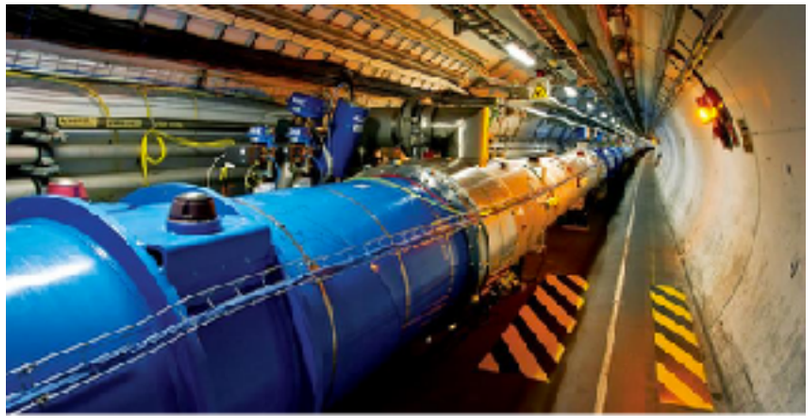
[301] Data Programming

Tyler Caraza-Harter

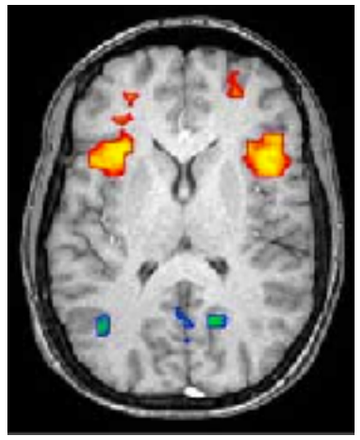
Welcome to Data Programming!

Data is exploding in many fields

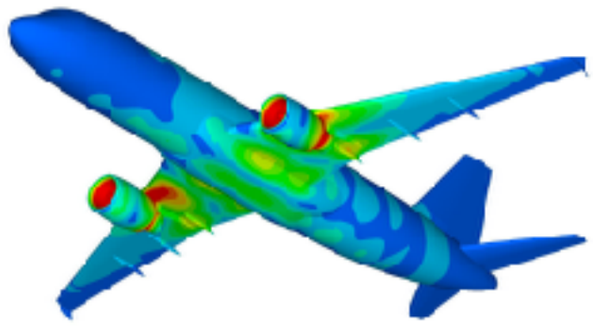
- Biology, physics, chemistry
- Psychology, sociology, economics, business
- Engineering (mechanical, electrical, industrial, etc)



https://home.cern/topics/large-hadron-collider



https://en.wikipedia.org/wiki/Neuroimaging



http://www.stressebook.com/finite-element-analysis-in-a-nut-shell/

https://science.howstuffworks.com/life/genetic/gattaca-gaptacaz-adding-letters-the-genetic-alphabet.htm

Welcome to Data Programming!

Data is exploding in many fields

- Biology, physics, chemistry
- Psychology, sociology, economics, business
- Engineering (mechanical, electrical, industrial, etc)

How can we gain insights from that data?

- With computation

Welcome to Data Programming!

Data is exploding in many fields

- Biology, physics, chemistry
- Psychology, sociology, economics, business
- Engineering (mechanical, electrical, industrial, etc)

How can we gain insights from that data?

- With computation

Approach 1: human computation

https://en.wikipedia.org/wiki/Human_computer

Welcome to Data Programming!

Data is exploding in many fields

- Biology, physics, chemistry
- Psychology, sociology, economics, business
- Engineering (mechanical, electrical, industrial, etc)

How can we gain insights from that data?

- With computation

Approach 1: human computation

https://en.wikipedia.org/wiki/Human_computer

Approach 2: machine computation

http://fortune.com/2015/11/15/intel-super-7/

Welcome to Data Programming!

CS 301 is about approach 2

- Faster, more reliable, can churn through more data

Approach 1: human computation

https://en.wikipedia.org/wiki/Human_computer

Approach 2: machine computation

http://fortune.com/2015/11/15/intel-super-7/

Welcome to Data Programming!

CS 301 is about approach 2

- Faster, more reliable, can churn through more data
- Requires being able to tell computers what to do!

society needs more domain experts in specific fields who can write code

Approach 1: human computation

https://en.wikipedia.org/wiki/Human_computer

Approach 2: machine computation

http://fortune.com/2015/11/15/intel-super-7/

Welcome to Data Programming!

CS 301 is about approach 2

- Faster, more reliable, can churn through more data
- Requires being able to tell computers what to do!
- Automate to save time

Approach 1: human computation

https://en.wikipedia.org/wiki/Human_computer

Approach 2: machine computation

http://fortune.com/2015/11/15/intel-super-7/

Welcome to Data Programming!

CS 301 is about approach 2

- Faster, more reliable, can churn through more data
- Requires being able to tell computers what to do!
- Automate to save time

“Find the leverage in the world, so you can be more lazy!”

~ Larry Page, Founder of Google

Approach 1: human computation

https://en.wikipedia.org/wiki/Human_computer

Approach 2: machine computation

http://fortune.com/2015/11/15/intel-super-7/

Why CS 301?

Common approach to introductory CS courses

- Use a programming language like C++ or Java
- Teach CS students and other majors together
- Emphasis on writing large programs (OOP, encapsulation) and theory (complexity analysis)
- Light on data

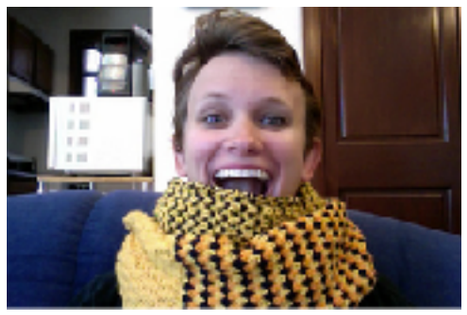
Why CS 301?

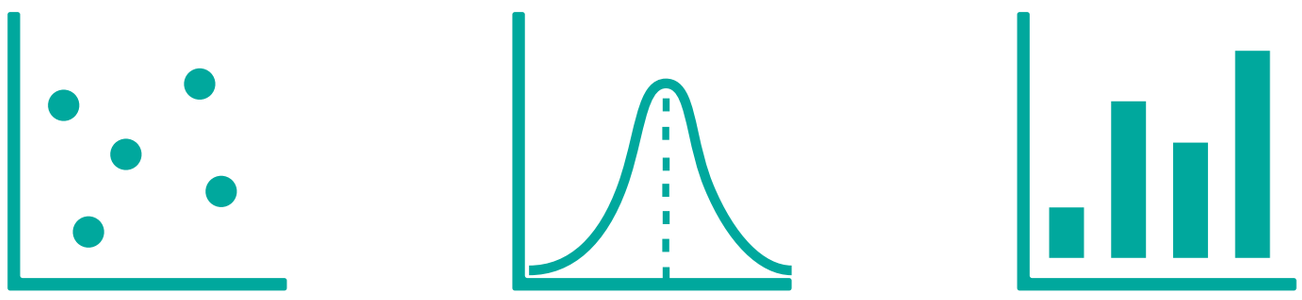
Common approach to introductory CS courses

- Use a programming language like C++ or Java
- Teach CS students and other majors together
- Emphasis on writing large programs (OOP, encapsulation) and theory (complexity analysis)
- Light on data

CS 301 approach

- Pioneered by Laura Hobbes LeGault
- Use Python (powerful but ~~easy~~ easier to learn)
- Goal: bring more programming into other fields
- Practical, minimal theory
- Emphasis on data, simulation, analysis, plotting





Today's Topics

Introductions

- Who am I? Who are you?

Website

Course overview

Computer basics

Who am I?

Tyler Caraza-Harter

- Email: tylerharter@gmail.com
- Just call me “Tyler”, no formalities necessary

Long time badger

- Did undergrad, masters, and PhD at UW-Madison
- Opportunity to teach classes I wish I could have taken

Return to teaching from industry

- Worked at Microsoft on SQL Server and Cloud
- Other internships/collaborations:
  Qualcomm, Google, Facebook, Tintri

Open-source projects

- OpenLambda project (Python-based platform)
- PivotLibre project (preferential-voting tool)

Who are You?

Year in school?

- 1st year? 2nd? Junior/senior? Grad student?

Area of study

- Natural science, social science, engineering, other?

How many have programmed before?

- Any language? Python? Taken a class?

Have specific datasets you want to leverage after 301?

Today's Topics

Introductions

Website

- Syllabus
- Schedule/calendar
- Datasets
- Lecture questions

Course overview

Computer basics

Course Website

There are 3 lecture sections for 301 this fall. I'm teaching 001 and 002 and Gerald is teaching section 003.

Website for my section (001 and 002):
https://tyler.caraza-harter.com/cs301/fall18/home.html

Walk through…

Today's Topics

Introductions

Website

Course overview

- Learning objectives
- Text book
- Class communication
- Grades
- Projects
- Exams

Computer basics

301 Learning Objectives

Learn basic Python

- Python is a good programming language for beginners
- We'll learn about input/output, functions, and flow of execution

Learn data structures

- When we have lots of data, we'll learn strategies for staying organized, by putting data in order (lists) or giving data names so we can find it easily (dictionaries)

Learn popular data formats

- We'll work with popular formats for sharing data, such as CSV, JSON, and HTML

Learn database basics

- A database is like a spreadsheet on steroids. We'll learn how to store data here and ask the database questions (called queries) to answer interesting questions

Learn how to create plots

- Plots and other visualizations are key to communicating well as a data scientist

Today's Topics

Introductions

Website

Course overview

- Learning objectives
- Text book
- Class communication
- Grades
- Projects
- Exams

Computer basics

Think Python, 2nd Edition

Note: Think Python does not cover all the topics we care about in CS 301. We'll provide other online readings, lecture notes, or slides to help with that material.

Why it's a good text

- Assumes no programming background
- It's very concise
- Extra problems for you to try (optional)
- It's free! (by the hardcopy if you choose)

Note on edition

- Get the 2nd edition, which is for **Python 3**!
- Don't get the 1st edition, which is for Python 2
- We'll be using Python 3 this semester (301 previously used Python 2)

Power of open source

- 1999: Downey wrote an introductory Java text
- 2001: Elkner translated it to Python
- 2003: Downey started using Python version
- 2016: Downey published 2nd edition
- Future for CS 301???

Today's Topics

Introductions

Website

Course overview

- Learning objectives
- Text book
- Class communication
- Grades
- Projects
- Exams

Computer basics

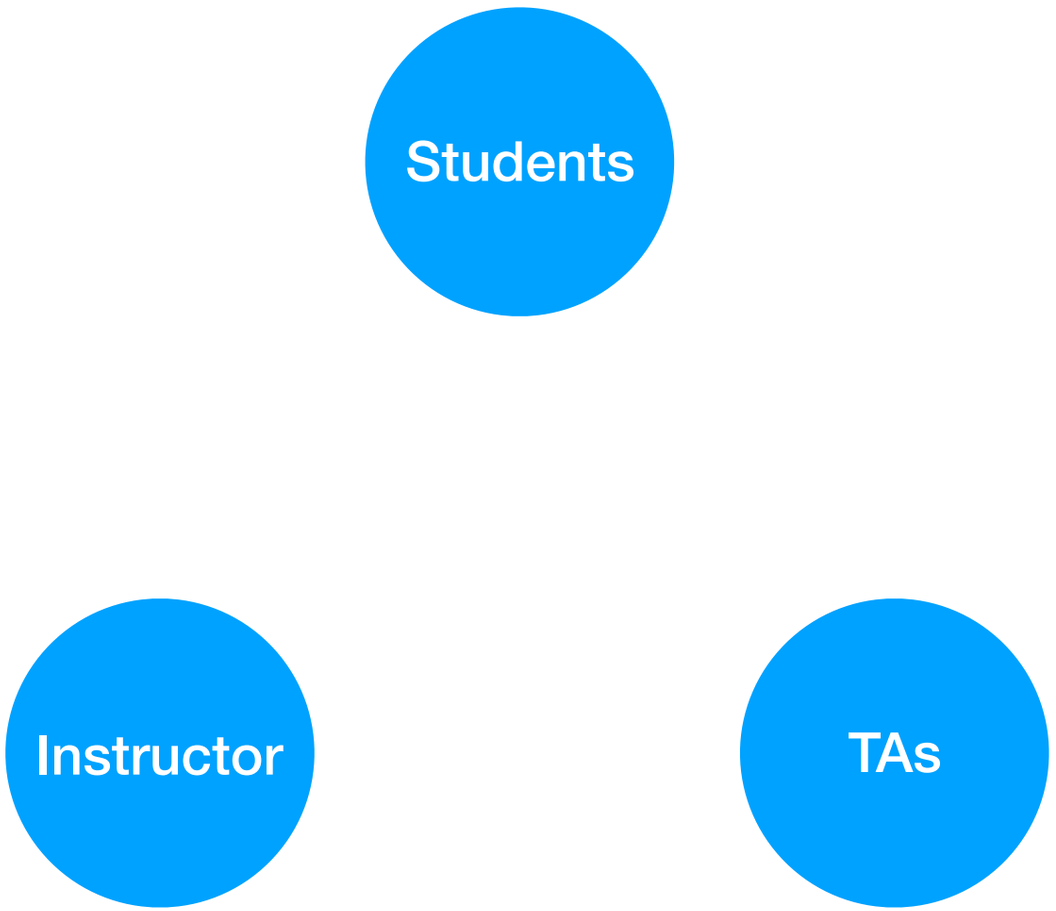
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



