Objectives

At the end of this section, you will be able to:

• Describe the two stories behind CPSC 301
• List some examples of and reasons why scientific research would be impossible without computers
• Identify the components of an algorithm
• List the steps of a simple algorithm for a simple task
• Describe a few of the top learning goals of CPSC 301
• Locate the key information sources for CPSC 301
Admin

• Beginning-of-term survey
  – [https://survey.ubc.ca/s/cpsc301_presurvey_2016w2/](https://survey.ubc.ca/s/cpsc301_presurvey_2016w2/)
  – closes Jan 17 @ 11:59pm
    • worth a 0.5% bonus to course grade
    • second survey at end of term worth save

• Waitlisted students
  – Only students who attend lectures and sign in will be moved off waitlist
  – Come sign in at front at the end of lecture
The Scientific Method

• Rough steps
  1. Ask a question
  2. Background research
  3. Construct hypothesis
  4. Design an experiment
  5. Perform the experiment and collect data
  6. Analyze results and compare with hypothesis
  7. Report results
  8. Depending on the outcome, repeat from 3 or 4

• Which steps can computers help with?
  – 2, 5, 6 & 7
  – This course will focus on how they help with 5 & 6
First Story of CPSC 301: Data, Data, Data

• Data is one part of the problem solving mechanism
• Using computers we can work with scientific data at
  – larger scales,
  – higher speeds
  – lower costs
  that would otherwise be impossible

• Computers allow scientists to
  collect, store, sort, search, summarize, extract, aggregate,
  analyze, transmit, ...
  large amounts of data, faster and cheaper.
Larger Scales: The Large Hadron Collider

• The LHC is the world’s largest and highest energy particle accelerator
  – A 27 km diameter ring below French-Swiss border near Geneva
• LHC Computing Grid is designed to distribute data from the experiments
  – Detectors create data at TB/sec; immediately filtered to GB/sec of “interesting events”
  – Experiments generated ~25 PB/year during the 3 year run that ended in Feb 2013, at which point the CERN data center had over 100 PB of data
  – Subset of data is stored at 12+ sites around the world

http://home.web.cern.ch/about/accelerators/large-hadron-collider
Higher Speeds: Top 500 Supercomputers

- List of the “fastest” computers in the world
  - Measurement is the Linpack benchmark: solve a large dense linear matrix equation

- Nov 2016 fastest:
  - 93.0 Petaflop/s
  - using 3,120,000 processor cores
  - A Samsung Galaxy S4 smartphone achieves around 3-4 Gflops/s
  - A typical laptop achieves around 50 Gflops

http://www.top500.org/
Lower Costs: Ecological Monitoring with Sensor Networks

• Cheap, low power microprocessors, sensors and wireless networking can collect data from remote locations without human intervention
  – PermaSense project in the Swiss Alps has instrumented a number of peaks to study climate change and rock fall in high altitude permafrost regions
  – Also demonstration projects for monitoring forests, herd animals, water, industry, and military

http://www.permasense.ch/
Larger Scales, Higher Speeds & Lower Costs

• Early gene sequencing
  – First complete genome sequence in 1976; second in 1977; first bacteria in 1995; first animal 1998; first plant 2000

• Human genome sequencing
  – Human Genome Project: 1990 to 2003 for $3 billion
  – 1000 Genomes project: 2008 - 2012 aimed to sequence 1092 genetically diverse genomes with cost under $100 000 each
Larger Scales, Higher Speeds & Lower Costs

• Those numbers were from 2009...
  – Typical cost for full human genome today is $3000 - $4000

http://www.genome.gov/sequencingcosts/
Small Scales, Slow Speeds and Moderate Costs

• Every scientist routinely encounters small data files
  – We want to extract some information
  – We need a slightly different format
• Too often these small tasks are performed by hand
  – Spot the relevant lines by scanning the file and use a calculator
  – Cut & paste the data into a new file in the correct format
• Why not let the computer do the work?
  – Less prone to one-off errors
  – Creates a record of what was actually done
  – If you do the same job twice or on more data, it will be faster
  – Is cut & paste really the best use of your time?

