

Biol 365 – Computer applications in biology

Class intro

Computers have become extremely powerful

- The computer used to operate the Apollo 11 moon landing (1969)
 - Used about 5,000 circuits (number of calculations per cycle possible)
 - Had a clock speed of 1.024 MHz (number of times the 5,000 operations can be done per second = 1,024,000)
- It had 4 kb of RAM (working memory)
- How does this compare to our modern machines?



Onboard computer from a Ford Ranger



- Runs at 40 MHz
- But, many more transistors on the CPU chip (can do many more calculations per cycle)
- Several MB memory
- Way more powerful than the Apollo moon landing computers

General-purpose, desktop computing has also advanced

- Desktop computing was just for hobbyists initially
- Hardware was not up to tasks like graphics, audio
- Not multi-tasking – could only run one program at a time
- Coincidentally, my own computer history parallels the development of desktop computers pretty well

My computing history

- First (family) computer – Apple IIe, late 1970's
- No hard drive, monochrome (green) display
- Two floppy drives recommended – one for programs, one for data
- Floppies held 512 KB
- No graphical interface
- 1.023 MHz clock speed
- 64 KB RAM
- Usable – word processing



College days – terminals, not home PC's



- PC's were very expensive, few students had their own in the mid/late 80's
- When we used computers, we used terminals like this one
- Monochrome display (white, green or amber)
- Printed to a central print facility – submit a print job, pick it up later
- Used for “real” work – PC's were still considered too limited for serious scientific computing

Mac SE

- Acquired in 1989 – senior in college
- CPU ran at 7.8 MHz
- 40 MB hard drive (!)
- 1.2 MB floppy drive
- 1 MB RAM
- Monochrome display (black, white, and shades of gray)
- Graphical, mouse-driven interface
- Multi-tasking – could have more than one program running at a time
- Access campus network (no Internet yet) through an external modem over phone lines (remember those?)



The mighty Gateway 486DX



- Upper-end PC in 1990
- 33 MHz CPU
- 250 MB hard drive
- 4-64 MB RAM
- MS-DOS or Windows 3.1
- 16 or 256 colors on display
- No network connection (floppy disks, tape drives, or 100 MB cartridge drives for data exchange)
- Not my own – used to do Geographic Information Systems analysis for my MS thesis
 - One operation (overlaying locations of bald eagle sightings with maps of vegetation) would take about 4 hours – lightning fast in 1990!

Mac IIci

- Acquired in early 1990's
- 25 MHz
- Separate FPU for fast math with decimal numbers
- 1-68 MB RAM
- 40 MB hard drive
- True color (like today, millions of colors but less screen resolution)
- Dial-up modem built in (AOL!)



Power Mac G3

- Acquired in mid-late 1990's
- 300 MHz
- 256 MB – 1 GB RAM
- Hard drives up to 27 GB
- Graphics card with 16 MB RAM
- Ethernet built in, but most people still used modems from home
- Started my PhD dissertation on this computer



Compaq Presario

- Acquired in 2001
- 1.33 GHz
- 256 MB RAM
- 64 MB video card
- 80 GB hard drive
- Ethernet built in, broadband internet was becoming common, but no WiFi yet
- Finished my PhD dissertation on this one



Dell Inspiron 531

- Acquired in 2006
- 2 GHz dual-core processor
(more than 1 CPU!)
- 1-8 GB RAM
- 160 GB hard drive



My current computer – Dell Optiplex 9010

- Standard faculty computer
- Octo core processor, 3.4 GHz
- 16 GB RAM
- 500 GB solid state drive



- GIS analysis that took 4+ hours on the Gateway in 1991 would take less than a few seconds on this machine

