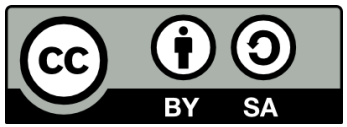
Statistics for Engineers

Overall revision number: 102 (January 2014)

Copyright, sharing, and attribution notice



We appreciate:

- if you let us know about **any errors** in the slides
- **any suggestions to improve the notes**

All of the above can be done by writing to

`kevin.dunn@mcmaster.ca`

or anonymous messages can be sent to Kevin Dunn at

http://learnche.mcmaster.ca/feedback-questions

If reporting errors/updates, please quote the current revision number: 102

Please note that all material is provided "as-is" and no liability will be accepted for your usage of the material.

Plan for today's class

1. Background and administrative issues
2. Course contents

Some background about myself (most of you already know)

- Undergraduate degree from University of Cape Town, 1999
- Masters degree from McMaster, 2002 (not a "doctor", please)
- Worked with a number of companies since then on data analysis and consulting projects
- Worked at GSK on a 1-year contract until June 2012
- Now working full-time at McMaster since July 2012
- Office hours: check my online calendar
- Office is in BSB, room B105
- Arrange a meeting: kevin.dunn@mcmaster.ca
- Cell: (905) 921 5803 and not ~~extension 27337~~

My objective

I hope to make this class **worthwhile** and **practically** applicable to you. Please let me know how I'm doing at any time; there will be anonymous course evaluations throughout the course for your feedback.

Acknowledgments – sources of reference for this course

- Dr. John MacGregor, who taught this course since 1983
  - Happy 31st anniversary 4C3/6C3
- McMaster Advanced Control Consortium (MACC)
- The many companies I've worked with over the past 11 years
  - their problems and data appear in the course, often in disguised form

Administrative issues

- TA introduction
- Video and audio
- Website
- References
- Software
- Expectations
- Grading

Let's meet your teaching assistants

Yanan Cao

- caoy4@mcmaster.ca
- JHE, room 369
- extension 24031
- Currently doing her Ph.D with Chris Swartz

Shailesh Patel (back on the 17th)

- patelsr@mcmaster.ca
- JHE, room 369
- extension 24031
- Currently doing his Ph.D with Chris Swartz

Office hours for both TAs are by email appointment

The course website is where everything important is announced and posted

http://learnche.mcmaster.ca/4C3

- *Not an Avenue website!*
- Slides will be added to the site before class
- Please print slides and bring to class
- Assignments and solutions will be posted there
- Data sets and many resources you will require are posted there

Website is the main reference for all things course-related

- expected to check it about 3 times per week (top left)
- and follow on Twitter to get updates: @stats4eng

The course website

Statistics for Engineers: CHE 4C3/6C3

Administrative

- Course outline
- Official course textbook
- Supplementary readings
- A self-paced software tutorial
- Course videos and audio from 2010, 2011, 2012, 2013 and 2014
- Copyright and other legal stuff
- Comments, questions, feedback? Don't wait for official course evaluations.

Announcements (previous ones) or

- 30 Dec: the course outline for 2014 has been posted.
- 30 Dec: consider following the Twitter account for this course: `@stats4eng` so you always stay up to date.
- 04 Jan: the first class is on 06 January, at 16:30 in MDCL 1110. See you there!

Class materials

Notes for 2014 class

Topic	Latest slides
0. Course outline and overview	PDF
1. Visualizing process data	
2. Univariate data analysis	
3. Least squares modelling	
4. Design and analysis of experiments	
5. Process monitoring	
6. Latent variable methods and applications	
7. Course wrap up	

Suggested readings

Assignments, projects, exams

Assignments:

1. Assignment 1 - 2014
2. Assignment 2 - 2014
3. Assignment 3 - 2014
4. Assignment 4 - 2014
5. Assignment 5 - 2014
6. Assignment 6 - 2014
7. Assignment 7 - 2014

How to submit an assignment electronically

Tests, exams and project

- Designed experiments project
- Response surface bonus project
- Final exam
- Practice questions for tests and exams