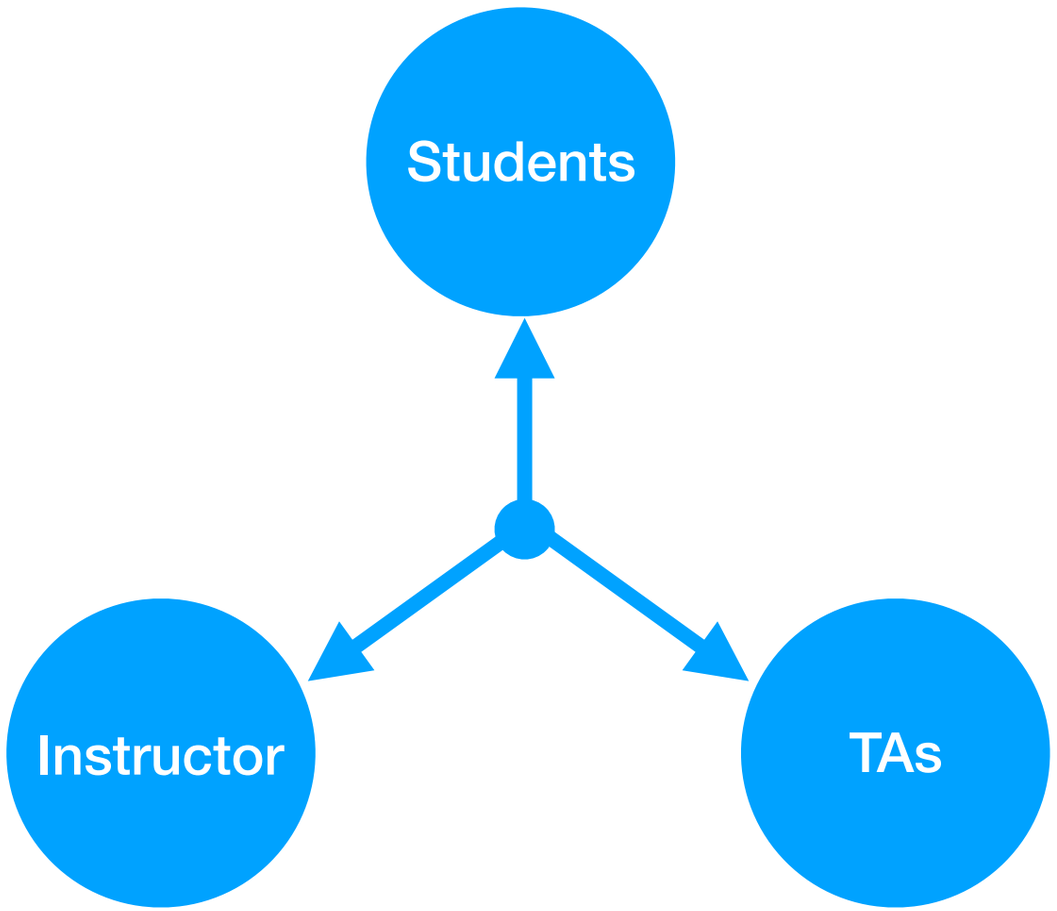
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



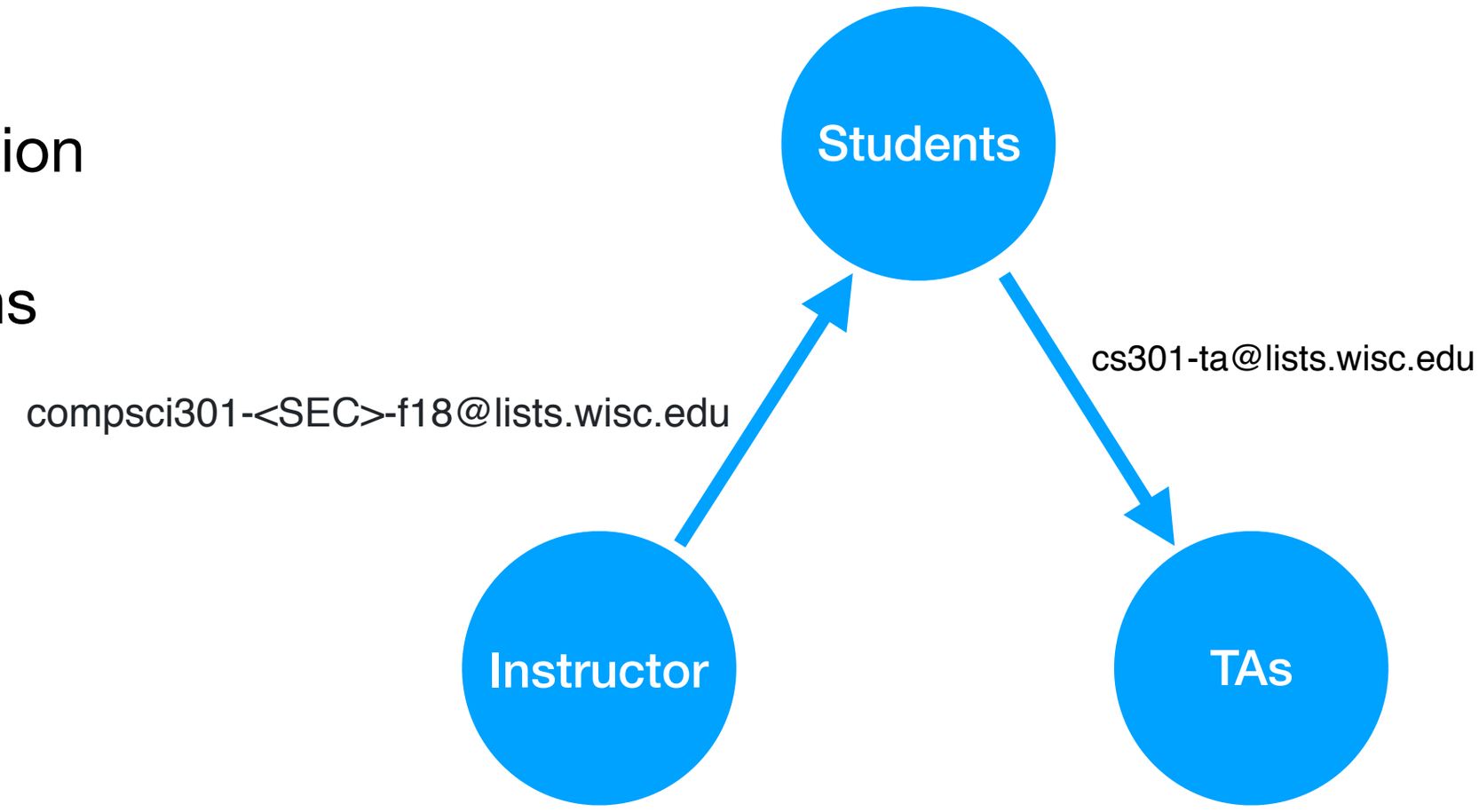
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



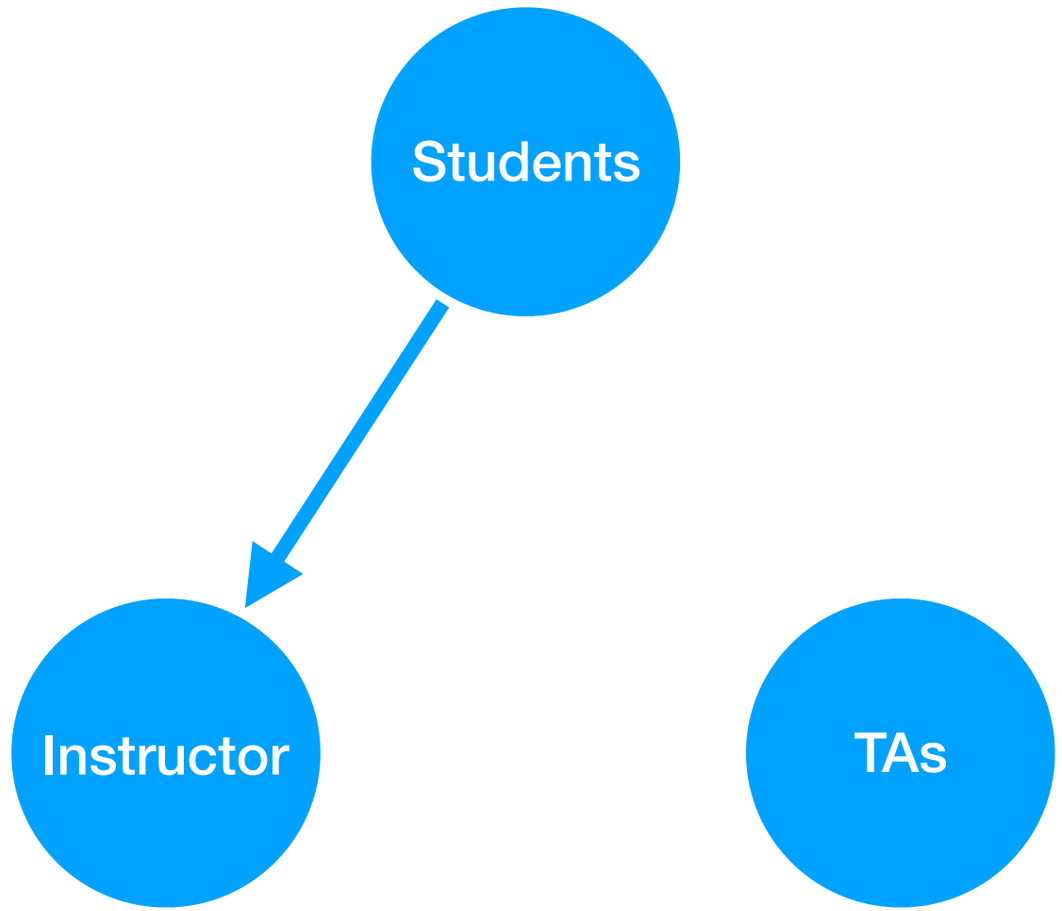
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



