Welcome to CSCI 1010!

An experimental journey through through data science, machine learning, and generative AI

Profs. Michael Linderman and Phil Chodrow
Department of Computer Science
Middlebury College

Learning From Data

We are surrounded by systems that **learn from data**:

- **Amazon** makes recommendations for things you might want to buy based on your browsing and purchase history.
- **Netflix** suggests movies and TV shows based on what you’ve watched before.
- **TikTok** curates your feed based on your interactions and preferences.
- **GPT/Copilot/Gemini/Claude** create text, code, images, and music in response to your prompts after training on huge datasets.

It’s 10am.

How many times have you interacted with an automatic recommendation, a personalized ad, a curated feed, or a generative AI model *since you woke up today*?

What is this class about and who is it for?

Data science, machine learning (ML), and generative artificial intelligence (GenAI) rely on computational methods for learning patterns in large, complex, datasets. We teach those methods here! But...

That coursework is often locked deep in the relevant majors. This course brings it to the “200” level. The goal is to build practical skills now and a foundation for building a deeper understanding in “CSCI311: Artificial Intelligence”, “CSCI451: Machine Learning”, “CSCI457: Natural Language Processing”, etc.

Is this course for you?

> This course is designed for students with introductory programming or data science preparation seeking to build foundational understanding and practical skills in machine learning and generative AI.

- ✅ You have some programming experience at the 100 level (in any language)
- ✅ You are curious about data science, machine learning, and generative AI
- ✅ You want to build practical skills and foundational understanding in these areas
- ✅ You want to explore how these technologies are impacting society and could be used (responsibly)
- ❌ You have already taken courses in data science or machine learning at the 200-level or above (e.g., STAT211, STAT218, STAT311, CSCI451)
- ❌ You are looking for a course that will teach you how to write good prompts for generative AI systems

Slide 4 Notes

If you have already taken courses like CSCI451, STAT218, etc. you have already advanced past this course! If that describes you, we want to supportively encourage you to take another class. You will likely be bored for a majority of the time, and we don't want that for you.

We have diverse backgrounds in the room

- ✅ Coming from R? We’ll learn how to translate your skills to Python.
- ✅ Never worked with a Data Frame? We don’t assumed prior experience with the Python data science ecosystem.
- ✅ ML, what? No problem, we will start from scratch.
- ✅ Some ML experience? Great! Let’s build on that foundation.

Slide 5 Notes

We have a wide variety of backgrounds:

- Some of you have lots of CS and Python experience, but may have not worked with its data science tools
- Some of you have data science experience in R, but not Python
- Some of your have machine learning (statistical learning) experience others don't

That is a purposely broad set of backgrounds. We will learn a lot, and do so with and from each other!

To these ends, we expect everyone to:

- Be responsible for independently picking up the details of unfamiliar tools or technologies.
- Put in the effort to make sure you don't get left behind. Use the resources on the course page (and others) and make sure to ask us and others for help when you need it. Don't be the person that can't contribute because you don't know what is going on!

If you have prior experience, we expect you to:

- Use your knowledge to actively help your classmates, not to sit back in judgement or frustration. Recall the often the best way to learn is to teach (e.g., "see one, do one, teach one"); you will get more out of the class if you actively engage with all of your classmates, including those with less experience.
- Understand you can't do it alone. It may seem like you can do the assignments or the project better or faster by yourself, but the end product will actually be worse if everyone can't (or doesn't) contribute.

What does it mean to be a ‘Responsible Computing’ course?

We should all “be able to recognize, identify, and make informed judgments about societal and ethical implications of the uses and development of computing technologies”.[1]

This is an interesting and uncertain moment in computing and society. As an industry, we are developing *and deploying* powerful new technologies at a breakneck pace. We have a responsibility to understand and consider the implications.