My smart phone

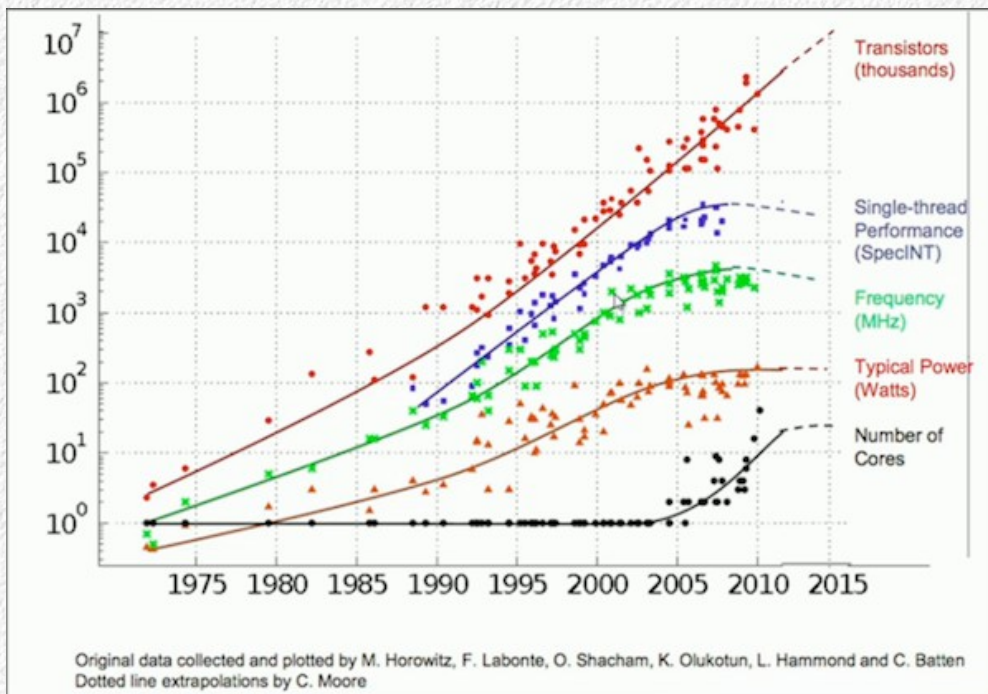
- 2.2 GHz octo-core CPU
- Millions of transistors
- 32 GB internal storage
- 6 GB RAM



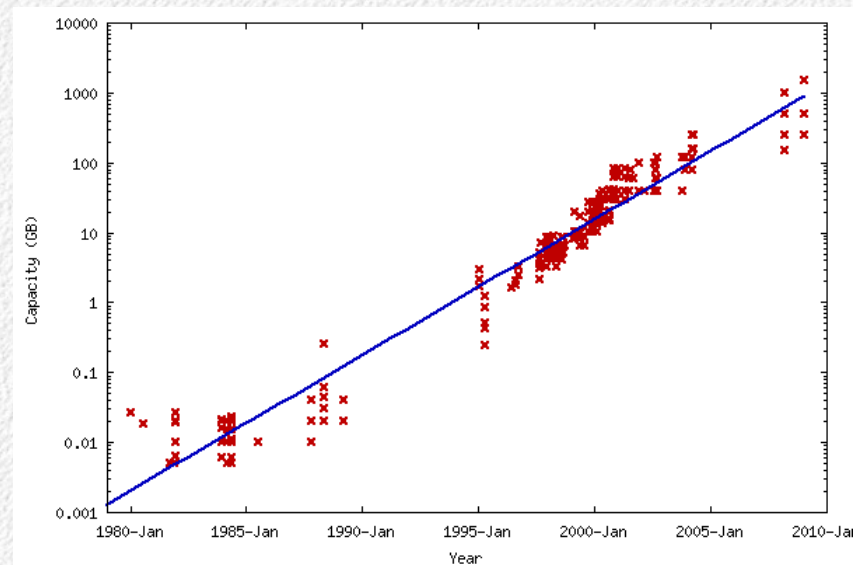
- Nowhere near the best available today, but still more than enough to go to the moon!

Increase in computing power over time

CPU



*Hard drive
capacity*



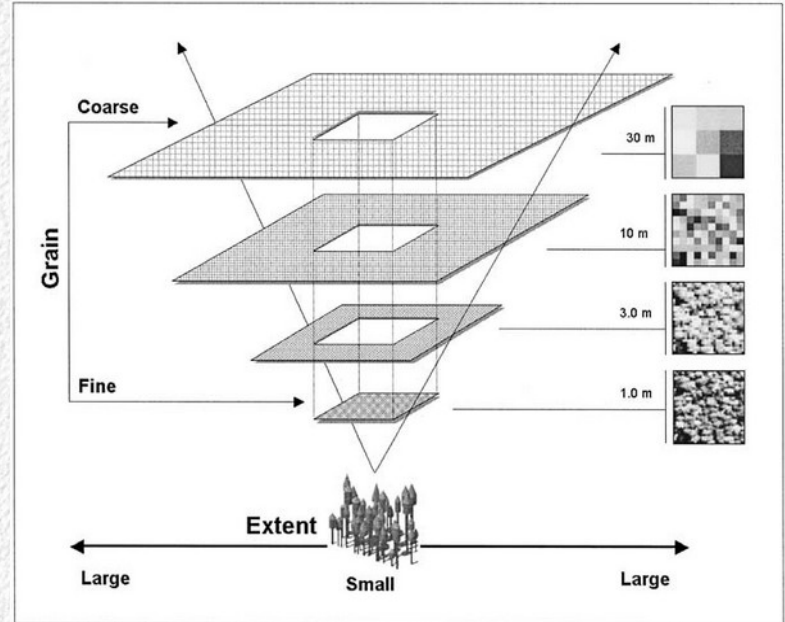
At the same time, cost has declined (both in absolute \$ and adjusted for inflation)