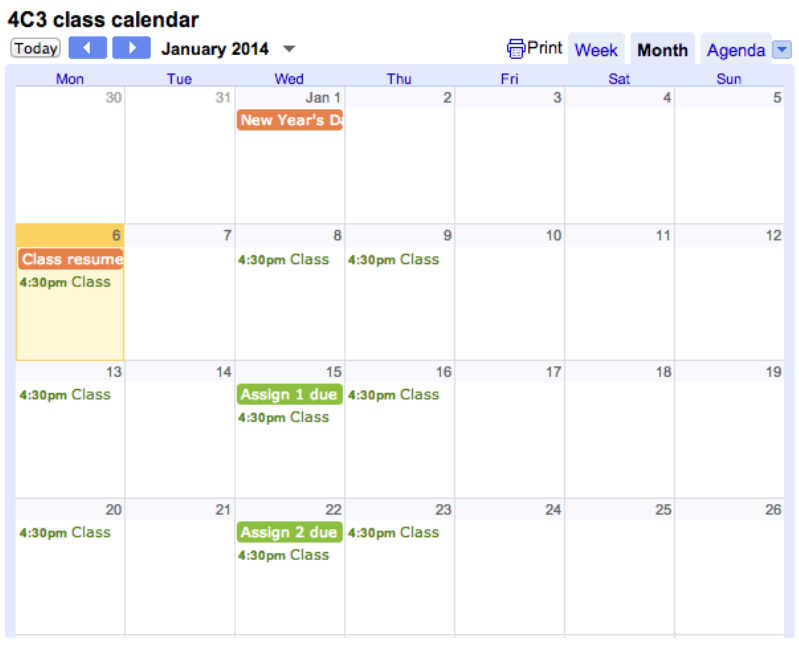
Tables

- Tables of the normal and *t*-distribution
- DOE trade-off table

Datasets

All course datasets

Course calendar



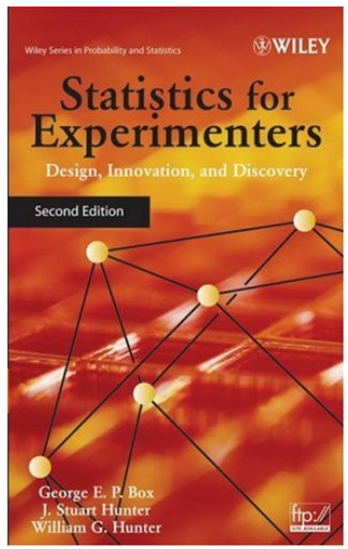
Video and audio recordings will be available

- Video/audio recordings from 2010, 2011, 2012 and 2013 on website
- Purpose: for your review, and to prepare for assignments and exams
- Might be useful if you miss a class
- As long as feasible, I will try to video record all classes in 2014
  - Try to record just myself, the board and the projector
  - Can't guarantee the quality will be very good (background noise, etc)
  - Video should be available within 24 to 48 hours after the class
- Audio recordings will also be made available, when possible

Course text book (required) and recommended references

- All you require are the slides and ...
  - **Process Improvement using Data**
  - Draft book; it gets updated every week while I'm teaching 4C3/6C3
  - **http://learnche.mcmaster.ca/pid**
  - Use website to report errors; suggest improvements for slides and book

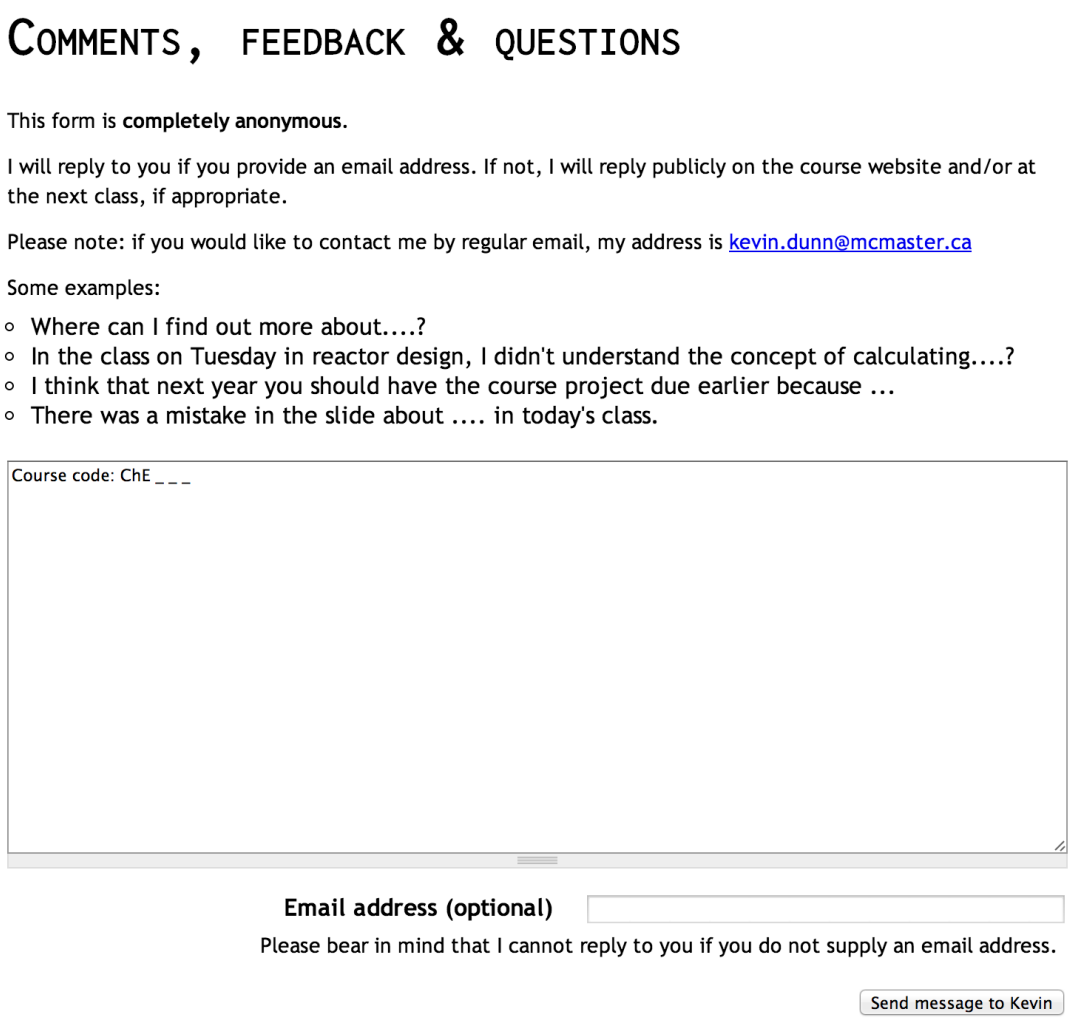
- Some suggested books on the course website:



- *Recommended*: Box, Hunter and Hunter, "Statistics for Experimenters: Design, Innovation, and Discovery", 2nd edition
- Other references on course website (self-directed learning)

Please keep submitting course feedback via the website

http://learnche.mcmaster.ca/feedback-questions



Comments, feedback & questions

This form is **completely anonymous.**

I will reply to you if you provide an email address. If not, I will reply publicly on the course website and/or at the next class, if appropriate.