[1] Middlebury Computer Science Learning Goals

Slide 6 Notes

As a practical matter this means we will specifically talk about and consider the ethical, societal, and environmental implications of computing technologies. In both class discussions and assignments, we will actively interrogate potential sources of bias, harm and disparity in data and models, and consider ways to mitigate those issues. In your course project, you will be expected to actively consider and to, the extent possible, address the ethical, societal and environmental implications of your work.

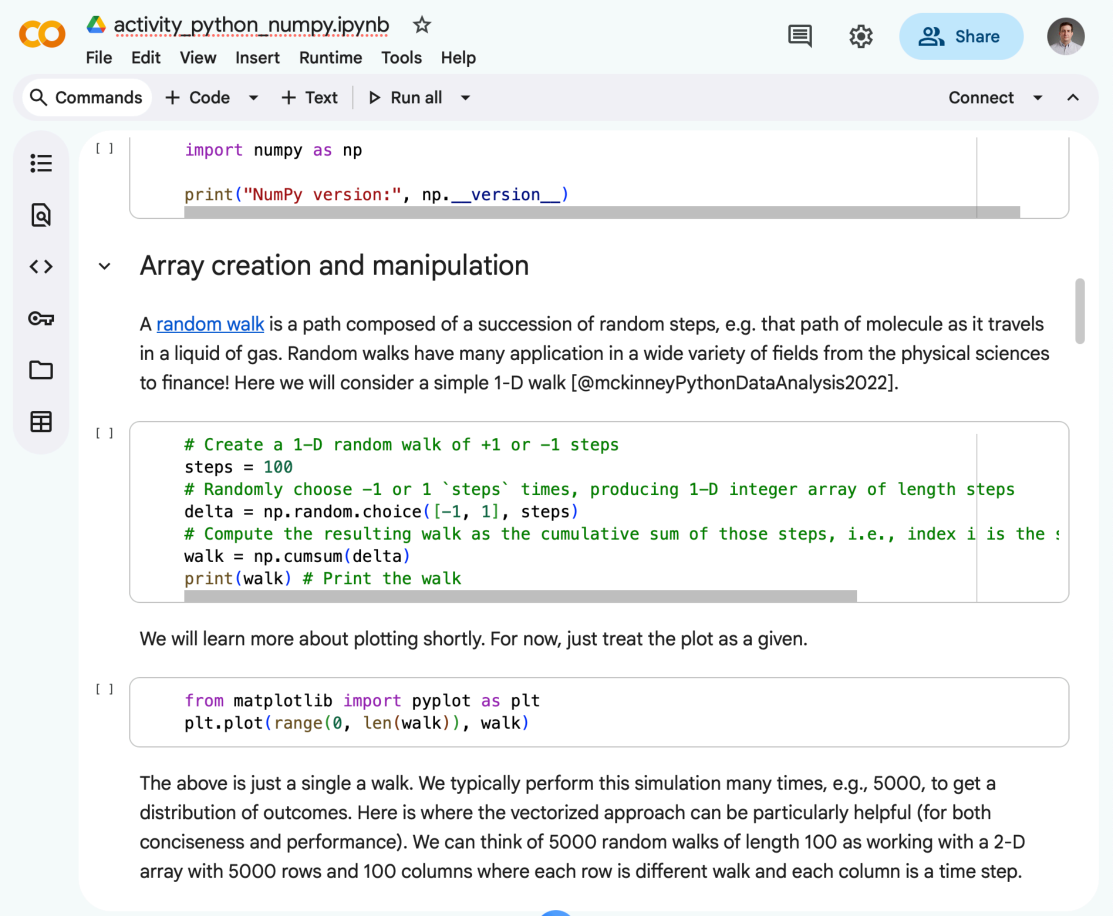
This not an aspect with right and wrong answers. Ultimately, you will have to decide for yourself what is right and wrong. Sometimes that answer might be easy, other times reasonable people might disagree. What, if any, tradeoffs might exist for/around interesting technologies? Is the work interesting or important enough to justify potential negative consequences?

