Statistics for Engineering

Course overview

Kevin Dunn

http://stats4eng.connectmv.com

2012

Plan for today's class

1. Background
2. Administrative issues
3. Course contents

About myself

- Masters degree from McMaster, 2002
- Not a prof, or doctor, please
- Run my own engineering software company, ConnectMV
- Work full-time at GlaxoSmithKline, Mississauga

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 at least twice throughout the course for your feedback.

Acknowledgments

McMaster Advanced Control Consortium (MACC)

- Dr. John MacGregor, who taught this course for many years

Administrative issues

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

TA introduction

- Yasser Ghobara
- Pedro Castillo

Announcement: Graduate preview

Graduate Preview event: 23 - 24 February

- Meet faculty and existing grad students
- Some entrance scholarships are available
- http://chegradpreview.mcmaster.ca

Apply by 16 January

Video and audio

- Video recordings from 2010 are available on course website
- Poor sound quality though
- Purpose: for your review, and to prepare for assignments and exams
- Might be useful if you miss a class, but previous students have told me that it doesn't always help
- Video recording might be possible this year
  - 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
  - Any objections? Please speak to me after the class; or email me.
- Audio recordings will be made available, when possible

Course website

http://stats4eng.connectmv.com

- Notes and slides (please print out before class)
- Assignments
- Assignment solutions
- Data sets

Website is the main reference for all things course-related

- expected to check it about 3 times per week

Reference text book

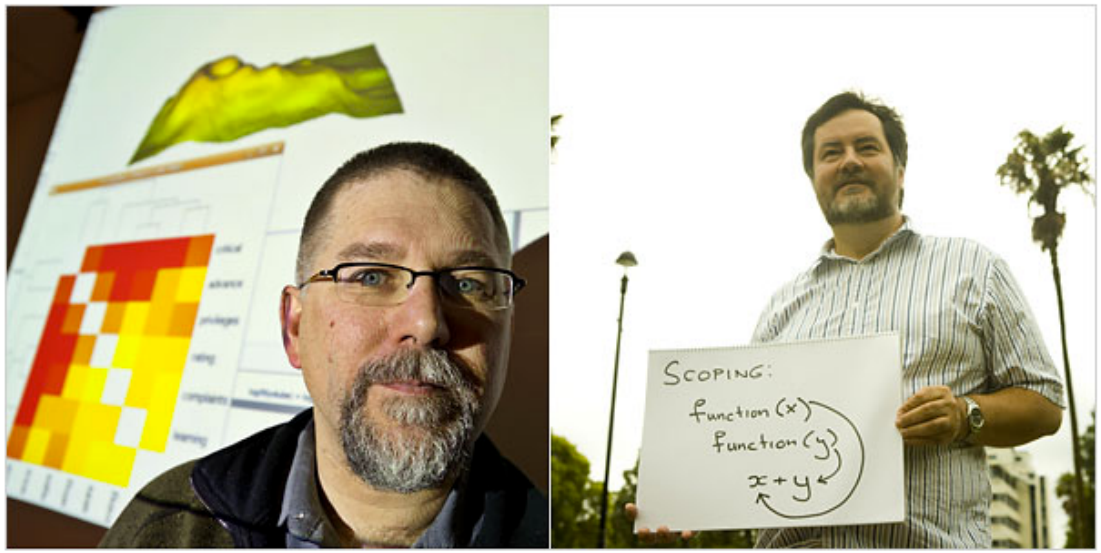
- No mandatory text, just the class notes on website:
  - **Process Improvement using Data**
  - Website: **http://pid.connectmv.com**
  - Draft book
  - Use website to report errors; suggest improvements
- Pre-printed copies available from Titles bookstore
  - $34.60 (just the cost of printing)
  - available around 10 January
- Some suggested books on the course website:
  - Box, Hunter and Hunter

Course software

- A computer is required for assignments, DOE project, and take-home exam
- 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://connectmv.com/tutorials**
    - R, MATLAB, and Python tutorials available there
  - How to install and use software
  - Example of loading data, plotting, data analysis, etc

Course software

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

Expectations outside class

- You can expect TAs and I to answer emails promptly
- If you have questions:
  1. Please email the TAs with CC to me (please send from your McMaster address)
  2. see TAs during office hours (times arranged next week)
  3. set up in-person meeting with TAs or myself
  4. my office hours: 15 minutes before and as long as required after class

Grading

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. accuracy.