• This case is why CPSC 301 will be useful to most of you
Problem solving requires a process

- Computers require a description of the process at a level of detail that leaves little room for different interpretations.

An algorithm is an effective method for solving a problem

- A sequence of steps which can be mechanized and which (if followed rigorously) will give a correct answer in a finite number of steps.

An algorithm consists of

- Input data (may be none)
- A finite sequence of well-defined operations (steps), and
- Final output (results).

Algorithms (sometimes hidden) are continually affecting your life

- Credit scores, Google rankings, Amazon recommendations, course grades, car engine controllers, cell phones, etc.
Example of an Algorithm

• Instruction for a proper Hand Washing:
  – https://www.youtube.com/watch?v=EwjDShmfFHM
Incidence of dementia expected to more than double in Canada in next 30 years

Patients with moderate to severe dementia require help with basic activities of daily living

Computerized handwashing assistant
  – Developed by Mihailidis & collaborators at U. Toronto
  – Overhead video camera tracks patient’s hands and towel
  – Video screen and speaker provide various levels of prompting when there appears to be no progress

Handwashing Assistant Video
CPSC 301: Computing in the Life Sciences

• We will consider two challenges:
  – How to represent data in the computer
  – How to describe to the computer the process that should be
    applied to the data

• We will design our data representation and process
  descriptions in three to four frameworks:
  – Scratch
  – Python
  – Microsoft Excel
  – BioPython

• We will draw upon examples from the life sciences to
  illustrate the approaches
Course Level Learning Goals

• Students who complete this course will be able to:
  – Create, identify, view and modify common data storage formats using common applications.
  – Choose an appropriate data format for a specified task, explain why that format was chosen over alternatives, and identify potential shortcomings of the chosen format.
  – Write, modify, debug, analyze and execute simple programs to create, collect, transform, transmit, manage, retrieve, analyze or visualize data in simple ways.
  – Develop and debug programs in an integrated development environment (IDE).
  – Find, validate and use well-designed scientific software from the Internet for more complex data and its manipulation.
  – Discuss how the capabilities and limitations of computers and networks might influence their use in a specified scientific task.
Course Structure

• Readings
  – complete readings AND next classes slides before we start that topic in lecture
  – pre-class quizzes (Connect) on there readings
  – slides will be posted ~2 days before first lecture on the topic

• Lectures
  – Less presentation more discussion
  – We’ll use clickers
    Register your clicker on Connect
  – In-class exercises

• Labs
  – Start next week!
  – Lab 2 and on will be done in groups of 2
  – Do the before-lab sections BEFORE

Course grade:
• 4% - pre-class quizzes
• 4% - in-class exercises
• 4% - clicker questions
• 25% - labs
• 20% - midterm exam
• 43% - final exam.

To pass the course, must
• get an overall passing grade in the labs
• pass the final exam.
Course Details: Getting Started

- **Find the course web page:**
  

- **Read through the Syllabus on the first page**

- **Log into Connect and register your i>clicker.**

- **Register into Piazza site for the course.**

- **Do the assigned reading and review the slides for Thursday.**
  - pre-class quizzes will start this week (due Thursday before class).

- **Download the lab: it starts next week.**
Modern science thrives on data

The material in CPSC 301 is chosen based on two themes:
  – Computers allow us to work with data at scales, speeds and costs which would otherwise be impossible
  – Computer science is the study of automating processes

CPSC 301 has
  – lectures (use clickers and group exercises)
  – labs
  – a midterm exam
  – a final exam

CPSC 301 uses three websites
  – A main course site with most of course materials
  – Piazza for discussions
  – Connect for grades
Change log

V2
- slide 19 updated to state quizzes start this week for Thursday
- added in the link to the pre survey.