Quantitative increases in computing power qualitatively changed Biology

- The increase in computing power has allowed us to use computers very differently
 - Not just for math, data management, statistical analysis
 - Used now for sound, images, video – big files, very processor intensive
- We are now able to address questions that couldn't be addressed before...
 - Genome-level analysis
 - Environmental change detection at global scales
- ...using methods that weren't practical before
 - Numerical methods of analysis
 - Stochastic simulation modeling

Sizes of data sets can be bigger

- Example: GIS – computer mapping
- For a given level of computing power, trade-off grain and extent
 - Can record fine detail over a small area
 - Can map a large area, but only at a coarser level of detail
- This trade-off is true no matter how powerful the computers are, but...

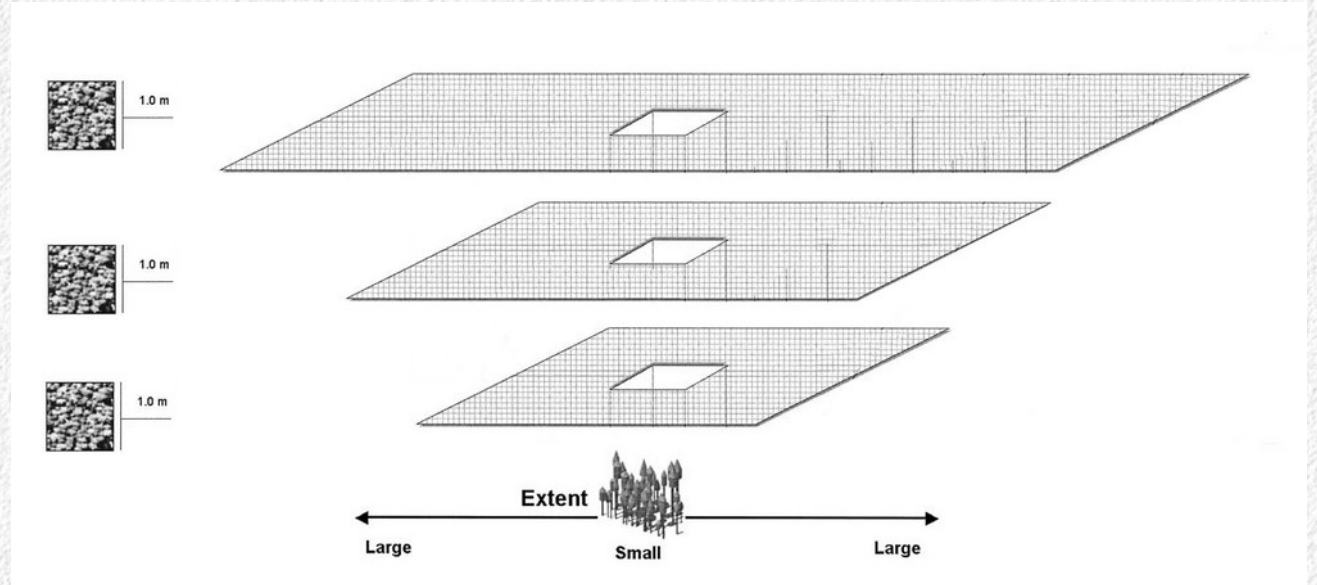


Sizes of data sets can be bigger

More powerful



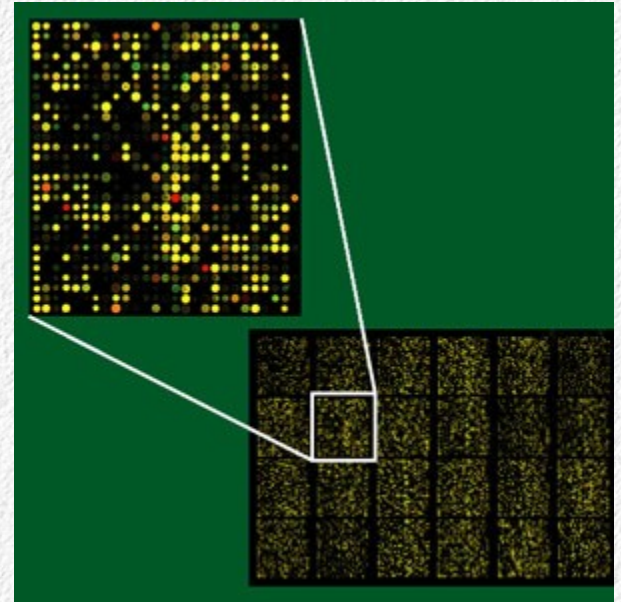
Less powerful



- As computers get more powerful, it's possible to record fine grain detail over larger extents