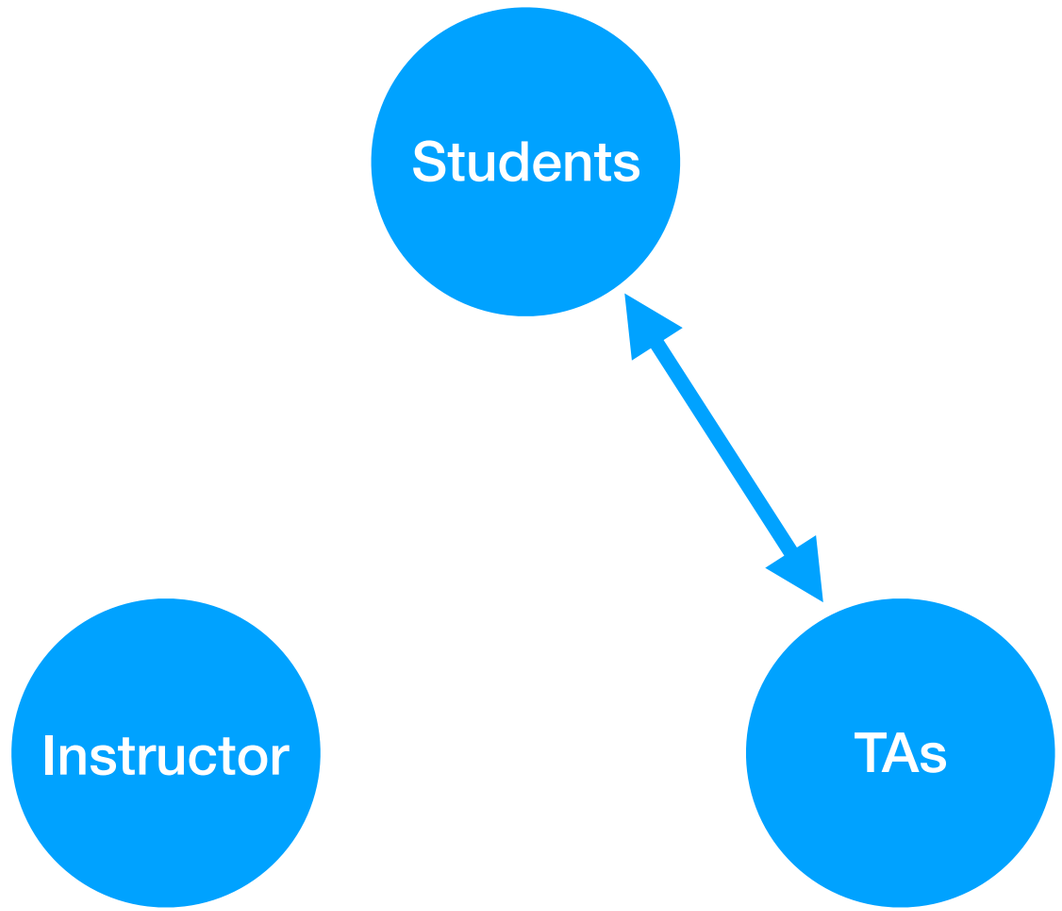
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



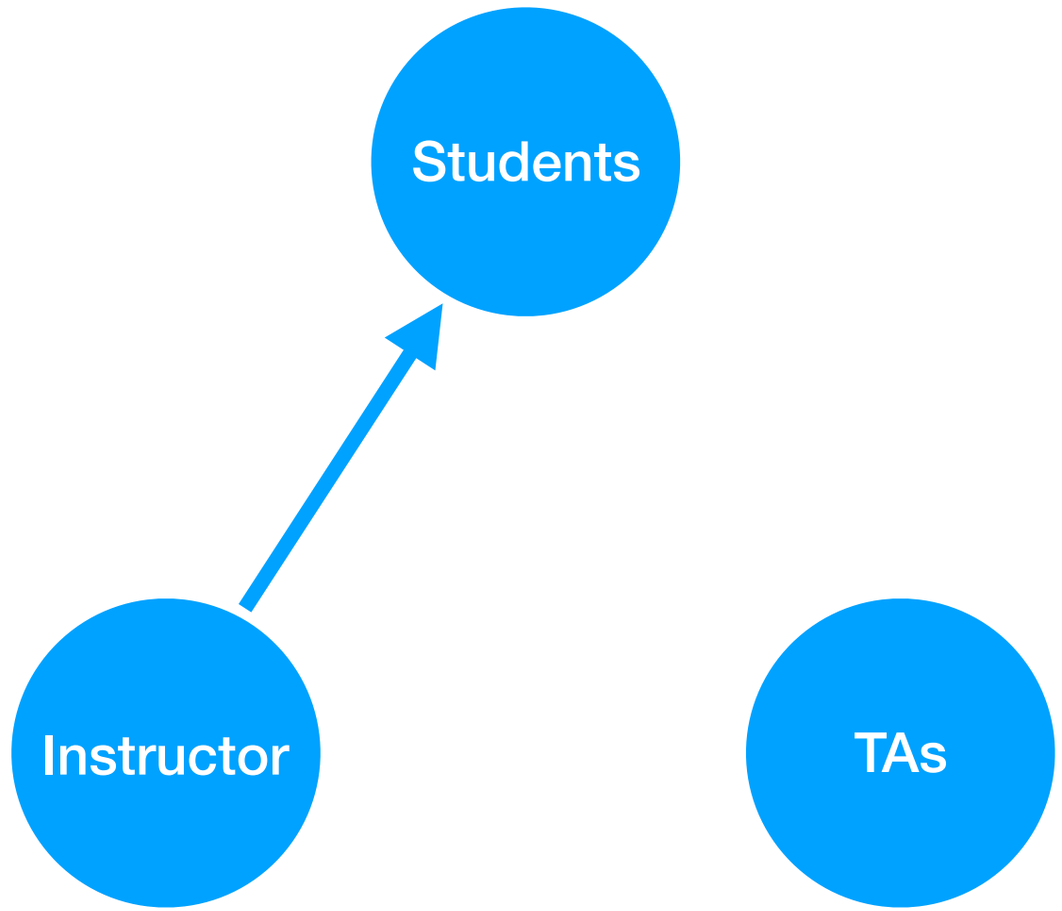
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



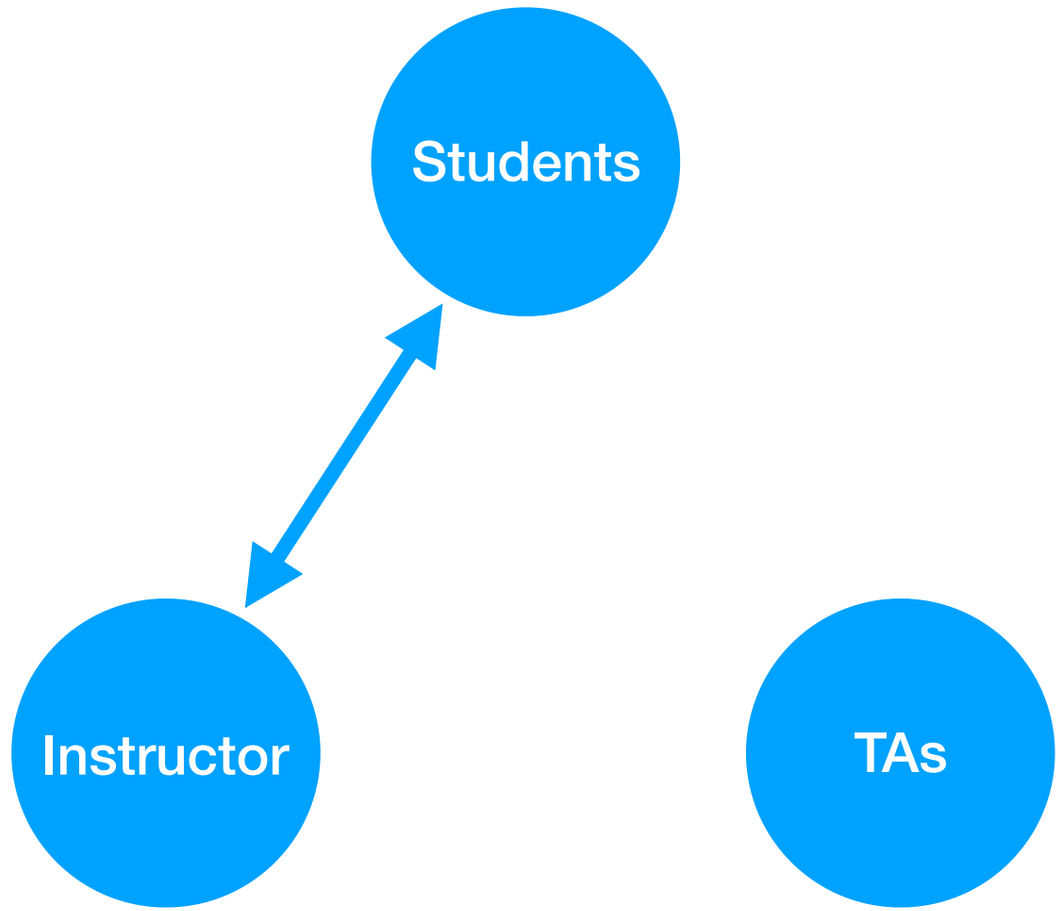
Communication is CS 301

Good communication is critical for a class of this size

- Who needs to communicate: students, TAs, instructors

Besides direct email, we'll use six communication tools

- Piazza
- Email lists
- Feedback Form
- Project Submission
- Canvas
- Clicker Questions



Today's Topics

Introductions

Website

Course overview

- Learning objectives
- Text book
- Class communication
- Grades
- Projects
- Exams

Computer basics

Grades

60% for programming projects

- 12 projects, each 5%
- automatic tests will essentially tell you what score you will get (with some minor exceptions)
- This is weighted heavily because **learning to write code well is our #1 concern in this course**

40% for exams

- 10% midterm 1 (in class)
- 10% midterm 2 (in class)
- 20% final

Final grading will be curved

Today's Topics

Introductions

Website

Course overview

- Learning objectives
- Text book
- Class communication
- Grades
- Projects
- Exams

Computer basics

Project Overview

Nearly all projects will relate to some dataset

- https://tyler.caraza-harter.com/cs301/fall18/datasets.html

Timeline

- Projects will be due most weeks, on Wed, at midnight
- Any lecture material necessary for the project will be covered the week before
- 10% penalty for turning it in late; not accepted after more than 1 week
- Under special circumstances (e.g., illness), ask me before, and we can discuss appropriate adjustments to your deadline

Getting help

- Piazza (don't share substantial code) or email (do share code)
- Monday lab sessions
- My office hours or TA office hours

Pair Programming

You can optionally work in pairs of two

- Partner with students in any of the three sections
- Change partners when you like (post to Piazza to find people)
- Work together at the same time in the same place.
- One person writes code while one watches. Share!
- Motivation: learning from each other, not splitting the work

Grading

Grading will be done with automated tests

- The tests will determine the exact grade, unless we see that the code does not follow the spirit of the assignment
- We'll share the tests
- Not getting 100% should never be a surprise

Today's Topics

Introductions

Website

Course overview

- Learning objectives
- Text book
- Class communication
- Grades
- Projects
- Exams

Computer basics

Exams

There will be two midterms and one final

- Check website for dates/locations
- One 8.5 by 11 in note sheet allowed only
- Exams will be multiple choice

Today's Topics

Introductions

Website

Course overview

Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Today's Topics

Introductions

Website

Course overview

Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Input/Output

I/O (stands for input/output)

- What are examples for human?

Input/Output

I/O (stands for input/output)

- What are examples for human?