Please note: if you would like to contact me by regular email, my address is kevin.dunn@mcmaster.ca

Some examples:

- Where can I find out more about....?
- In the class on Tuesday in reactor design, I didn't understand the concept of calculating....?
- I think that next year you should have the course project due earlier because ...
- There was a mistake in the slide about .... in today's class.

Course code: ChE _ _ _

Email address (optional)

Please bear in mind that I cannot reply to you if you do not supply an email address.

Send message to Kevin

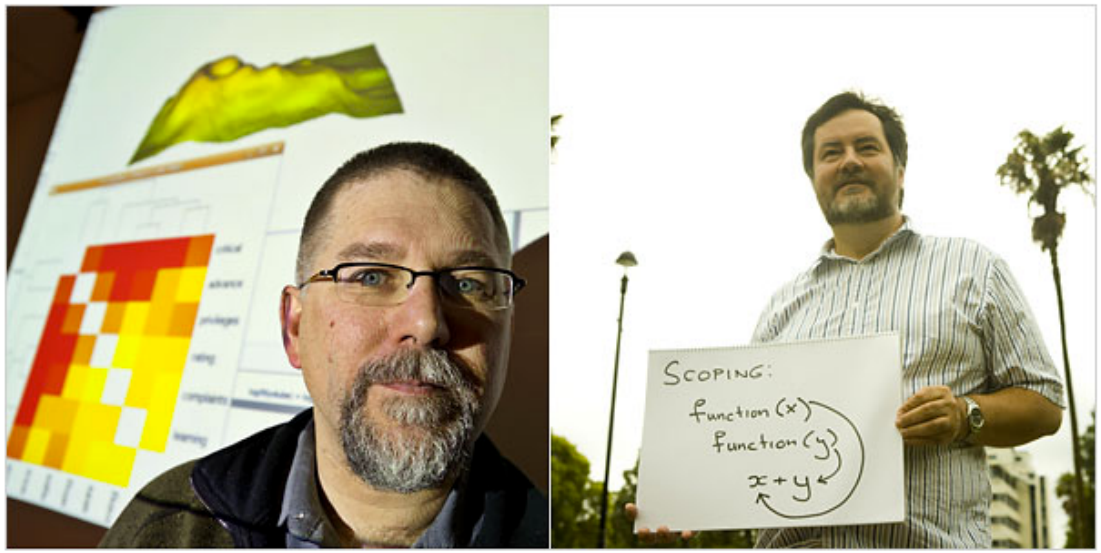
- I might not have explained something clearly;
- You didn't get a chance to ask a question, *etc*

Course software that you've used before: R; and why it's good to keep using it

- Main software: R statistical computing language; we also support Python, Minitab and MATLAB
- Why use R?
  - Widely used: Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group, Shell.
  - Runs on Windows, Linux and Mac computers
  - Excellent add-on libraries available for almost anything related to data analysis
  - Free (both for academic and commercial use): you can use it after you graduate
  - Promotes good statistical practice: write self-documenting code
- Tutorial on website
  - **http://learnche.mcmaster.ca/4C3/Software_tutorial**
    - Step-by-step R tutorial is available there
  - How to install and use software
  - Example of loading data, plotting, data analysis, etc

The course software is well-known and widely used

Data Analysts Captivated by R's Power



Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

SIGN IN TO RECOMMEND
TWITTER
SIGN IN TO E-MAIL
PRINT
SINGLE PAGE
REPRINTS
SHARE

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

Related

Bits: R You Ready for R?

The R Project for Statistical Computing

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

But R has also quickly found a following because statisticians, engineers and scientists