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

- *Appropriate* group work is highly encouraged
  - Up to 40% of course grade
  - *Learn with each other*
  - Assignments done in groups of 2 or by yourself
  - Hand-in assignments in one submission
- 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
  - Counts **20%** of course grade
  - Best $N - 1$ assignments ($N \approx 7$) will be used
- Assignment dates: see website

Grading for exams

- Written midterm on 16 February: **15%**
  - optional
  - no make-up
- Written final exam: **45%**
  - Covers all material
- Midterm and final exam:
  - Open notes – anything on paper is allowed
  - No electronic devices unfortunately
  - Any calculator
- Take-home midterm exam due on 22 March: **20%**
  - Requires computer software
  - You have 5 days
  - Includes submitting your own experiment and analysis of it (given 4 weeks in advance)
  - Can collaborate, **but only within your group**: not between groups

Important dates

- 16 February: (mid-term), T28 room 1, 19:00 to 21:00
- 22 March: take-home exam and project due
- 29 March: last class
- 05 April: optional review class (*just after* last official class)
- April: final-exam

- Due dates for assignments: see course website

What is appropriate group work ?

- Roughly equal sharing of responsibility *within the group*
- Have an initial meeting about the assignment
- Allocate tasks to each group member
- Meet to discuss your solutions
- Work together to submit your hand-in
- **Do not share files or written work** *between* groups

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. *Process monitoring*: tracking process behaviour to detect abnormalities
4. *Least squares models*: correlation, covariance, ordinary and multiple least squares models
5. Design and analysis of *experimental data* and response surface methods to improve a process
6. Introduction to *latent variable methods*: a general overview

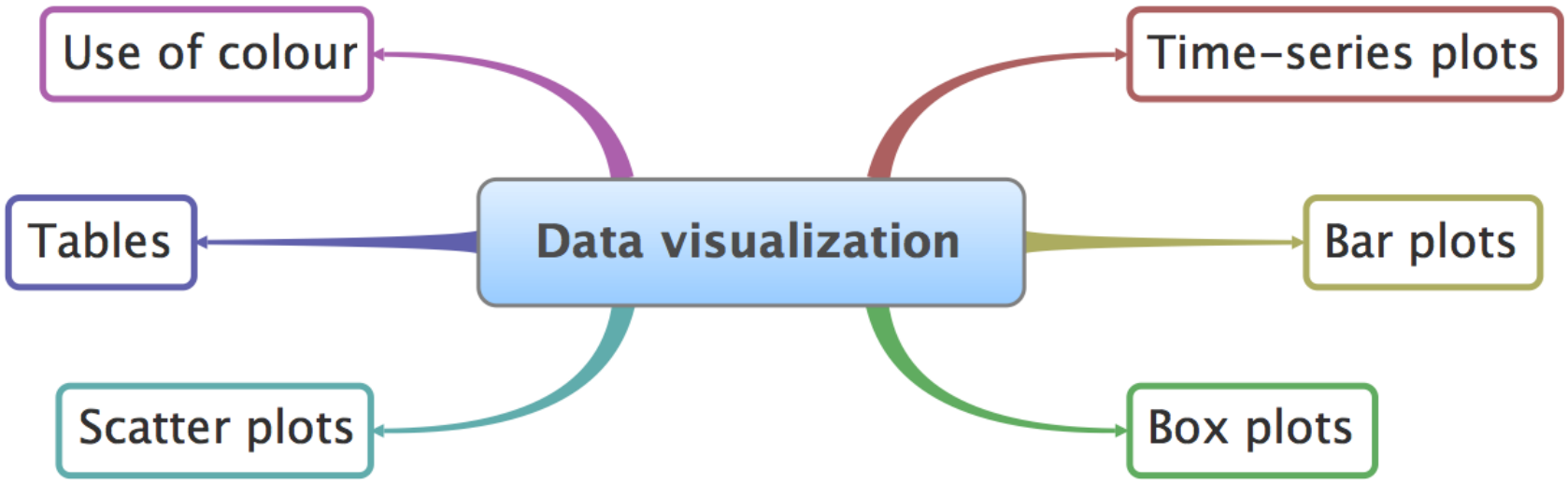
What this course is about

Extracting value from data
Ankit - 4C3 student in 2010 - now at Tenova:

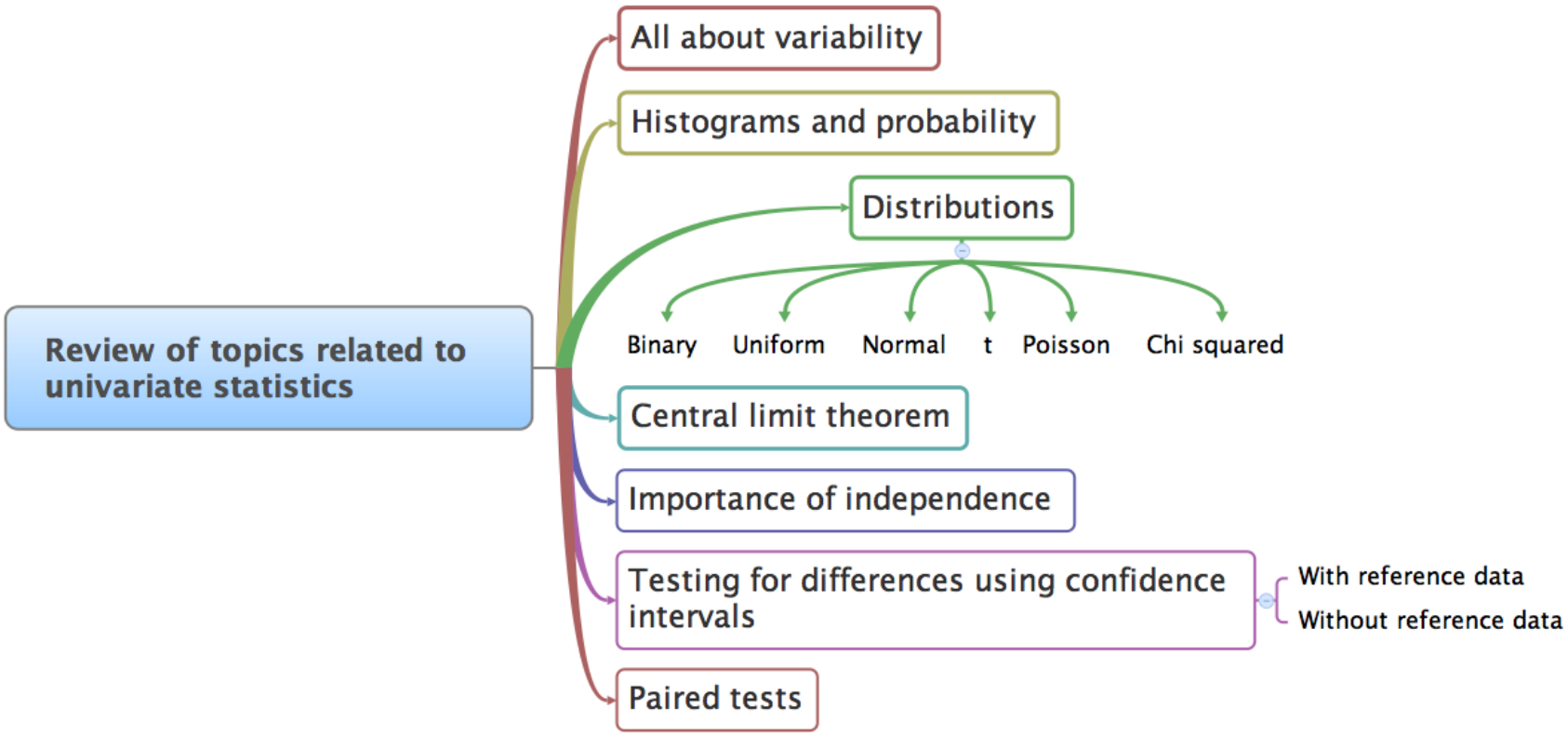
- *Now, having worked for over a year, I find myself referring back to my notes all the time and appreciating the concepts about how to look at data and represent the data in the best possible manner, especially since on a daily basis I look at a gigantic amount of data and am required to make sense 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.*