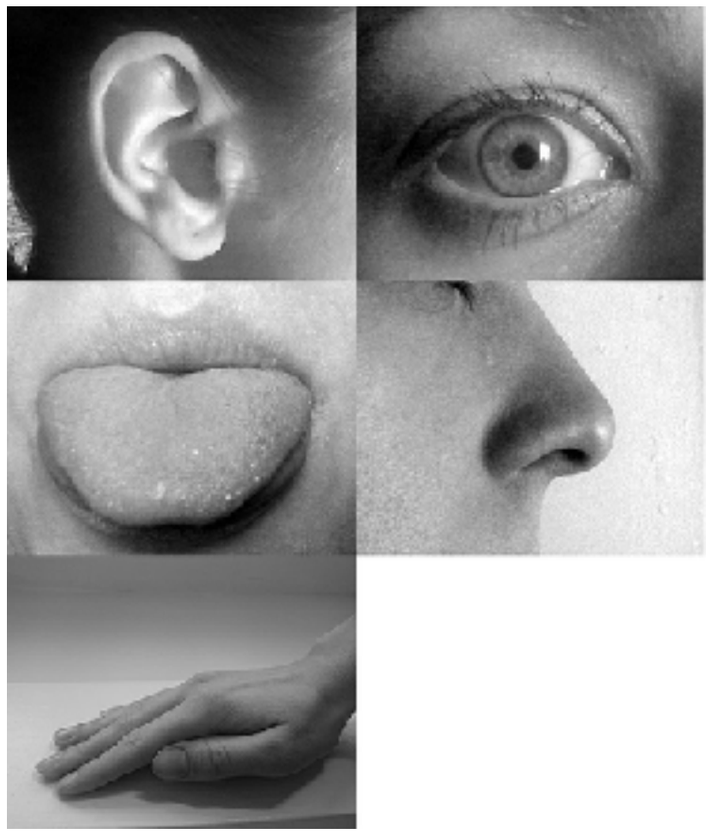
input: senses

Input/Output

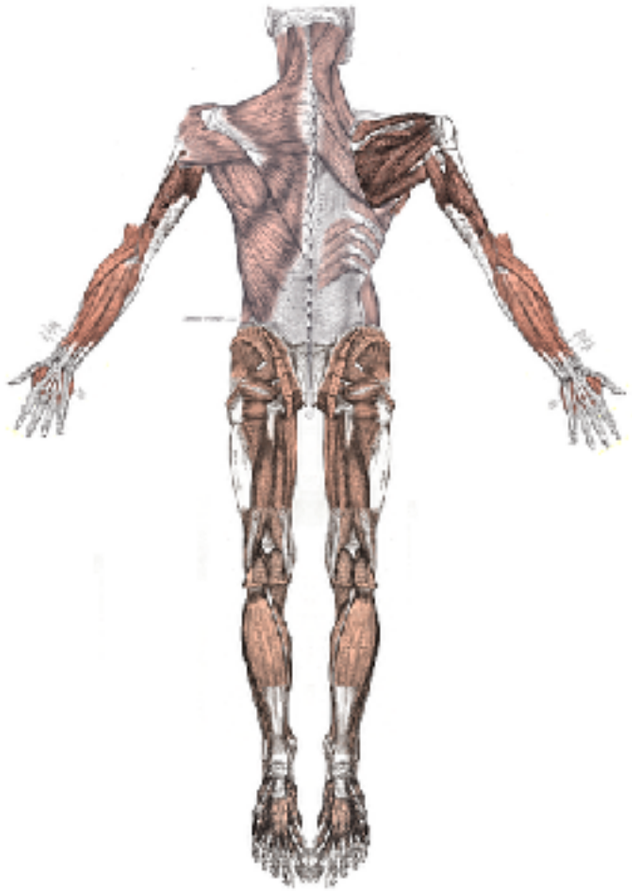
I/O (stands for input/output)

- What are examples for human?

input: senses



output: muscles

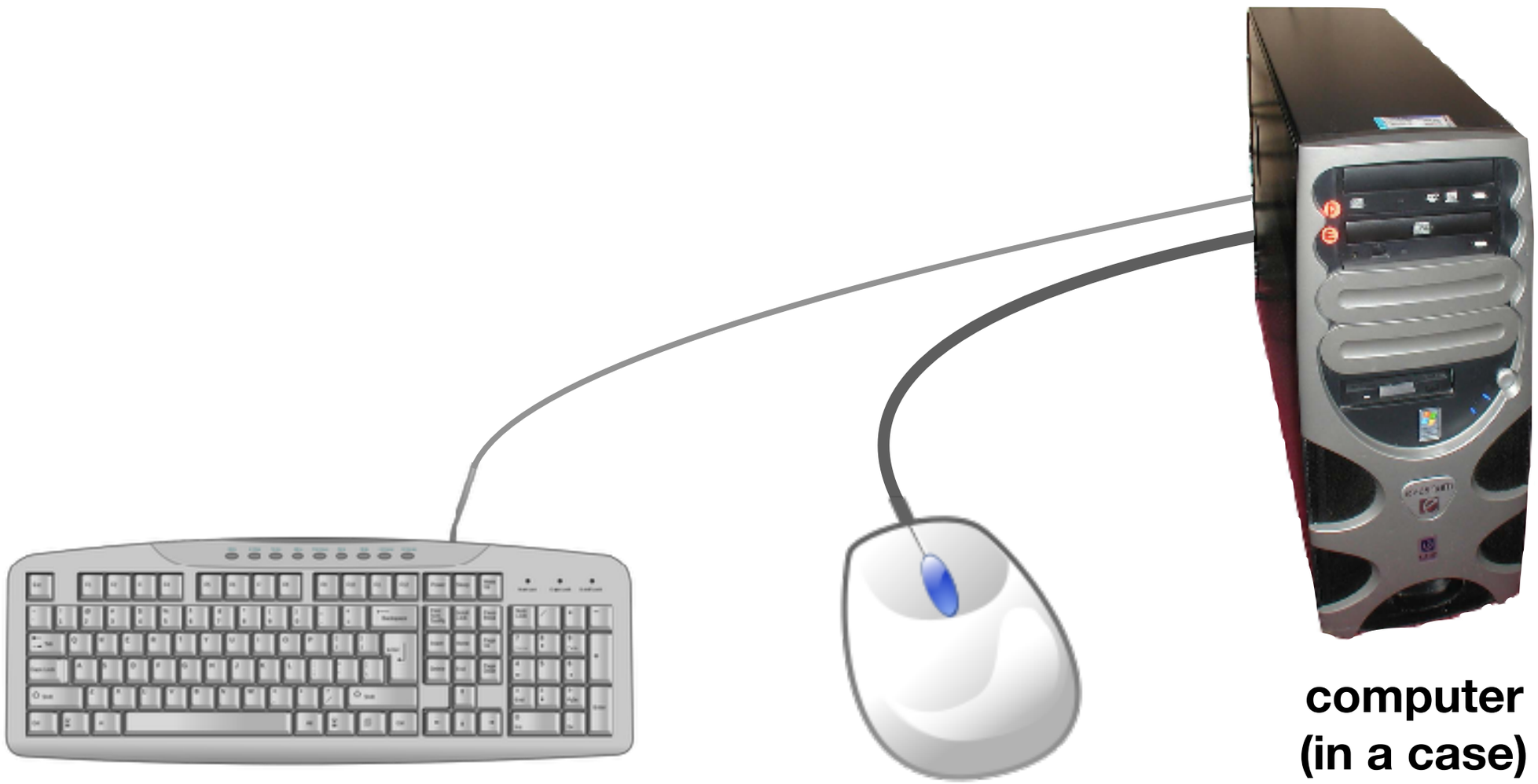


Computer Input/Output

computer
(in a case)

what are some common compute inputs?

Computer Input/Output



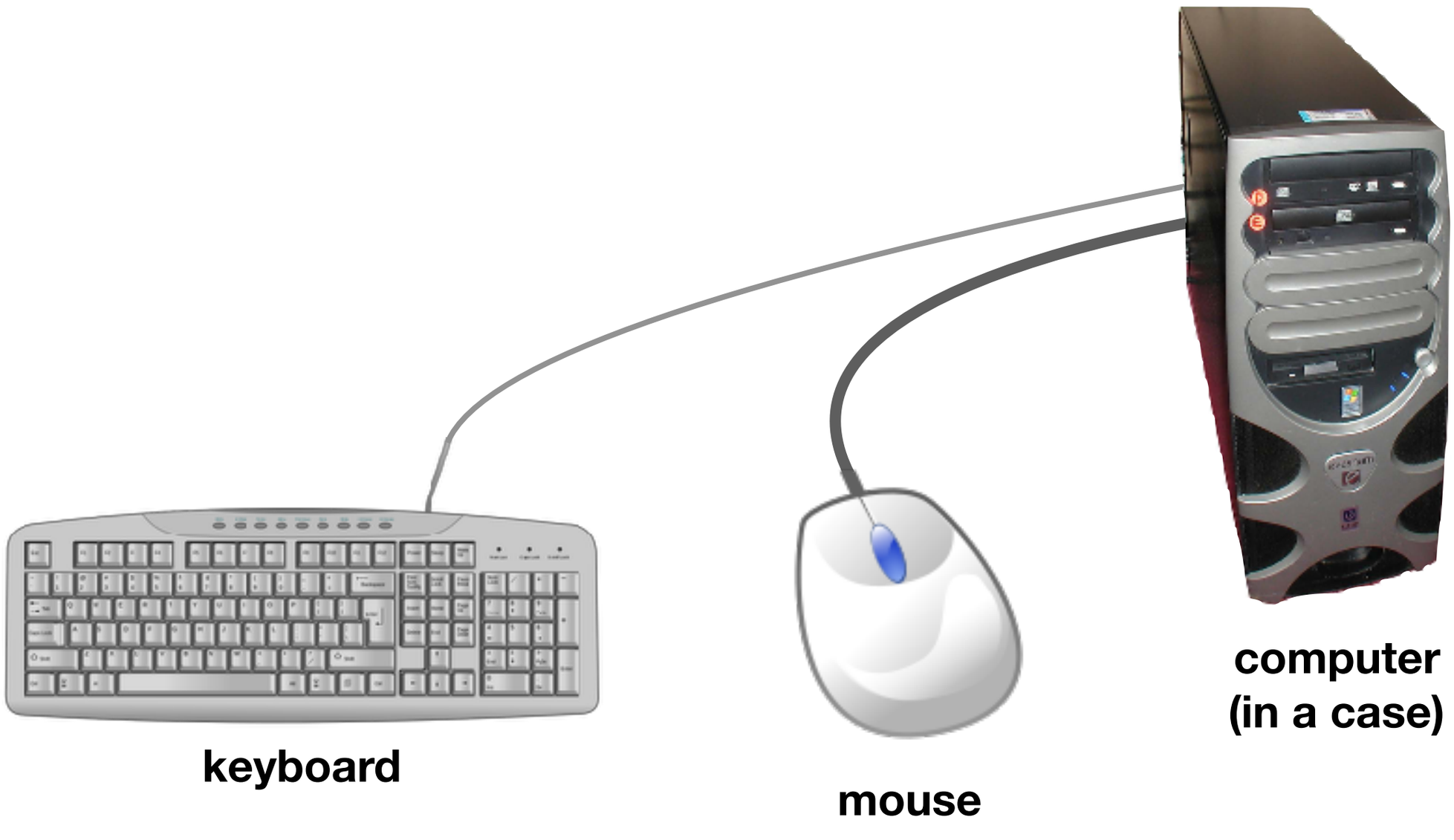
keyboard

mouse

computer
(in a case)

Computer Input/Output

what are some common compute outputs?