Office hours for the TAs and myself

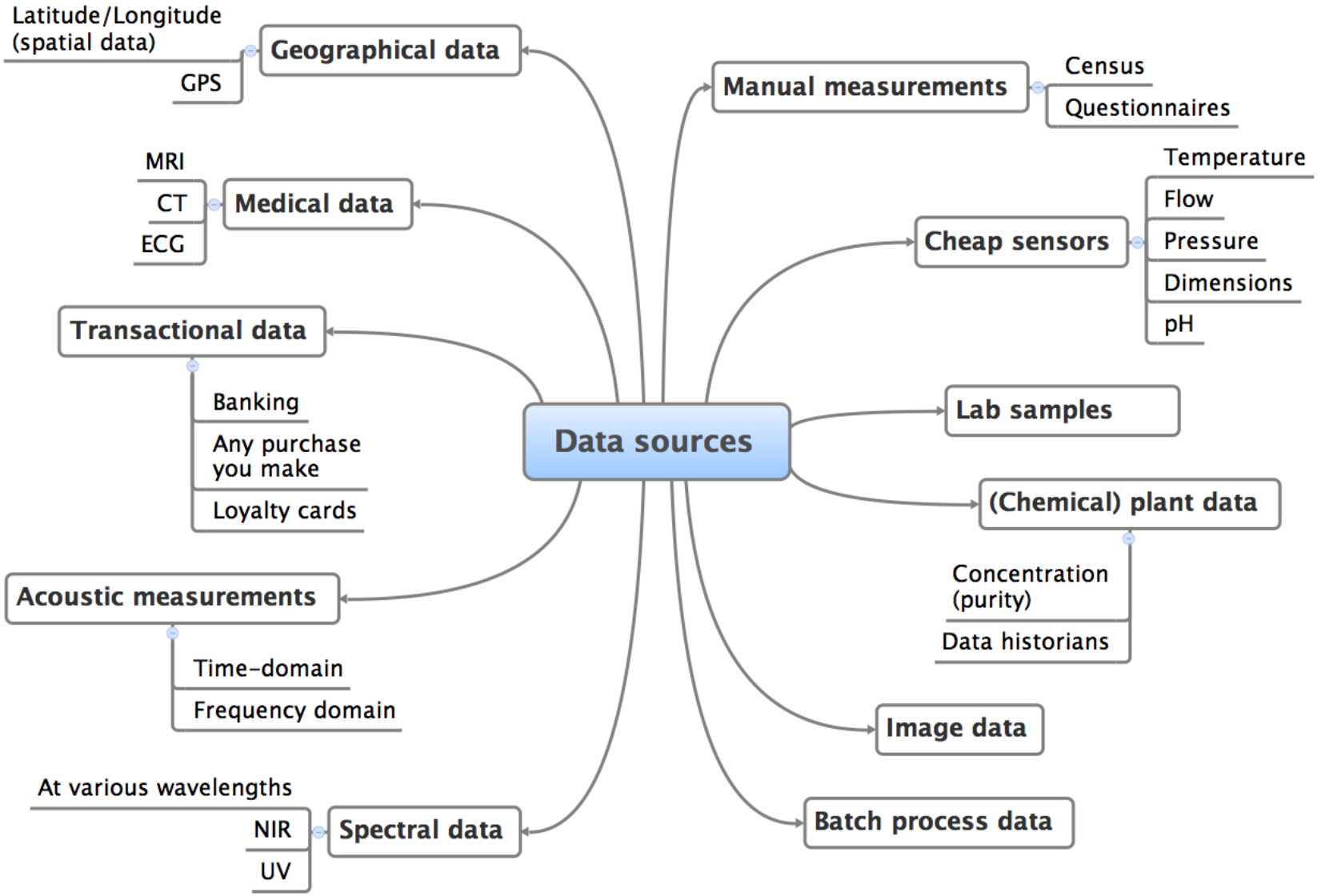
- You can expect TAs and I to answer emails promptly
- If you have questions:
  1. Please email the TAs with CC to me ⟵ hopefully this solves your problem
  2. Please send from your McMaster address
  3. Set up in-person meeting with TAs or myself
  4. My office hours: http://learnche.mcmaster.ca/contact-info

What this course is about

There are 6 main sections, spread over 12 weeks

1. *Visualization*: high-density, efficient graphics
2. *Univariate data analysis*: probability distributions, confidence intervals
3. *Least squares models*: correlation, covariance, ordinary and multiple least squares models
4. Design and analysis of *experimental data* and response surface methods to improve a process
5. *Process monitoring*: tracking process behaviour to detect abnormalities
6. Introduction to *latent variable methods*: a general overview

We are surrounded by interesting data



What prior students have said about this course

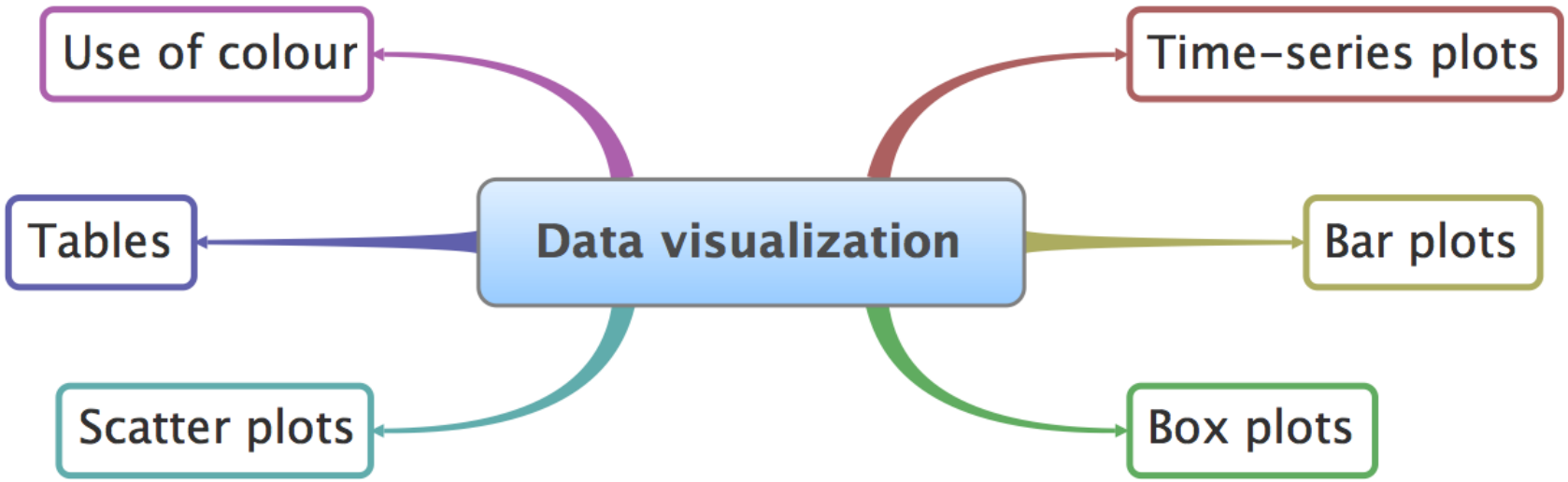
Extracting value from data

Ankit - student in 2010 - now at Oneira Corporation (Oakville):