- Result is better cover type mapping, change detection

Example: genomic analysis

- We can look at expression at huge numbers of genes (10,000 +) simultaneously
- Conducting the experiments, and then collecting, managing, and analyzing these data would not be possible without computers



Example: statistical analysis

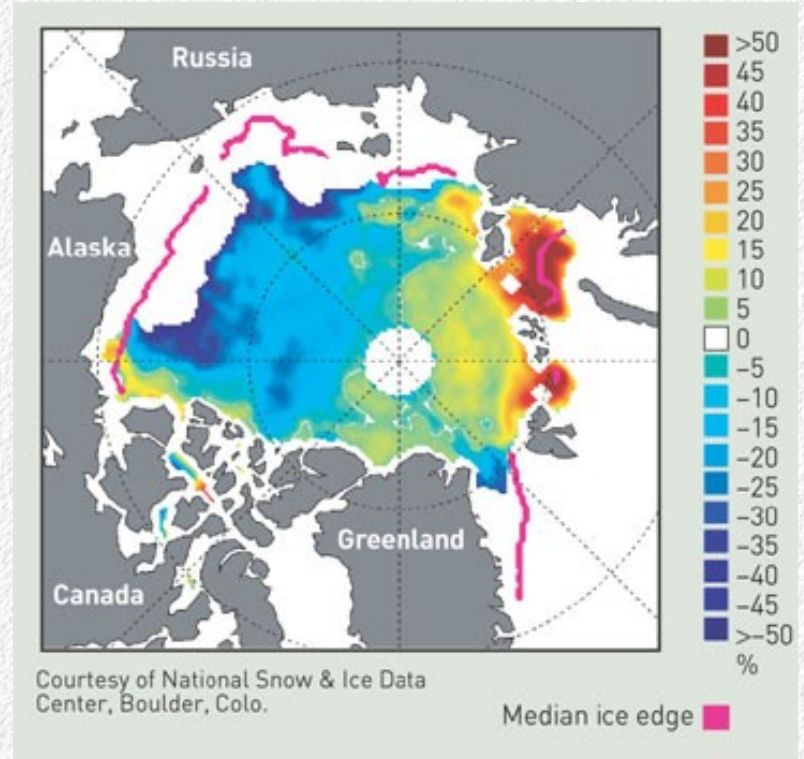
- Complex data sets – we have known how to analyze them for a long time, but couldn't
 - Multivariate data – require calculations that would take years to complete by hand
 - Less than a second on modern computers
- Computer-intensive methods:
 - Randomization tests
 - Neural networks
 - Tree-based methods
 - Maximum likelihood

Computers allow automation of data collection

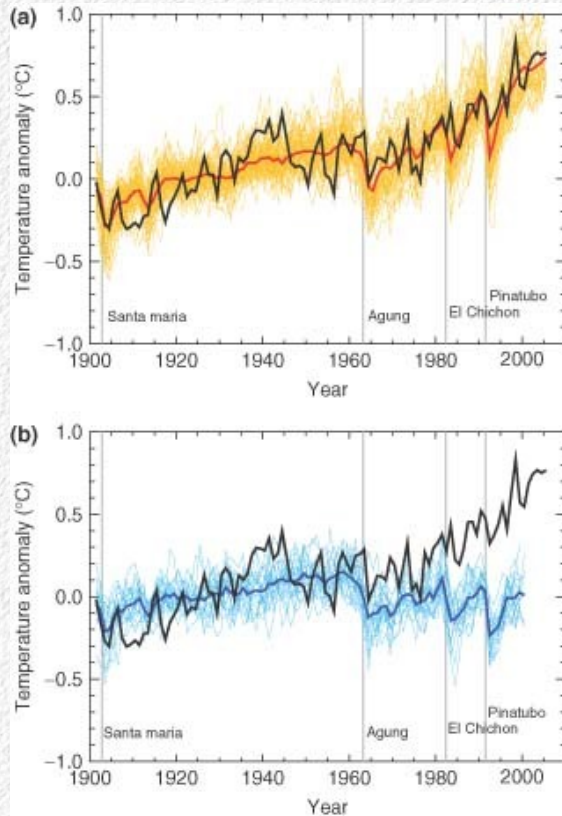
- Digital sensors of various kinds have been developed that can collect vast amounts of data
- Examples:
 - Imagery – microarrays, LandSat, x-rays
 - Position – GPS collars
 - Physiological state – continuous monitoring of heart rate, environmental conditions