To give you a concrete example, some of you may have heard of Joe Redmond. He is one of our more prominent CS alums. He developed the YOLO deep learning object recognition system which is now used in many applications. He walked away from the entire field of image recognition over concerns about the military applications.

We are *not* trying to encourage you to make any particular choice or choices. The choices you make will and should be personal. However, *we do* think it is your and all of our responsibilities to make a considered choice. Practicing *Responsible Computing* requires thinking about society and the consequences of the work we do. As Kranzberg's First Law of Technology states: "Technology is neither good nor bad; nor is it neutral." That is technology is only as neutral as its creators, which is to say not "neutral" at all. The work we do in here is not somehow separate from the world, but very much a part of it. Thus, it can both reflect society's challenges, however you define them, *and* be a vehicle, a powerful one, for effecting change. Thus what we are really asking you to do is not put your values down when you pick up your keyboard.

Computing environment

We will primarily use Jupyter/IPython notebooks implemented via Google Colab (https://colab.research.google.com)


activity_python_numpy.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
Connect

```
import numpy as np

print("NumPy version:", np.__version__)
```

Array creation and manipulation

A random walk is a path composed of a succession of random steps, e.g. that path of molecule as it travels in a liquid of gas. Random walks have many application in a wide variety of fields from the physical sciences to finance! Here we will consider a simple 1-D walk [@mckinneyPythonDataAnalysis2022].

```
# Create a 1-D random walk of +1 or -1 steps
steps = 100
# Randomly choose -1 or 1 `steps` times, producing 1-D integer array of length steps
delta = np.random.choice([-1, 1], steps)
# Compute the resulting walk as the cumulative sum of those steps, i.e., index i is the
walk = np.cumsum(delta)
print(walk) # Print the walk
```

We will learn more about plotting shortly. For now, just treat the plot as a given.

```
from matplotlib import pyplot as plt
plt.plot(range(0, len(walk)), walk)
```

The above is just a single a walk. We typically perform this simulation many times, e.g., 5000, to get a distribution of outcomes. Here is where the vectorized approach can be particularly helpful (for both conciseness and performance). We can think of 5000 random walks of length 100 as working with a 2-D array with 5000 rows and 100 columns where each row is different walk and each column is a time step.

Slide 7 Notes

We will primarily use Jupyter/IPython notebooks. These are very similar to R markdown or other notebook-style computing environments you might have used in other settings. Compared to writing a stand-alone program or scripts, notebooks are designed to be interactive computational documents that combine code, visualizations and narrative text.

Google Colab permits sharing, reading/writing to Google Drive, and free access to GPUs for machine learning tasks. Hopefully the free-tier will be sufficient for our needs. If not, we can explore other options. We will also leverage other cloud resources as needed, such as APIs for large language models.

Course logistics

- The course website go/cs1010 is the authoritative source for the syllabus, schedule, assignments, etc.
- Classes will typically include a short lecture/demonstration followed by interactive data science activities in an IPython notebook
- Use go/cs1010-inclass for in-class polls/reactions and go/cs1010-qa for Campuswire Q&A
- Activities will be submitted via Gradescope, with an initial due date (typically 24 hours) for immediate feedback and a later due date for credit
- Pre-class readings to prepare for accelerated class sessions
- Regular written quizzes and a combined written and oral exam in week 3
- Final project culminating in a class poster session

Slide 8 Notes

This is both a brand new course and a brand new kind of course for us (and the CS department). Please bring your sense of adventure and willingness to experiment. The final experience will almost certainly look different than the initial plan... Everything in the syllabus, etc. has an implicit “subject to change” attached. We will regularly solicit your feedback, and welcome comments, suggestions, ideas, etc. at any time.

Course Grading

You get a score (out of 100) for each of two buckets:

- **Understanding (U)**: Quizzes (35%), midterm exam (65%).
- **Application (A)**: Daily activities (35%), final project (65%).