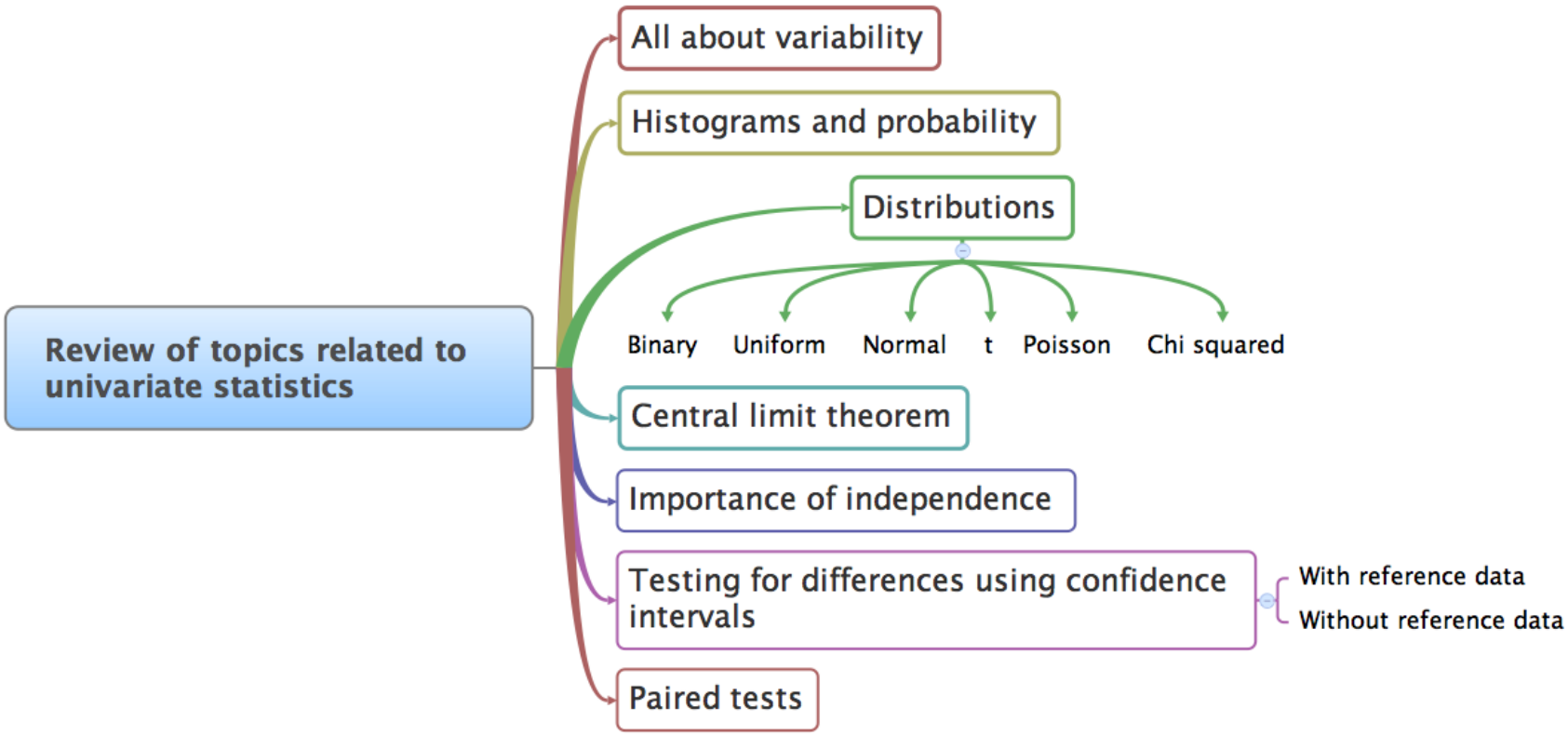
- *Now, having worked for over a year, I find myself referring back to my notes all the time and appreciating the concepts about how to look at data and represent the data in the best possible manner, especially since on a daily basis I look at a gigantic amount of data and am required to make sense of it.*
- *I think what I loved most about the course was the emphasis on the* **thinking** *and* **process of getting to a solution** *instead of the the final solution itself which has been an important attribute to becoming a good engineer or a problem solver/troubleshooter.*