Example: change detection

- Improvements in technology (remote sensing) along with computing power allows us to detect changes over large areas over time
- Changes in arctic ice shown here



Example: simulation modeling



- When you can't do experiments, next best thing is to simulate the system in a computer
- Global climate models can be built that:
 - Include all potentially important variables
 - Use high enough spatial resolution to account for geographic variation
- Here we're seeing that we can't reproduce historic temperatures unless we include anthropogenic CO₂ emissions

Numerical methods become practical

- Statistical methods
 - Maximum likelihood
 - Randomization methods (bootstrap, Monte Carlo)
- Numerical optimization
- Sequence alignments

Identifying function of unknown genes

- You sequence a gene, but don't know what it does
- How do you find out?
- Need to check if the function of the gene is already known in your own organism, or another
- Even if it's only known from another species, good chance that it has a similar function in your organism

BLAST searches

- Aligning DNA sequences – need two components
 - A database (or “library”) of sequences and their known functions
 - A routine (or “algorithm”) for aligning an unknown sequence with a known sequence
- BLAST is a common alignment method
 - Basic Local Alignment Search Tool
 - Aligns the unknown sequence to those in the database
 - Gives a measure of sequence similarity
 - Genes that are similar above a user-defined threshold are likely to have similar function
- Any database can be used, but large databases built from contributed sequences from all over the world are the most useful

The BLAST algorithm – matching one sequence to another

- Assign a score to the unmodified sequence
 - Matching bases get a +1
 - Mismatching bases get a penalty, $-\mu$
 - Score = # matches $-\mu$ (# mismatches)
- At the first mismatch try an insertion or deletion, see how the score changes
 - Inserting or deleting a base to match gets a penalty, $-\sigma$
 - Score = # matches $-\mu$ (# mismatches) $-\sigma$ (# indels)
- If the score improves, keep the indel
- Repeat with each mismatch, until the score is as high as possible = the score for 1 match

Example – possible matches between 1 and 2

Seq. 1: AGCTTATAAGCCAA

Seq. 2: AGTTATAAGACCAA

Matches: 8

Mismatches: 6

Seq. 1: AGCTTATAAGCCAA

Seq. 2: AG-TTATAAGACCAA

Matches: 11

Mismatches: 2

Indels: 1

Seq. 1: AGCTTATAAG-CCAA

Seq. 2: AG-TTATAAGACCAA

Matches: 13

Mismatches: 0

Indels: 2

Repeat for all the genes in the database

- The sequence may have thousands of bases
 - Alignment against one sequence in the database takes time
- May be thousands of sequences in the database
 - Multiply time per sequence x thousands
- Pick the gene with the highest score as the match
- Huge amount of computation – could be done by hand, but would take inordinate amounts of time
- Thanks to computers, this is now a routine part of molecular genetics work

Computers are ubiquitous

- As of today, 91% of US households have a computer (desktop, laptop, tablet, smartphone)
- Powerful computers are now found in everyday items (smartphones, cars, appliances)
- It's increasingly difficult to avoid computer use entirely, even for people who do not personally own one
- Kids are growing up using them → no fear (but maybe some loathing)

What's missing

- Students today are more comfortable with using computers, but are not expected to learn as much about them
- Computer programming classes are not required for Bio majors
- Computer literacy courses no longer required at CSUSM
 - Did not cover MS Excel, MS Access for data management and analysis even when it was required

In this class:

- We will learn some of the ways computers are used in Biology
- We will use a spreadsheet as the primary computational tool
 - The “swiss army knife” of scientific computing – general purpose, can do most things
 - Will mention more specialized tools for particular purposes, but won't use them in class