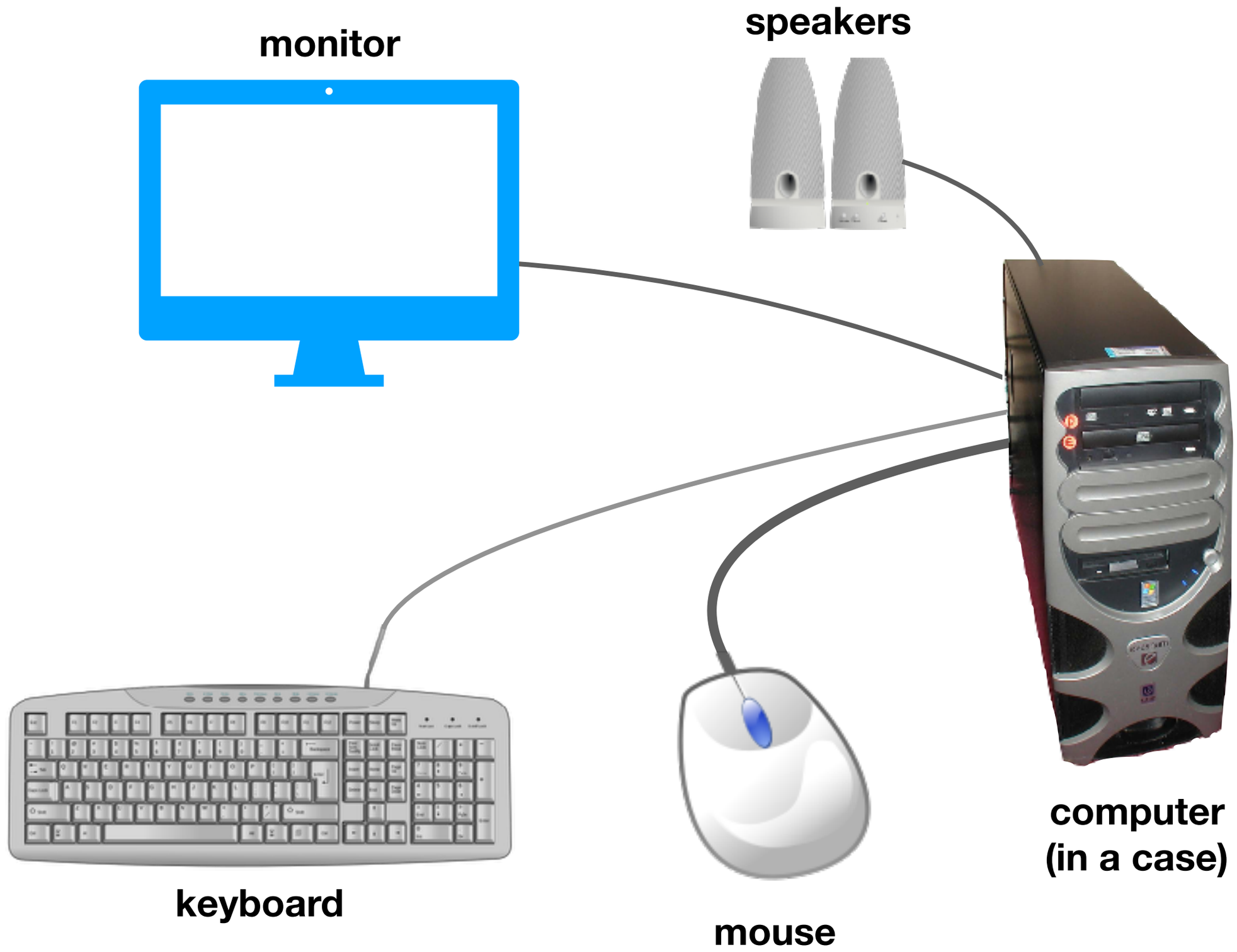


keyboard

mouse

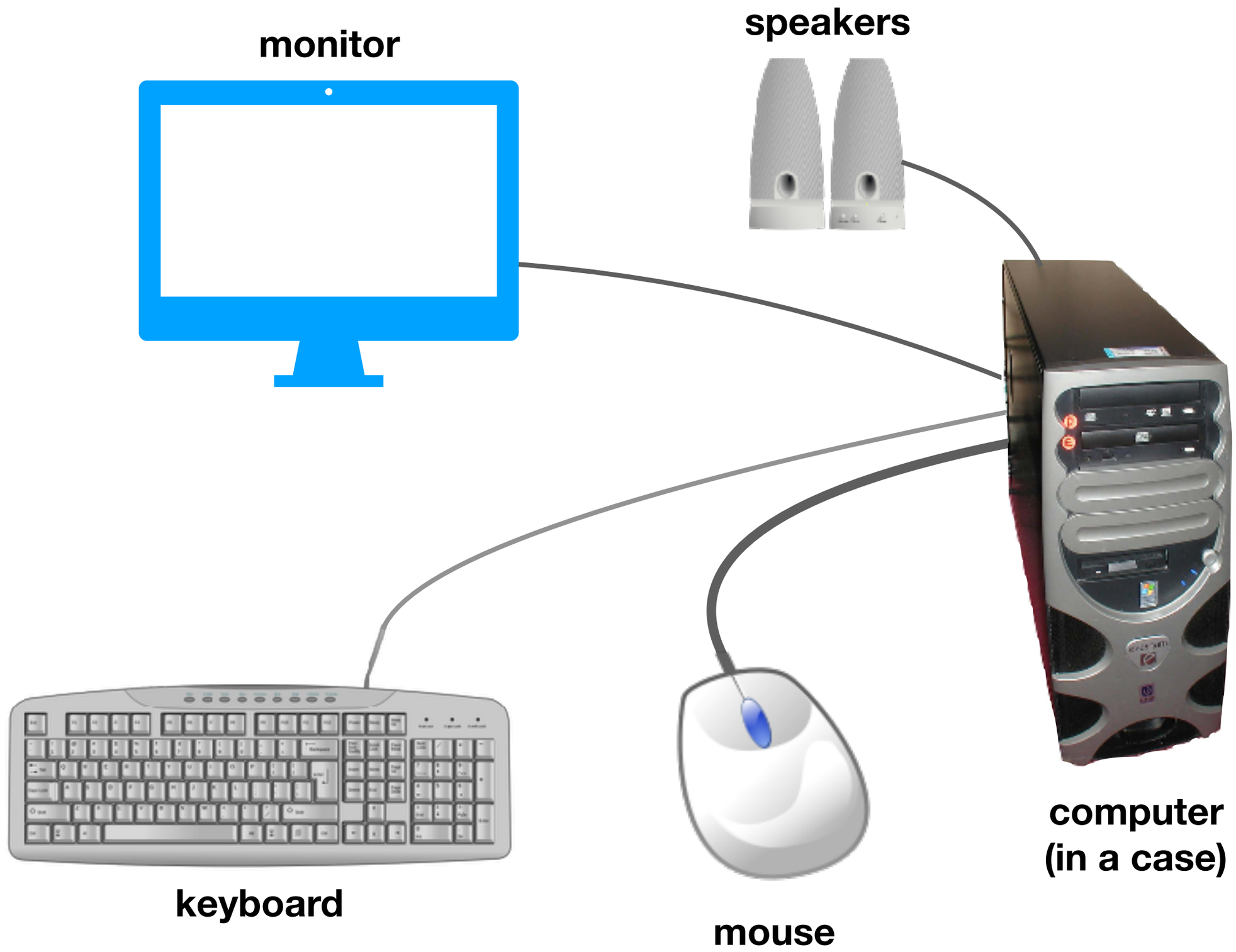
computer (in a case)

Computer Input/Output



Computer Input/Output

I/O devices attach via “ports” (e.g. USB) in back of computer



Computer Input/Output

monitor

speakers

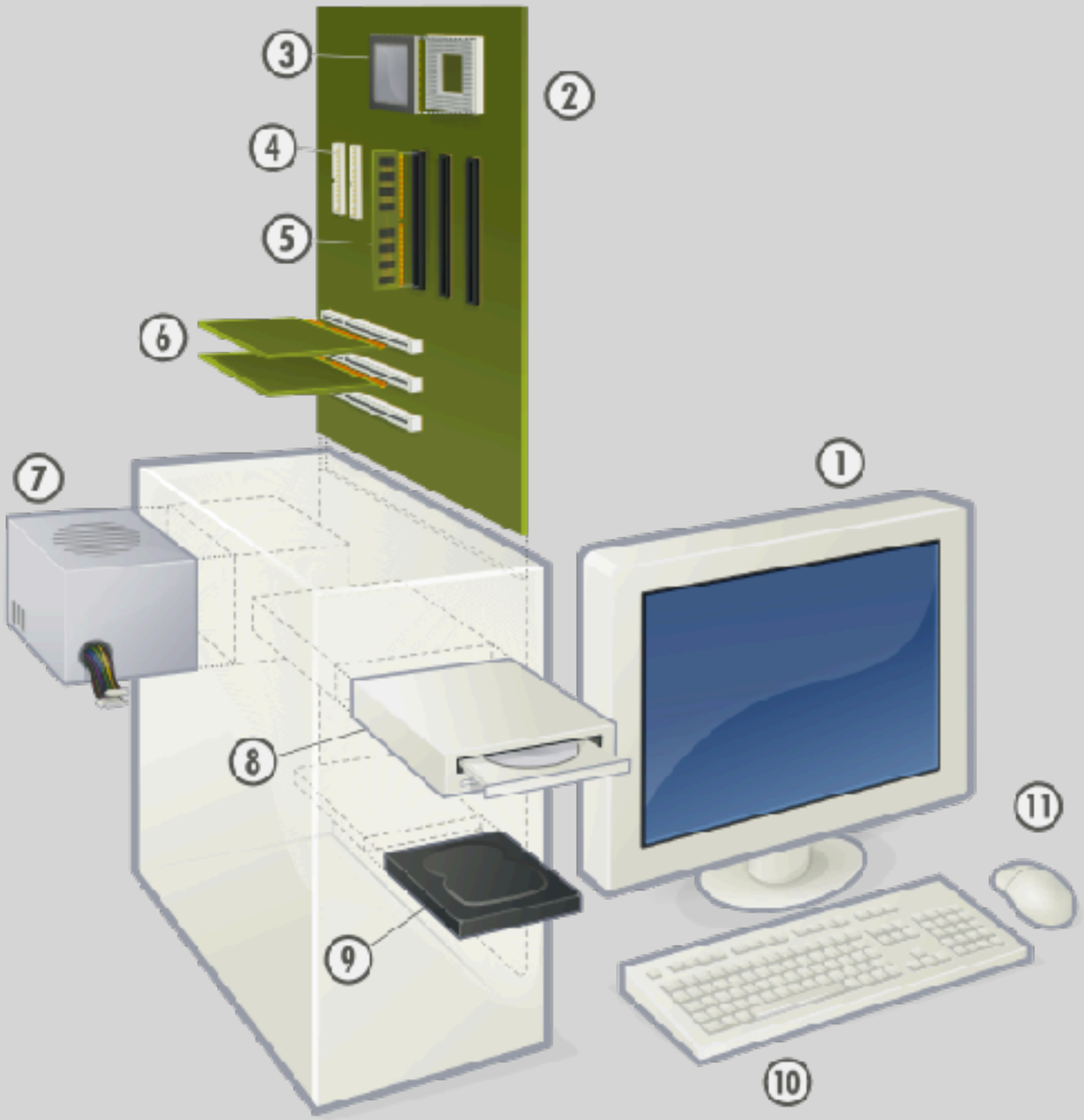
What's inside a computer case?

keyboard

mouse

computer (in a case)

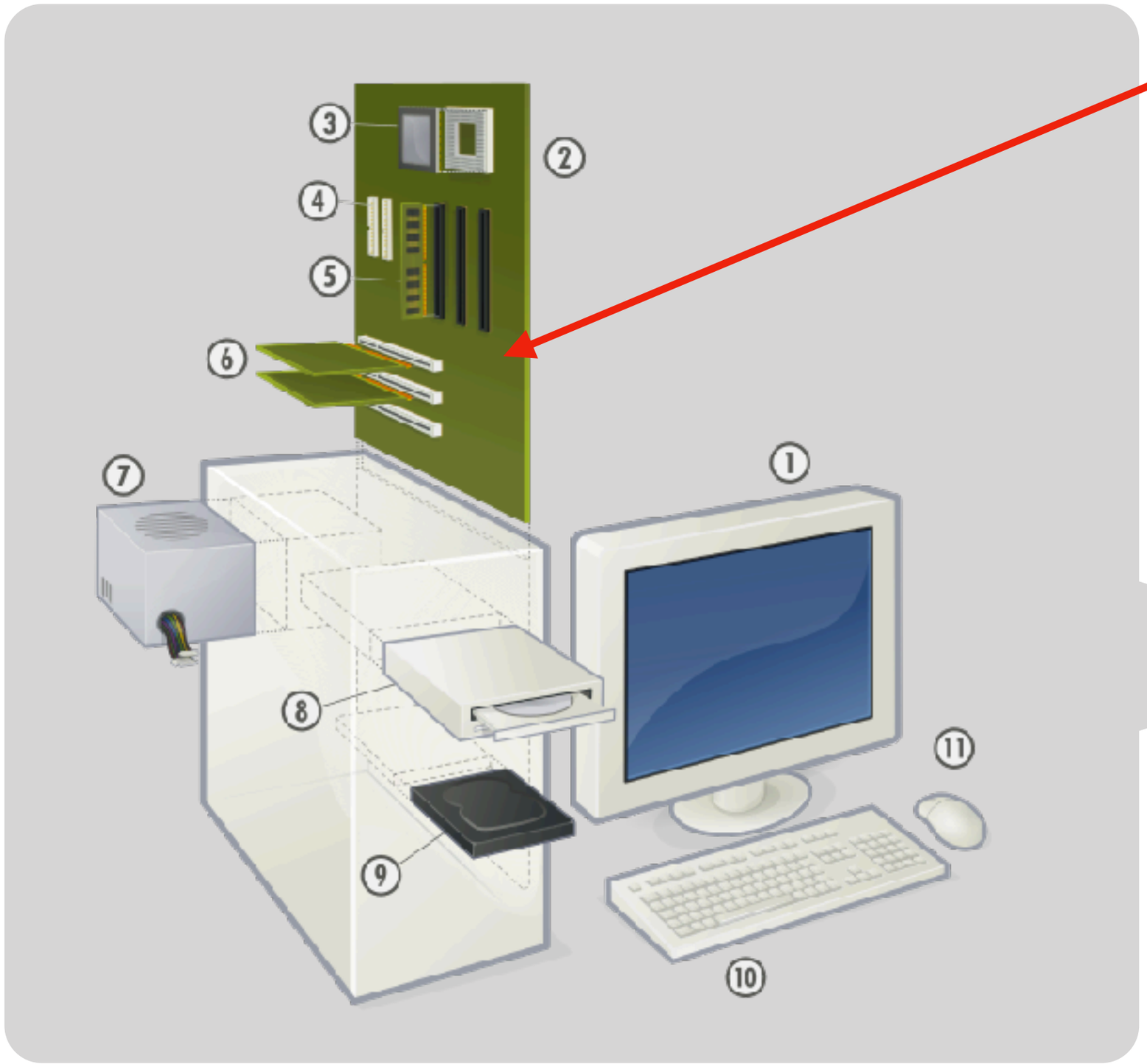
Computer Internals



https://commons.wikimedia.org/wiki/File:Personal_computer,_exploded_4.svg

Computer Internals

Motherboard: main circuit board to which other components connect, via sockets/slots