Tiffany at PepsiCo and many other students

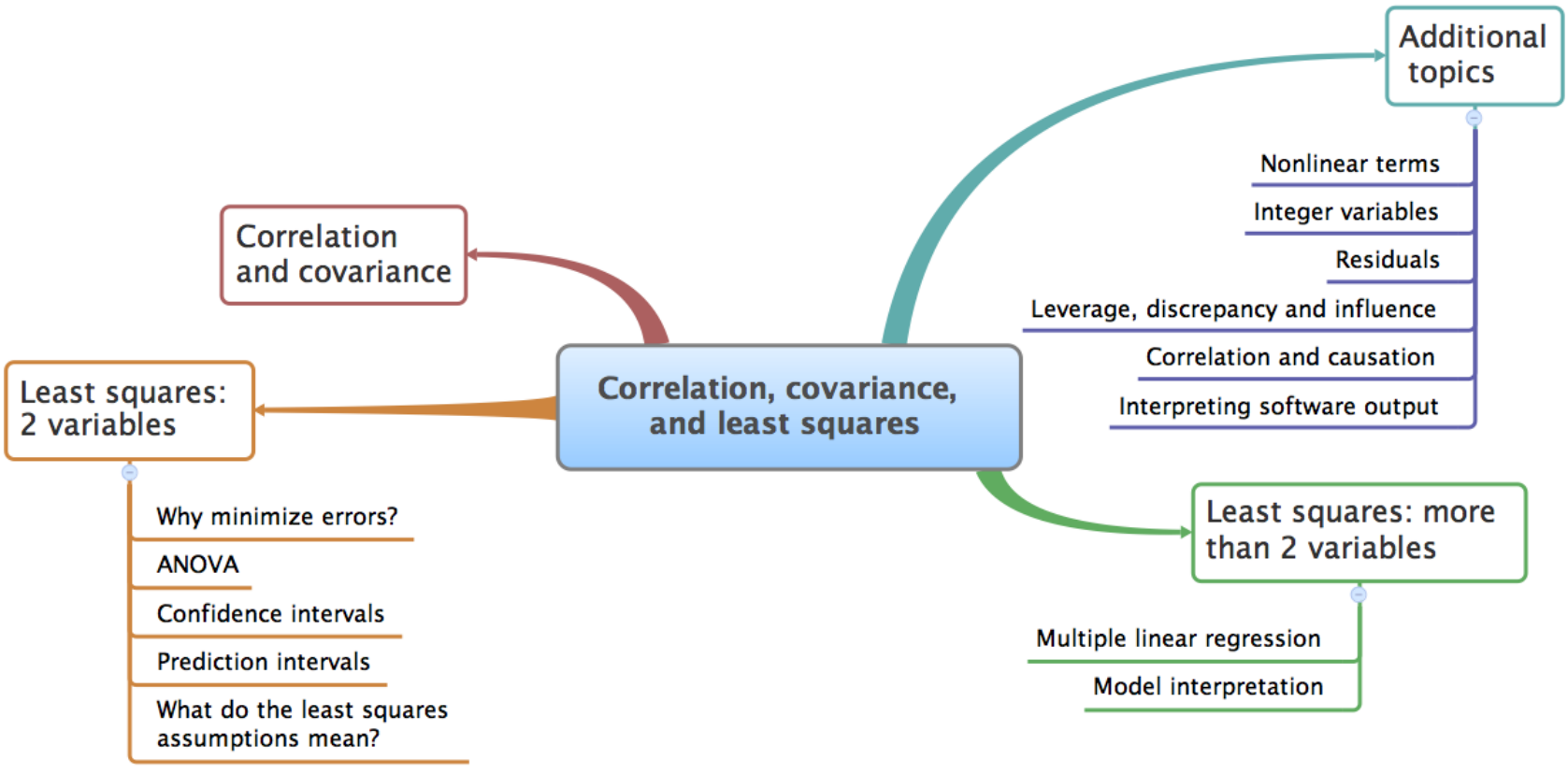
Section 1: data visualization



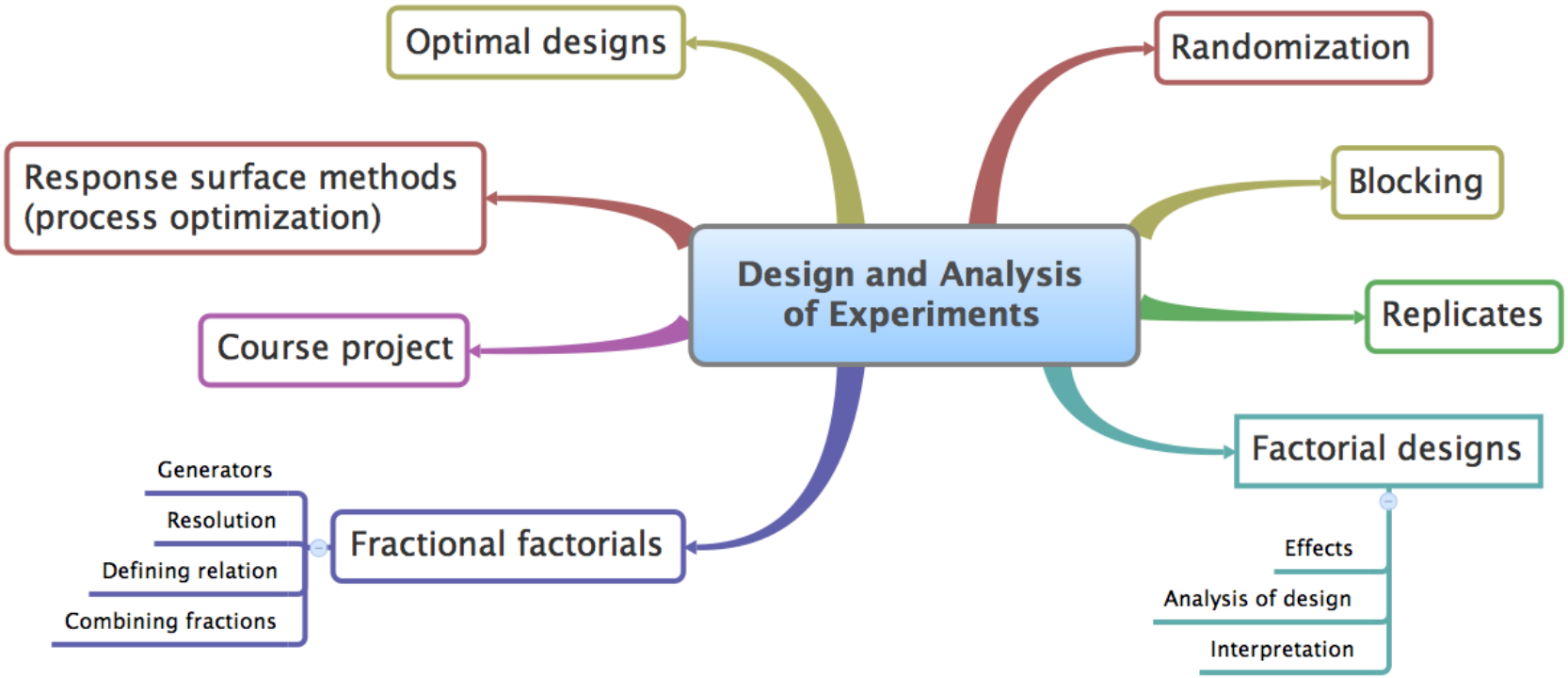
Section 2: univariate concepts review



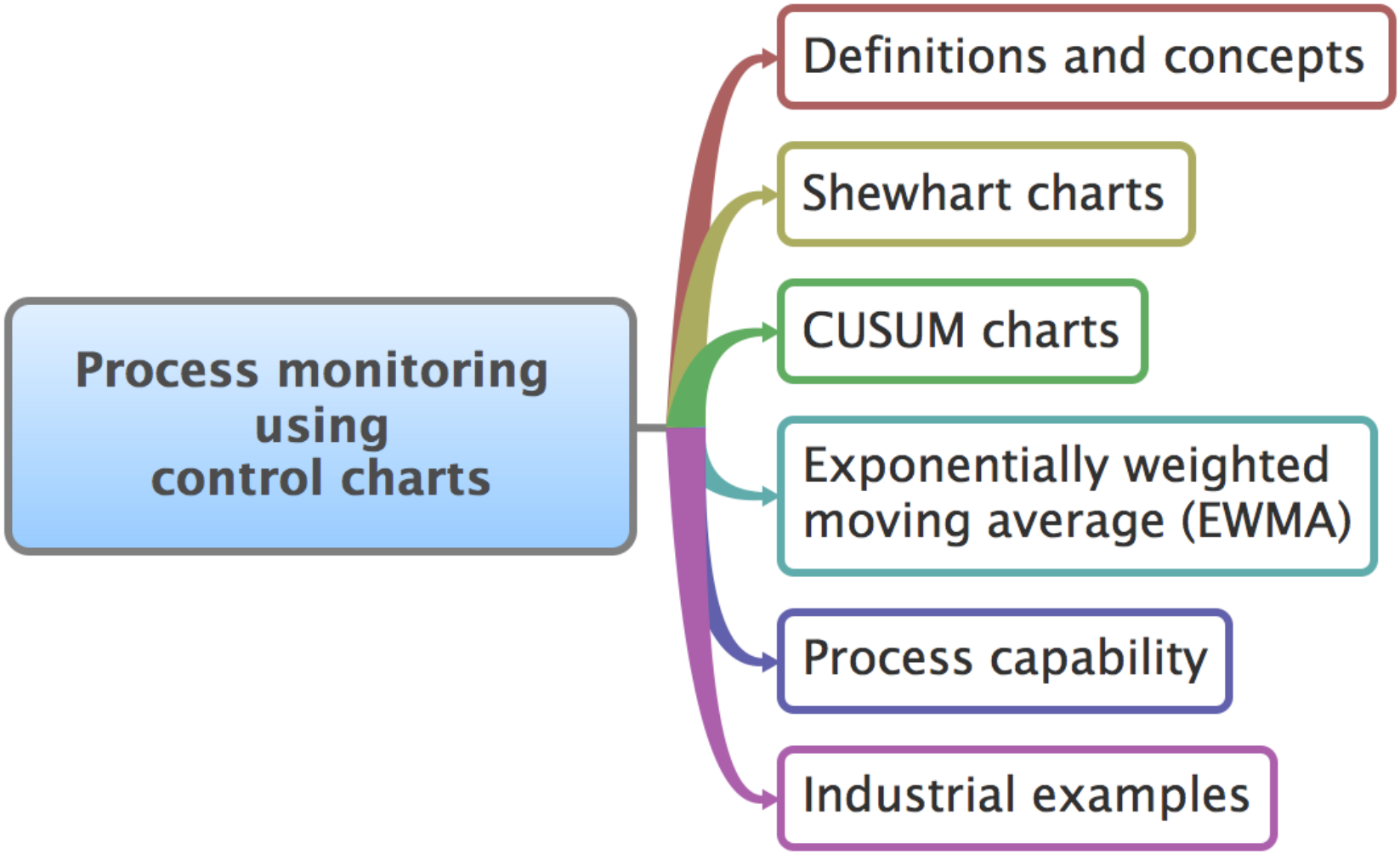
Section 3: least squares



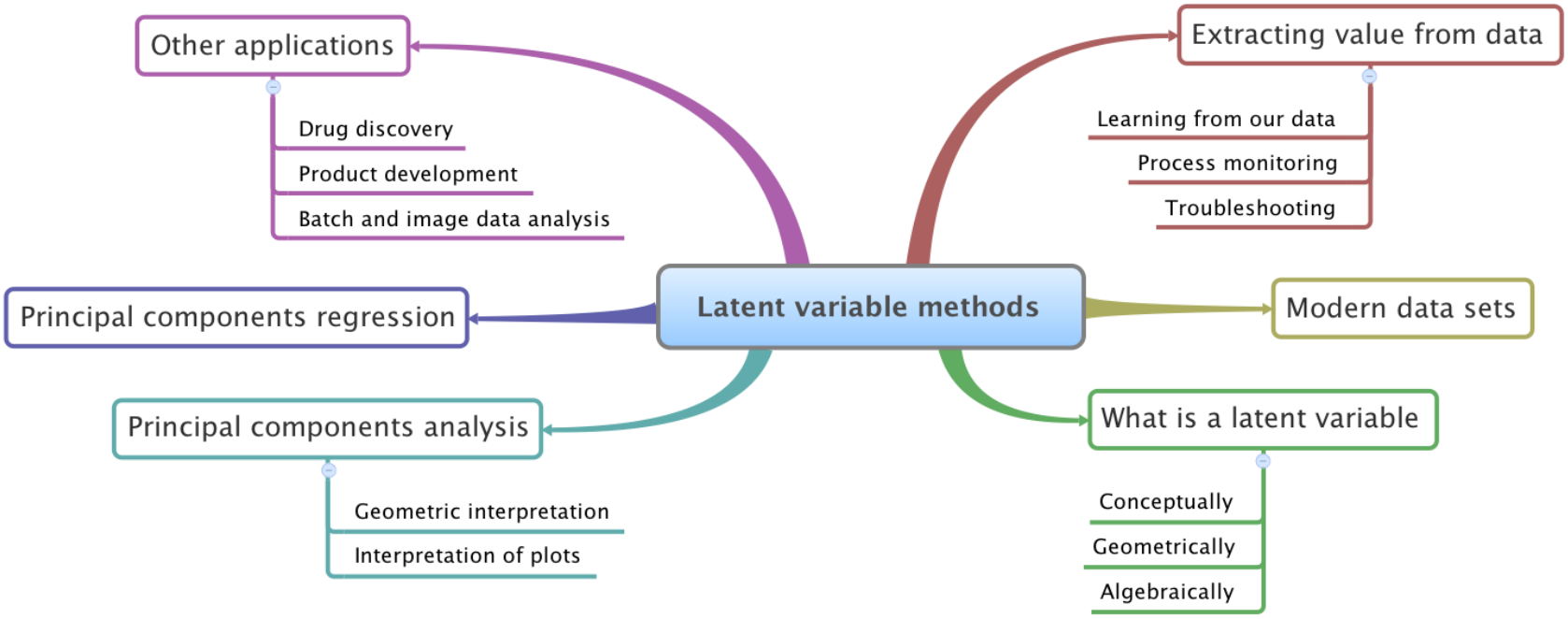
Section 4: design of experiments



Section 5: monitoring any process



Section 6: intro to latent variable methods



Enrichment topics that will be interspersed

- robust methods
- cross-validation
- nonparametric methods
- real-time applications of statistical methods
- missing data handling

What is our grading philosophy?

What we look for in the grading is demonstration that you/group:

1. understand the concept,
2. have the ability to apply the concept to new instances,
3. think creatively about problems,
4. my questions are seldom "plug-and-chug",
5. numerical accuracy,
6. grammar and spelling.

Not all questions will be engineering related:

- mostly (chemical) engineering questions
- but also expect: current world events, public policy, bioengineering, anywhere there are data to analyze

Grading policy in this course

- *Appropriate* group work is highly encouraged
  - Up to 30% of course grade (assignments and project)
  - *Learn with each other*
  - Assignments done in groups of 2 or by yourself
  - Hand-in assignments as one submission
  - Extra credit for 4C3 students if you do the 6C3 questions (where indicated; not always)
  - 6C3 students: you are held to a higher level of quality
- Late grading
  - -30% per day
  - 2 "late day" credits for assignments
  - solutions posted after ≈ 2 days of due date