Your final score is a weighted average of these two scores, closer to the lower one:

$$\text{Final Average} = 0.7 \times \min\{\text{U}, \text{A}\} + 0.3 \times \max\{\text{U}, \text{A}\}$$

We assign letter grades based on the scale shown in the syllabus.

Excellent presence in group work, participation with questions and discussion, attendance at Student Hours, etc. ***may*** *result in a 1/3 grade bump.*

(AI) Course policies (in an AI world)

- ✅ Working together
- ❌ Copying another person’s work in whole or part and submitting it as your own
- ✅ Searching online for docs, suggestions, StackOverflow, etc.
- ❌ Searching for or using previous solution to problem, even if freely available online
- ✅ GenAI (e.g., ChatGPT, Copilot) with citation
- ❌ Not citing work generated by someone or something else

You are intellectually responsible for all the work that you submit.

Slide 10 Notes

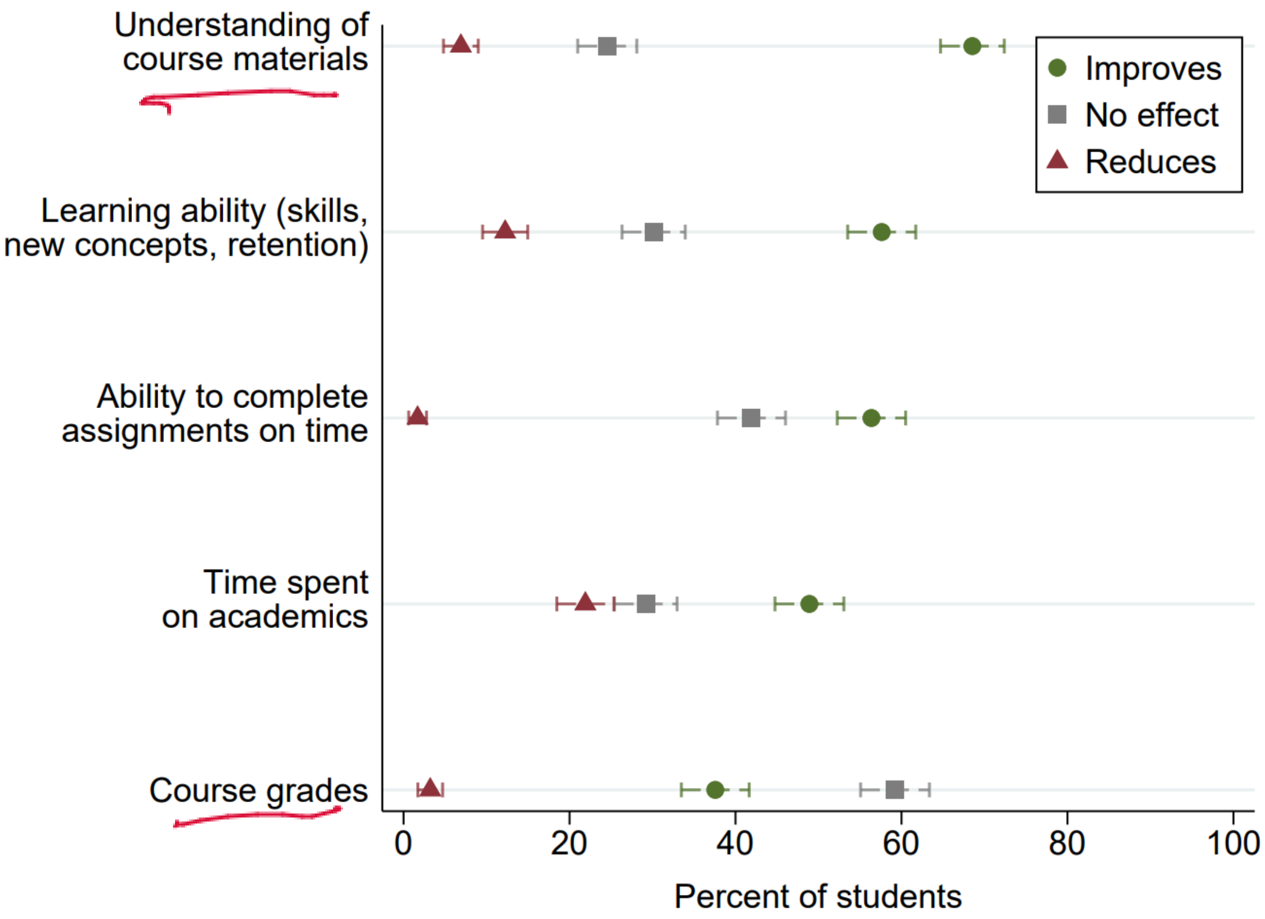
GenAI is a permitted resource in this course, and as described in the course catalog, use of GenAI, e.g., Google Gemini and other tools, will be part of the course content, assignments. If you are currently opposed to using those tools in any context, or become so later in the semester please come talk to us immediately about this is the right course for you.

