step in working on the final project is to submit a list of project proposals.

First, determine who you'll be working with. You are required to work in a group of four unless you receive prior approval from the course staff (that is, we specifically authorize you to work individually or in a smaller group before you submit the project proposal). This requirement is logistical – we simply don't have the capacity for everyone to work in smaller groups. By working in a group of four, you also give us the flexibility to support students who, due to their circumstances, don't have the option to do so.

Next, we'd like you to list, in ranked order, four choices of topics for your final project. We'll give you a list of suggestions in another handout, but you're welcome to choose any data-structure-related topics you'd like as long as the following are true:

- **The topic isn't something we've already covered.** However, you are welcome to pick topics that follow up on lecture or problem set topics. For example, feel free to tell us about progress on RMQ since the Fischer-Heun paper, or about improvements to SA-IS since the original 2008 paper, or on B⁺-trees as an optimization of B-trees.

- **That topic is focused on a specific data structure or research paper.** Every project needs a clear focus, and we've found that the best way to achieve this is to ground each project on a particular data structure or paper. For example, “binary search trees” would be too general, since there are hundreds of variations on this theme. A project on the paper *The Geometry of Binary Search Trees* is very reasonable. “Purely functional data structures” is too broad, since that's a whole area of research. A project on “purely functional red/black trees” is at the right level of detail.

For each topic that you rank, we'd like you to find at least two sources on that topic as part of your project proposal. One of those sources must be an academic paper, and the other can be any source that you'd like. We recommend finding a set of course notes or a follow-up paper as your second source; these often do a good job summarizing the key ideas from the original paper.

Once we've received everyone's submissions, we'll run a matchmaking algorithm to assign topics to teams. To ensure that we have a wide array of topics presented, no two teams will be assigned the same topic. In case you're curious, we'll be using a stable marriage algorithm so that no two teams have a mutual incentive to swap project topics. Once your topic is assigned, you will need explicit permission from the course staff to change it – again, this is so we don't end up with duplicate topics.

Some things to keep in mind:

- Take the source-gathering seriously. You will ultimately end up becoming an expert on your topic, and some topics are much easier to read up on than others. Don't just find a paper and call it a day – spend some time reading over that paper to make sure you have a sense of what you'd be signing up for and how good a resource it is. If you haven't read a research paper before, we highly recommend checking out Keshav's meta-paper *How to Read a Paper*.

- Because we're using stable marriage, if your top-ranked pick is something that no one else in the course proposed, you are guaranteed to get that topic. Every quarter we've had several new data structures proposed and discussed, and they're often quite exciting. However, be careful not to pick something obscure for the sake of its obscurity – often times, there are compelling reasons why some topic isn't more mainstream. 😊

- If you're unsure whether something would qualify as a good topic, please feel free to ask us! We're happy to weigh in and offer suggestions. We'd recommend doing this before the proposal deadline so that we have time to do some reading and get back to you.

There's no specific grade associated with the project proposal, but you're required to submit one so that we know what topics folks are working on.
Deliverables: The First Checkpoint

Once you have topics assigned, you'll need to start working to build a mastery of your topic. Your first deadline is the checkpoint, which consists of four milestones:

- **Submit a progress report.** Write a brief (1,000 – 2,000 words) summary of your progress in understanding the topic. In particular, explain the parts of your topic that you understand as well as you can explain them, and identify the areas you don't yet understand and talk through what you've done to try to understand them. The goal of this report is to give us a sense of where you stand and what specifically you'll need help on.

- **Address questions from the course staff.** Around the time you receive your topic assignment, the course staff will reach out to you with questions about your particular topic. These questions are designed to help you focus your efforts as you're learning more about that topic. Your checkpoint should include the best answers to the questions that you're able to provide. If you're able to answer the questions, great! If you have a hunch about the answers but aren't sure, that's okay too! Write down what you've thought about so far and what you're blocked on. If you are completely stuck, no worries! Tell us, in detail, all the leads you tried out and why you keep getting stuck.

- **Ask the course staff questions.** In the course of doing the above steps, you will likely have come across some questions you don't have the answer to. Write out a list of technical questions you're not sure about. We can then review them and offer some input and advice as you're moving forward.

After you’ve submitted your checkpoints, the TAs and I will review what you’ve submitted and set up a time to meet to discuss it. That will let us offer suggestions and advice about how to proceed and provide an opportunity for you to chat with us as well.

The first checkpoint is graded on an S/NC basis. We're expecting to see a solid effort and some good progress toward understanding your topic, as well as a draft of the explanatory article that shows you have some nontrivial handle on your topic. You are not required to have a mastery of your topic, nor are you required to have figured out all the details. If you've put in a solid effort into each part and taken things seriously, you should be in good shape even if you're struggling to understand your topic.

If you earn an NC here, we'll tell your group what it should do to make that up in the second checkpoint so that you have an opportunity to make things up.
Deliverables: The Second Checkpoint

The second checkpoint is designed to let you show us what you’ve done for your “interesting” component. Your deliverables are

• **A draft of your explanatory article.** You should submit a rough draft of your explanatory article. It will likely be incomplete because you’ll still be working on your “interesting” component, but should start to take on the contours of its final version. Aim to do the best job you can here; the TAs and I will offer feedback on this draft and the more work you put in the more useful our feedback will be.

• **A progress report on your “interesting” component.** This should be a brief (1,000-word) writeup summarizing what you’ve done so far for your “interesting” component, along with plans for your next steps. This will give us an opportunity to check in with what you’re doing and offer some suggestions and guidance before you finalize things.

As before, we’ll review your checkpoint and meet with your team to offer feedback and suggestions going forward. Our expectation here is that you’ve made nontrivial progress on your “interesting” component and have gotten your explanatory article into reasonable shape. If that’s the case, we’ll give you an S grade here. If not, we’ll give you an NC grade here and explain what you’ll need to do with your final submission to make that up.

Deliverables: The Final Submission

At the end of the quarter, you’ll submit your final explanatory article. This should include a summary of your “interesting” component in addition to the general description of your topic. And that’s it – you’re done! Congratulations on finishing your project!