- Assignment grading:
  - No make-ups for assignments
  - Assignments count **20%** of course grade
  - Best $N-1$ assignments ($N \approx 7$) will be used
- Assignment dates: see website

Group-based assignments

- "Appropriate" group work is highly encouraged (about 30% of course)
- Learn with each other: groups of 2, no larger

- Optimal group work: *an example of one approach*
  - Sarah and Brad work on an assignment
  - Both Sarah and Brad do **all questions** in draft: quick notes at home, on the bus, *etc*, ± 4 days before assignment due
  - Meet in the library next day and go over each other's notes
  - Explain to the other why you disagree
  - e.g. Sarah sees a mistaken interpretation in Brad's work
    - She explains why it is a mistake to Brad: Sarah learns
    - Brad also learns: he's heard this in class, and from Sarah now
    - If neither can resolve it? speak with TA or Kevin
  - Write up a joint solution; e.g. Sarah Q1 and 2, Brad does Q3
  - Both review it before submitting

- Other approaches are possible: your group decides
- What doesn't work: Sarah does Q1 and Q2, Brad does Q3; staple and submit
- **Do not share files or written work** *between* groups

There are weekly tests in this course: why?

- **Fact**: frequent small tests help understand and retain material
- This approach has been successfully used at McMaster in other courses.

- "*Testing effect*": probability of remembering a tested item ("an item that has to be recalled from memory") is greater than an item that was simply *studied*.

- "Spacing effect": breaks between multiple reviews of material, increasing the duration of the breaks, improves recall.