You are intellectually responsible for all the work that you submit for assessment. You must understand and be able to explain the code and experimental results you submit. You will complete paper-based assessments (quizzes and exams), an oral exam and orally present your project work without access to reference materials, AI systems or other outside resources. Successfully doing so requires a real understanding of the material and any submitted work.

All submitted work must be properly attributed. **Your work must not contain unattributed parts written or generated by someone or something else (person or AI system).** Misrepresenting another person's work or AI system's generated output as your own original work is an academic integrity violation. Clearly cite all sources and acknowledge all contributors (both people and AI systems). In the particular case of AI systems, make sure you are meaningfully "digesting" any output by adapting, editing and testing it such that any submitted work reflects your own understanding. Note that "AI systems" are ubiquitous and include stand-alone tools (e.g., ChatGPT's web application), but also embedded tools (e.g.,, GitHub Copilot in VSCode, Google Gemini in Colab) and search results (e.g., AI results in Google search).

As an informal guideline, if you are proud of how you used AI tools and would eagerly share your experience with the instructors and your classmates that is an indication you are using those tools appropriately. You should be happy to cite your usage (and we are eager to learn from how you are using these tools!). If instead you find yourself hiding, or thinking about hiding, your usage that indicates you may not be using those tools appropriately.

Do LLMs help you learn?



Middlebury Student Beliefs about the Impact of AI on their Academic Performance. Figure 7 from Contractor and Reyes (2025).

Slide 11 Notes

Figure 7 from Contractor and Reyes, Generative AI in Higher Education: Evidence from an Elite College (2025), a survey of generative AI use among Middlebury students. The unsurprising top-line finding: "Overall, 82.5 percent of students report using generative AI for academic purposes,..." I suspect the current number if even higher.

Note the large fraction of students who reported that AI improves their "Understanding of course material" and other aspects of their learning. But at the same time, many fewer students reported that AI improved their "Course grades". We are curious about that difference. Do you think it because the grades (the underlying assessments) don't reflect learning? Is there limited "dynamic range" for grades to improve? Is self-assessment of improved learning overly optimistic? Something else?

These are self-assessments. The empirical evidence for externally assessed learning is mixed, and *very* limited:

- "...because LLM summaries lessen the need to discover and synthesize information from original sources —- steps essential for deep learning -— users may develop shallower knowledge compared with learning from web links. When subsequently forming advice on the topic, this manifests in advice that is sparser, less original—-and less likely to be adopted by recipients. Results from seven experiments support these predictions, showing that these differences arise even when LLM summaries are augmented by real-time web links, for example."

  Shiri Melumad, Jin Ho Yun, Experimental evidence of the effects of large language models versus web search on depth of learning, PNAS Nexus, Volume 4, Issue 10, October 2025, pgaf316, https://doi.org/10.1093/pnasnexus/pgaf316

- A study of GPT-4 use by high school math students reported that "unfettered" access to GPT-4 improves performance, but if that access is taken away "students actually perform worse than those who never had access".