https://commons.wikimedia.org/wiki/File:Personal_computer,_exploded_4.svg

Today's Topics

Introductions

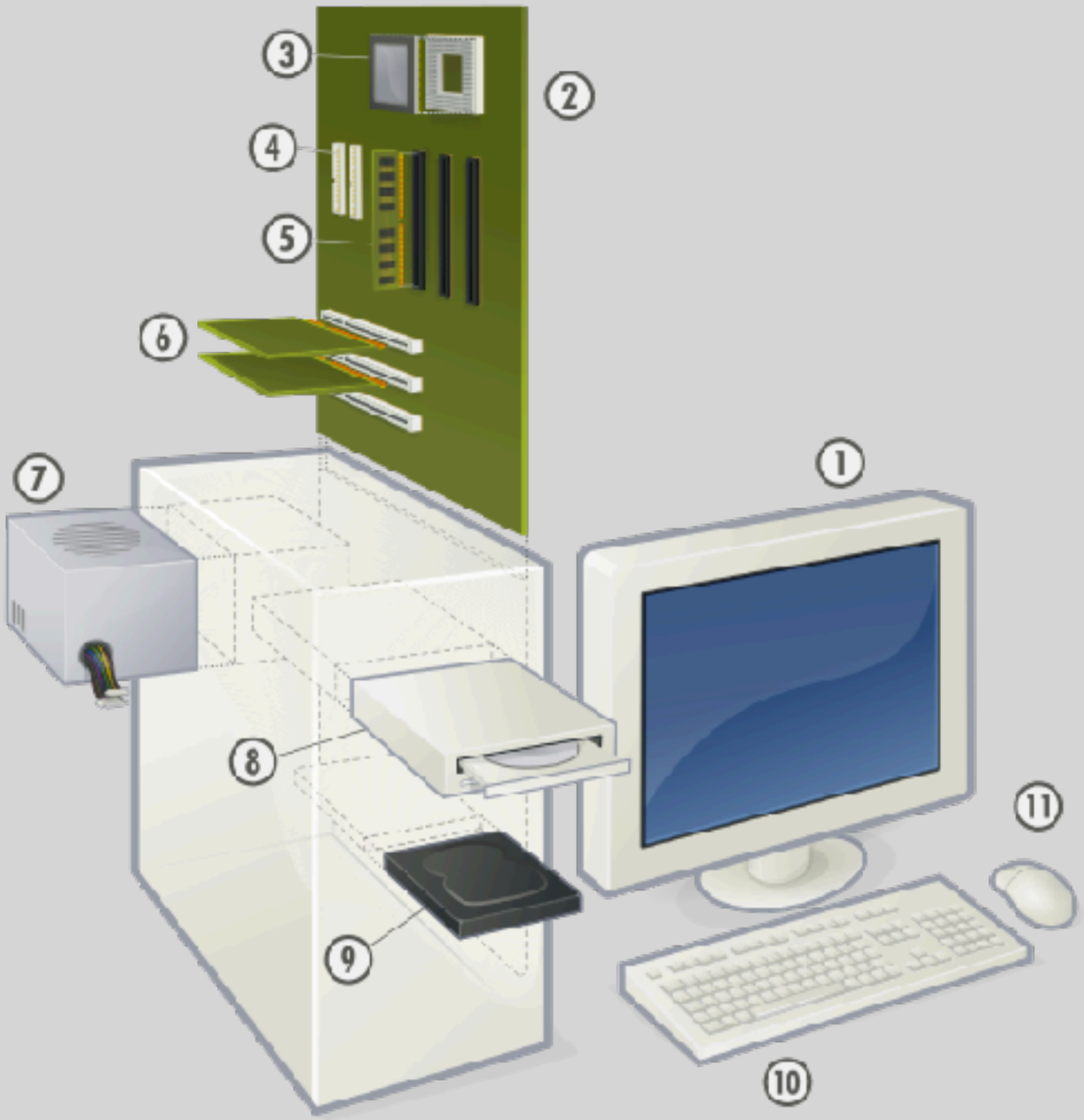
Website

Course overview

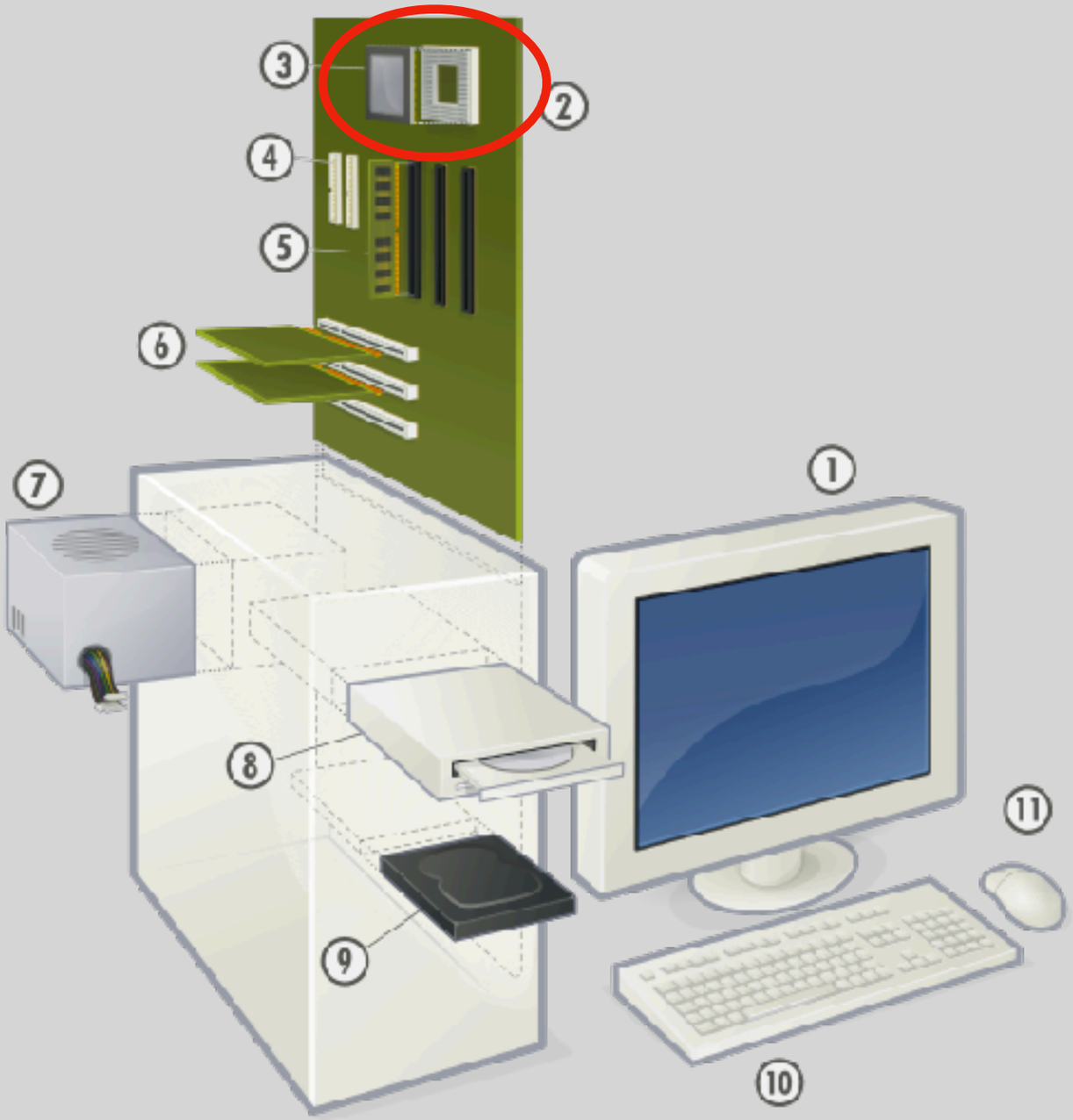
Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Central Processing Unit (CPU)



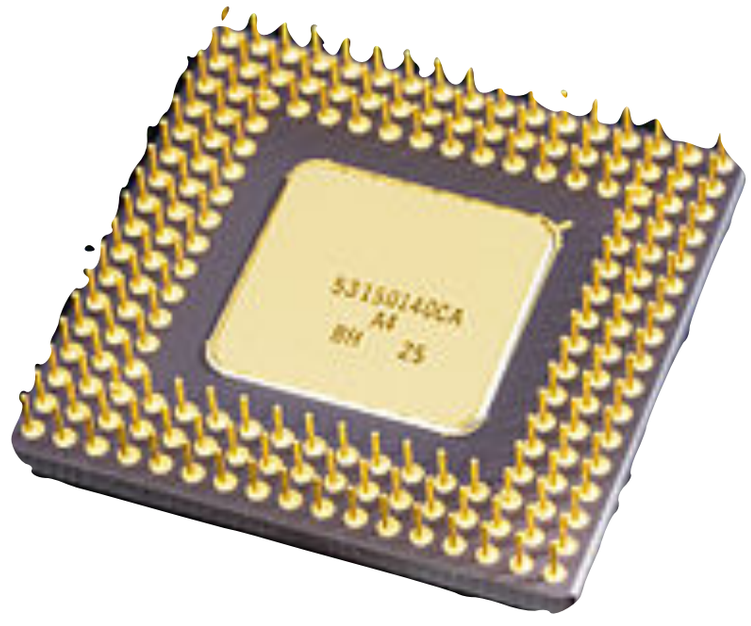
Central Processing Unit (CPU)



CPU

Responsible for computation

- Runs code
- Performs addition, other math
- Compares numbers, text
- Receives input, sends output
- Some compare it to a “brain”



Runs on a clock

- Typically a couple GHz (i.e., billions of ticks per second)
- High-speed makes CPUs hot, require fans/cooling
- Overclocking: running on faster clock than CPU is designed for

Computers often have multiple CPUs

- Motherboard may have multiple sockets
- Single chip may contain multiple CPUs
- Allows computers to do more things simultaneously