  - "*tent-in-the-wind-with-pegs effect*"

- Sport/practice analogy

Weekly tests

- Testing window: Monday night until Wednesday afternoon (a 40-hour window)
- Test is only about 1 hour in duration, depending on questions
- Covers work reviewed in class
- There will occasionally be questions from prior weeks
- Solutions will be automatically given immediately after the window closes
- Each test counts $\approx 1\%$ of your grades; total of **13%**
- No make-ups; **much lower stakes than the midterms**
- Any notes, materials, websites, electronic documents are allowed, and encouraged.
- **Code of honour**: you must do the test on your own

Grading for project and exams

- Midterm:
  - **12%** of course grade
  - *optional*
- Written final exam: **45%**
  - Covers all material
  - **You must achieve 50% or greater in final exam to pass 4C3/6C3**
- Midterm and final exam:
  - Open notes – anything on paper is allowed
  - No electronic devices unfortunately
  - Any calculator
- Experimental report due on 31 March: **10%**
  - Perform you own *designed experiment*
  - More details on the course website already
  - The earlier you start, the better (start early March)
  - You can, and should, provide an outline of your experimental plan for me to review by 10 March
  - Can collaborate, **but only within your group**: not between groups

Important dates

- **12 February, midterm on Wednesday evening**
  Notify me of clashes this week!
- 10 March: project outline due
- 31 March: course project due
- 07 April: last, review class
- NN April: **Final-exam**

- Due dates for assignments: see course website