  H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, & R. Mariman, Generative AI without guardrails can harm learning: Evidence from high school mathematics, Proc. Natl. Acad. Sci. U.S.A. 122 (26) e2422633122, https://doi.org/10.1073/pnas.2422633122 (2025).

- A recent, very preliminary study, conducted a RCT of 16 SW developers completing tasks on mature projects on which they have multiple years of experience. The headline results: Before starting developers estimated using AI will be 24% faster but in practice were 19% slower with AI. But I also note there are many caveats, possible artifactual explanations for these results and/or threats to validity and generalization.

  J. Becker, N. Rush, E. Barnes, & D. Rein, Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, arXiv 2507.09089 (2025).

- Profs. Reyes and Contractor performed subsequent randomized experiment with Middlebury students that indicated students who had access to generative AI tools outperformed those who did not on when completing a writing task on an unfamiliar topic.

Our “compact”

We (Profs. Linderman and Chodrow) are and will do everything we can, leverage every bit of our combined decades of experience, to help **you** learn.

We hope you will do everything you can to take advantage of us, the course materials, and your classmates to learn and grow (as a Data Scientist, a Computer Scientist, a person …).

Slide 12 Notes

We know everyone has different goals, different responsibilities, different constraints. We have tried to design a course that enables intentional choice making about your approach and your definition of success. However you can, we encourage you to "do the class"; to engage with the material, with us and with each other as fully and deeply as you can.

- Treat the assignments as the practice opportunities they are. Take a moment to reflect on the paths to your most cherished accomplishments; I suspect you didn't start off knowing how to do that thing. You grew into that accomplishment over time, through repeated practice and reflection.
- We are here for you (and don't judge). We want to talk with you about the class material, and anything else that interests you. The magic of student hours isn't in the answer to your initial question, but in the unplanned conversation that follows. We want to get to know you, and we hope you feel the same way.
- Get to know your classmates. I know that sounds "cheesy", but believe us that the relationships you build here will last far beyond this class (and likely the technologies we will talk about). Working together is *the* "pro-tip" for better learning, and a better life. We have the following request of you: Ask *a person* for help. Doing so is investing in your learning and your relationships. And if someone asks for your help, recognize that ask is about their specific question *and* about building a connection. Please don't respond with "What does ChatGPT say?" or similar. They are reaching out to you specifically because they value you and your perspective.