Today's Topics

Introductions

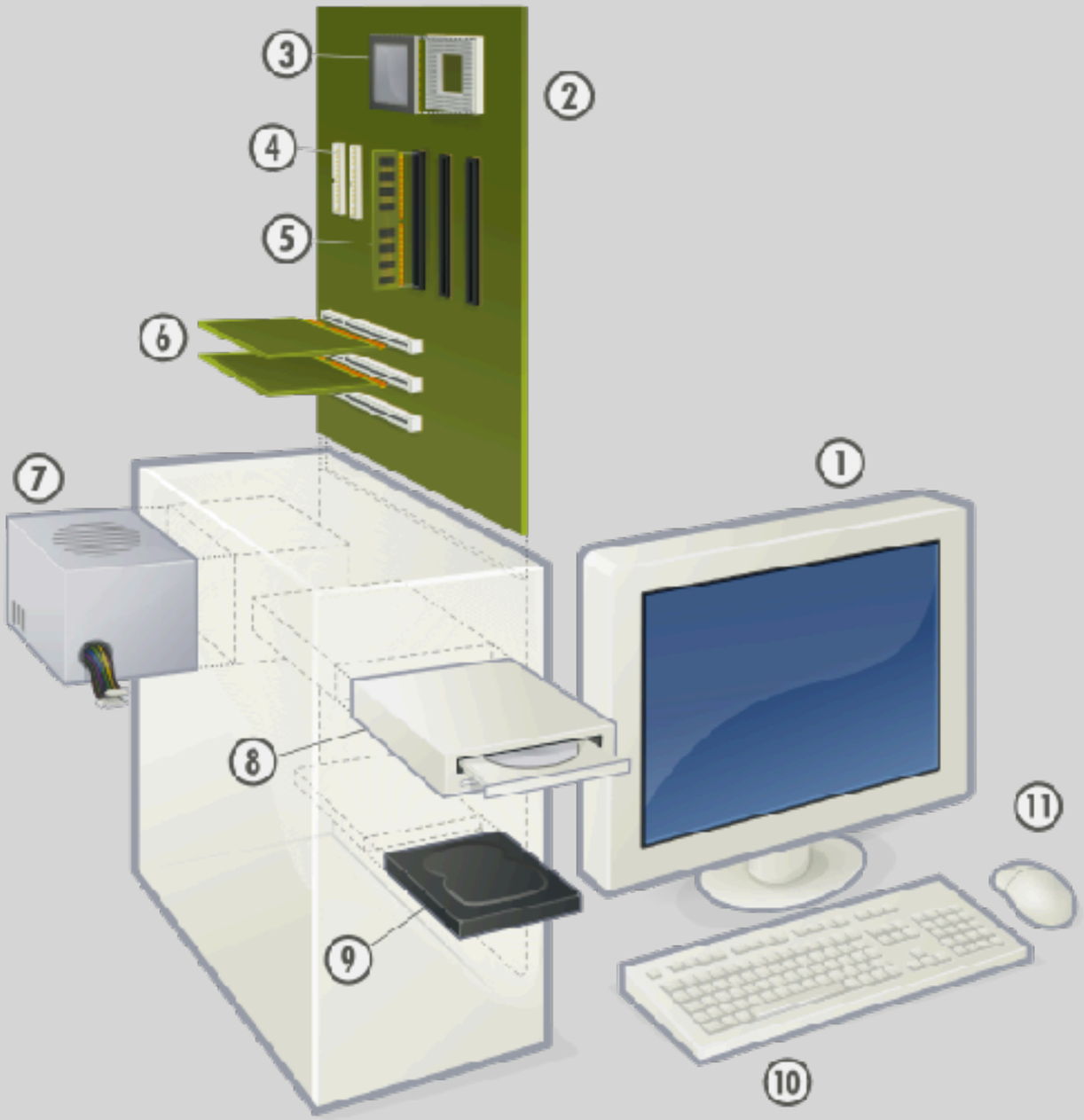
Website

Course overview

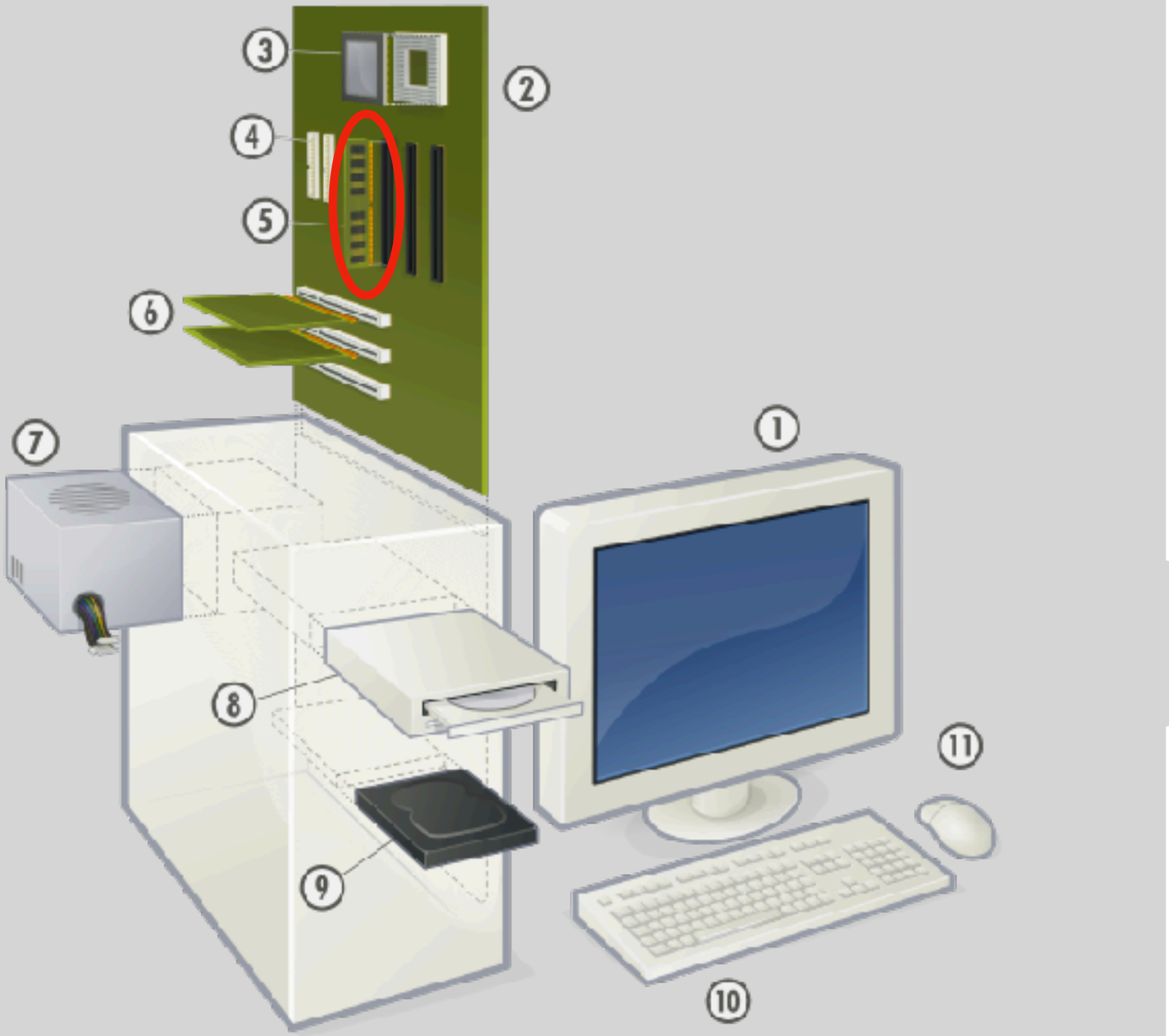
Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Random Access Memory (RAM)



Random Access Memory (RAM)



Memory

Memory stores data for short term

- RAM is most common form today (don't worry about specifics)
- CPU sends data to/from memory
- Accessing it is very fast
- It is "volatile" — meaning you lose this data when you power off your computer
- You don't save "files" in memory, otherwise they would be gone!

Stores bytes of data

- We'll talk about bytes more later; for now, one byte ≈ one letter
- The text "hello" requires 5 bytes
- Typical personal computer has few to tens of gigabytes (billion bytes) of memory