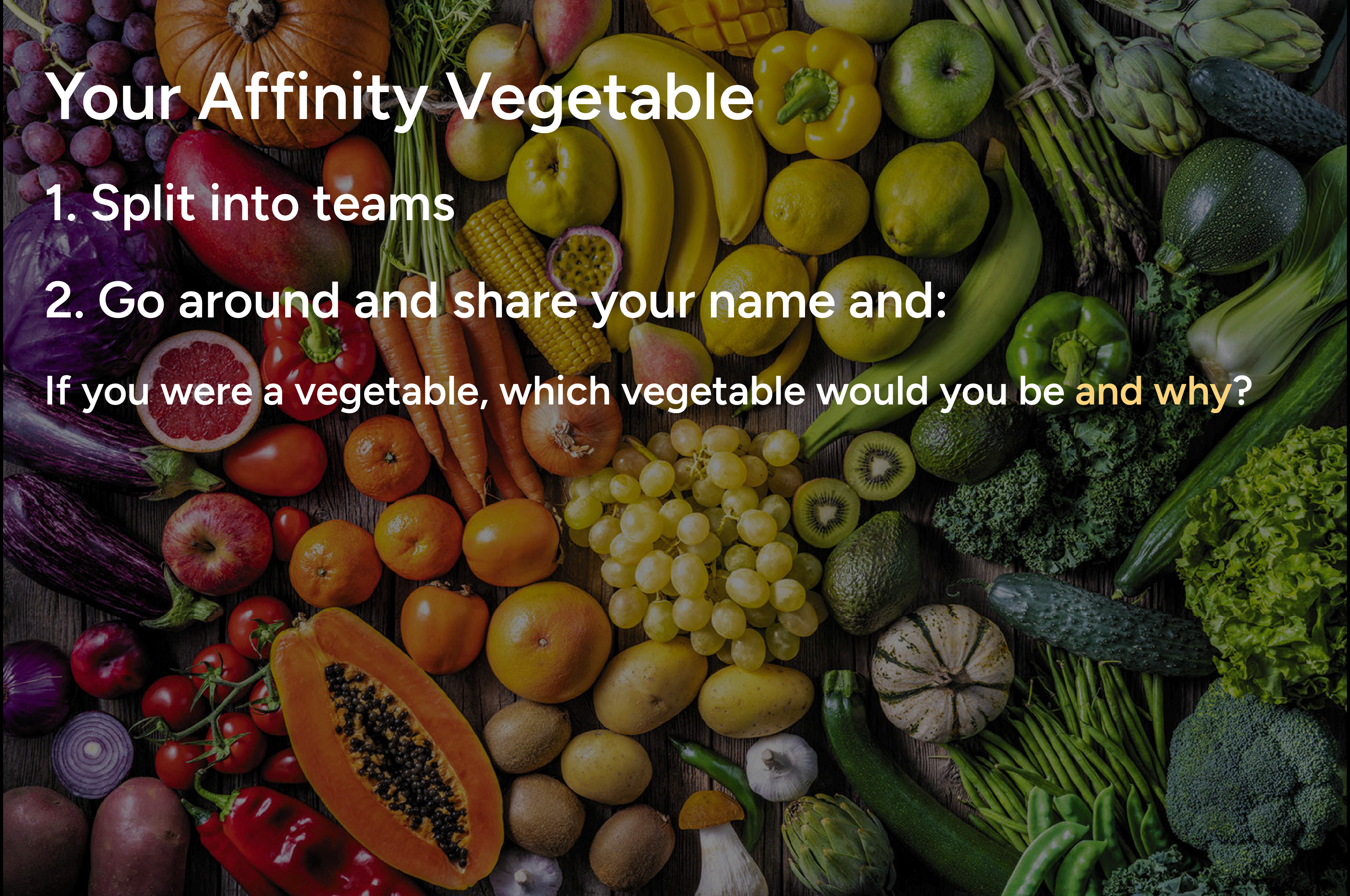
That last request is partly motivated by a recent interview study of undergraduate computing students help seeking which reported:

> Our findings suggest that help-seeking requests are now often mediated by generative AI. For example, students often redirected questions from their peers to generative AI instead of providing assistance themselves, undermining peer interaction. Students also reported feeling increasingly isolated and demotivated as the social support systems they rely on begin to break down.

We don't want that for you (and us)!

I. Hou, O. Man, K. Hamilton, S. Muthusekaran, J. Johnykutty, L.i Zadeh & S. MacNeil. 'All Roads Lead to ChatGPT': How Generative AI is Eroding Social Interactions and Student Learning Communities. In ITiCSE 2025, 79–85. https://doi.org/10.1145/3724363.3729024 (2025)

Your Affinity Vegetable



1. Split into teams

2. Go around and share your name and:

If you were a vegetable, which vegetable would you be and why?

Slide 13 Notes

This is not what vegetable you like to eat the most, but rather what vegetable describes you best. Your instructors:

PC: Onion ML: Zucchini

Your Affinity Vegetable

3. As a group identify a *delicious dish* that incorporates all of your vegetables.

Be ready to share!

References

Contractor, Zara, and Germán Reyes. 2025. “Generative AI in Higher Education: Evidence from an Elite College.” https://arxiv.org/abs/2508.00717.