Today's Topics

Introductions

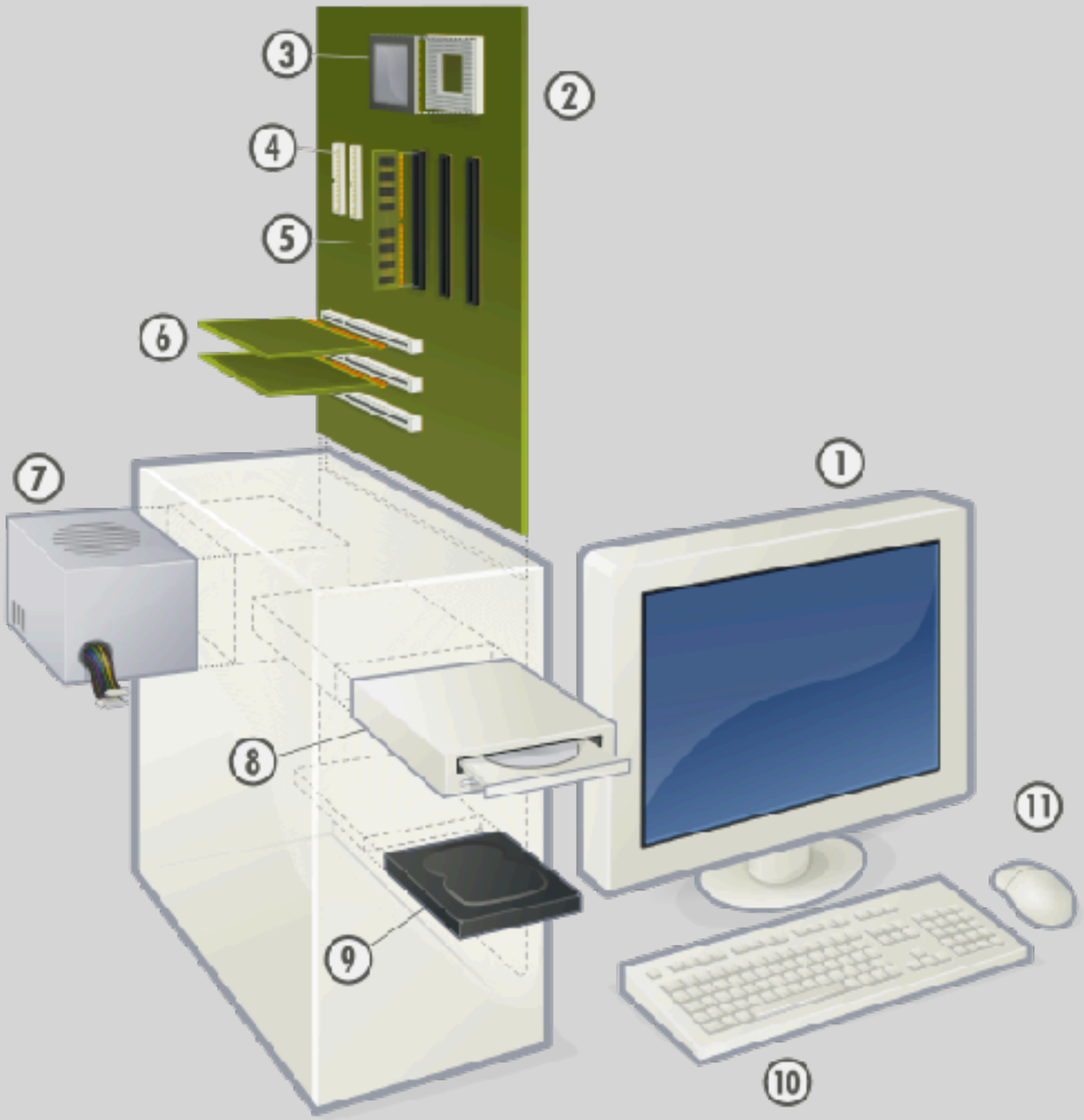
Website

Course overview

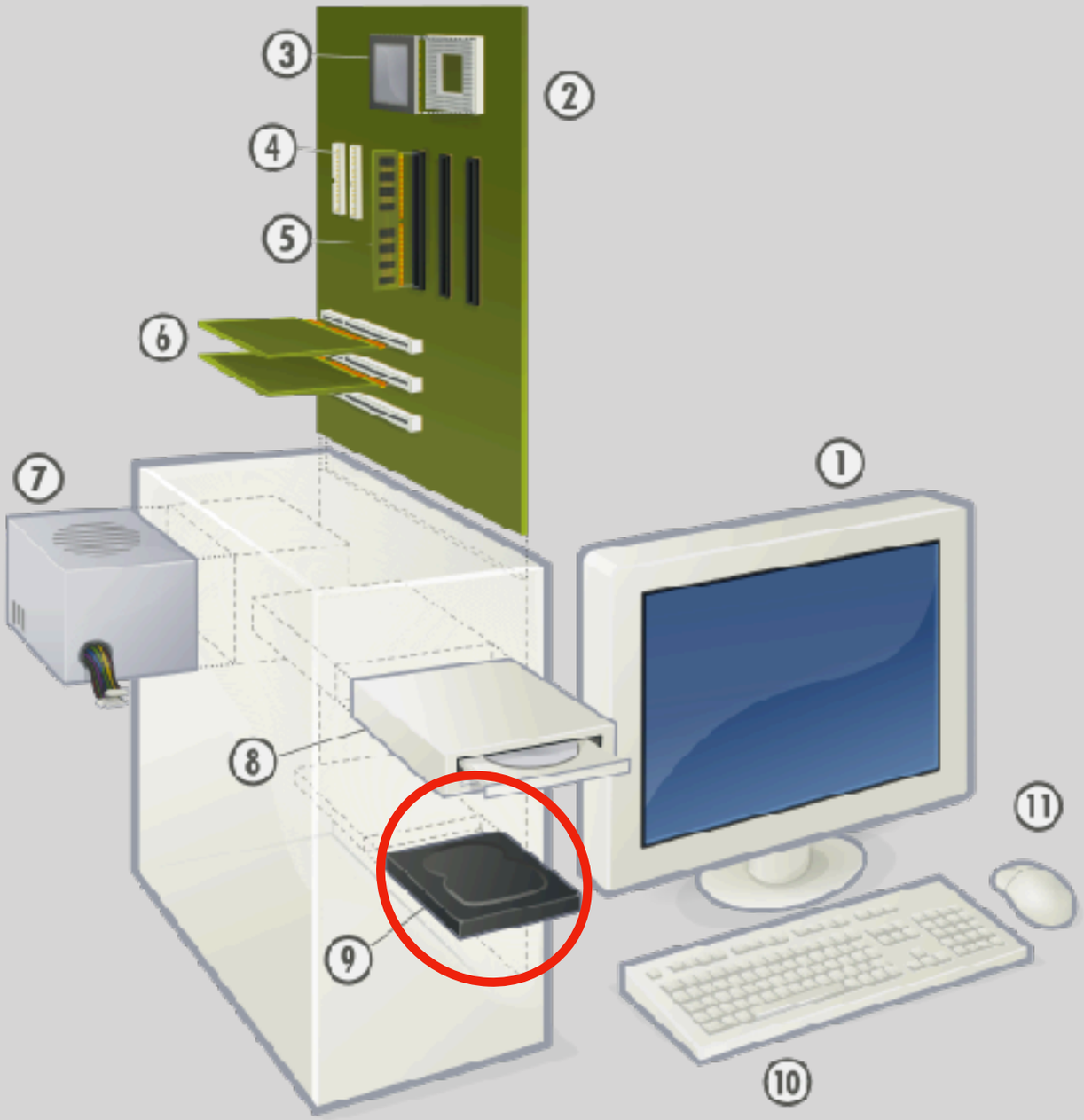
Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Storage Drives



Storage Drives



Storage Drives

Two common devices

- HDD (hard disk drive), has moving parts, cheap, slow
- SSD (solid state drive), no moving parts, expensive, fast
- RAM is much faster than either HDDs or SSDs

Storage devices used to save data after power down

- Persistant medium, in contrast to volatile RAM
- Typical capacity: hundreds of gigabytes

When you make a directory/folder or save a file, that data is ultimately getting recorded to your storage device

- Sometimes computers save to RAM first, and only to the device later; power down cleanly to avoid losing your data!!!

Today's Topics

Introductions

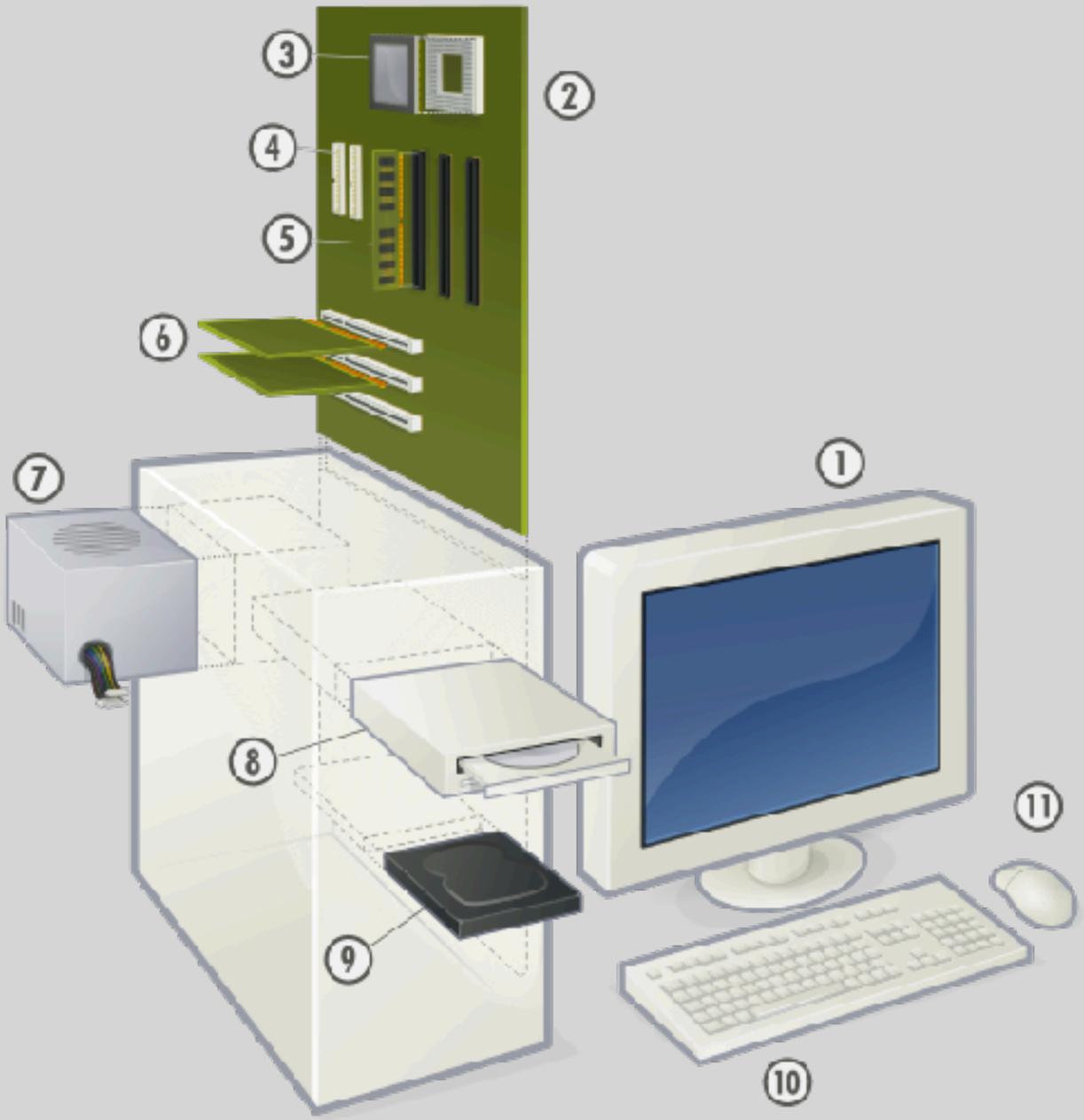
Website

Course overview

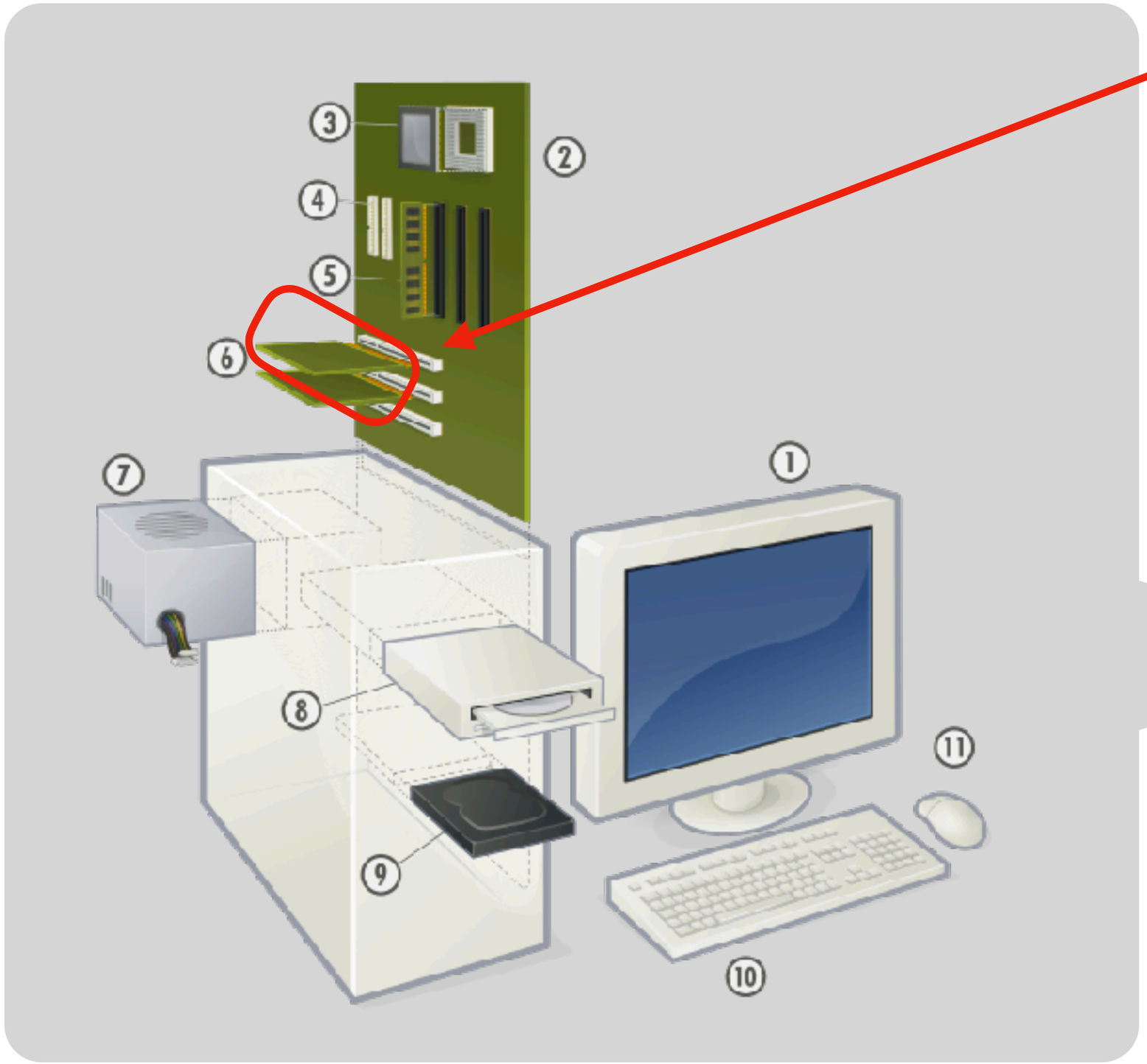
Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Network Interfaces



Network Interfaces



Network: often based on extension card or built into the motherboard itself

Networking

NIC (Network Interface Controller)

- Provides computer communication to other computers, and the Internet

Wired vs. Wireless

- Wired ethernet is common for cable-based connection
- Wi-Fi is common for radio-based wireless connection

Terminology

- **Server**: program/computer that runs, waiting for for incoming requests, to which it responds
- **Client**: program/compute that sends requests to a server

Today's Topics

Introductions

Website

Course overview

Computer basics

- Input/Output
- CPU
- Memory
- Storage
- Networking
- Important software

Important Software

Operating System (OS)

- Controls access to hardware
- Makes hardware easier to use
- Provides ability to run multiple programs
- Examples: Linux, Mac, Windows

Editor

- Allows you to type and save code
- May help you run code and add colorization
- Examples: VS Code, Notepad++, emacs, vim

Browser

- Client program that helps you load/view webpages
- Examples: Edge, Chrome, Firebox, Safari

Conclusion

Today we covered

- Course resources
- Policy
- Computer hardware basics

Action steps for you:

- Familiarize yourself with the syllabus and rest of the website https://tyler.caraza-harter.com/cs301/fall18/home.html
- Sign up for Piazza
- Add exams to your personal calendar now, and notify me of conflicts as soon as possible
- Start meeting your fellow students and thinking about project partners