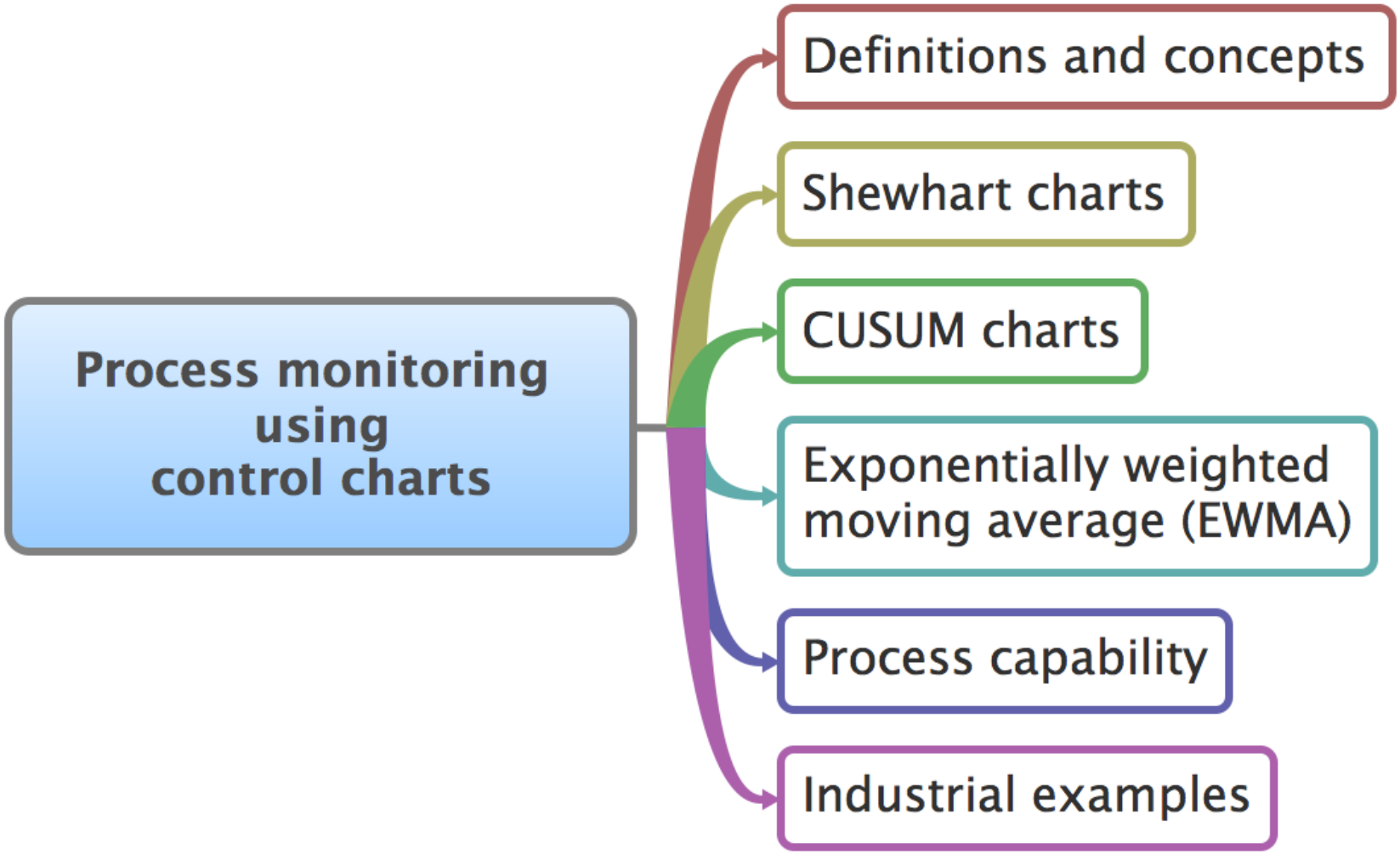
Section 1


Use of colour
Time-series plots
Tables
Data visualization
Bar plots
Scatter plots
Box plots

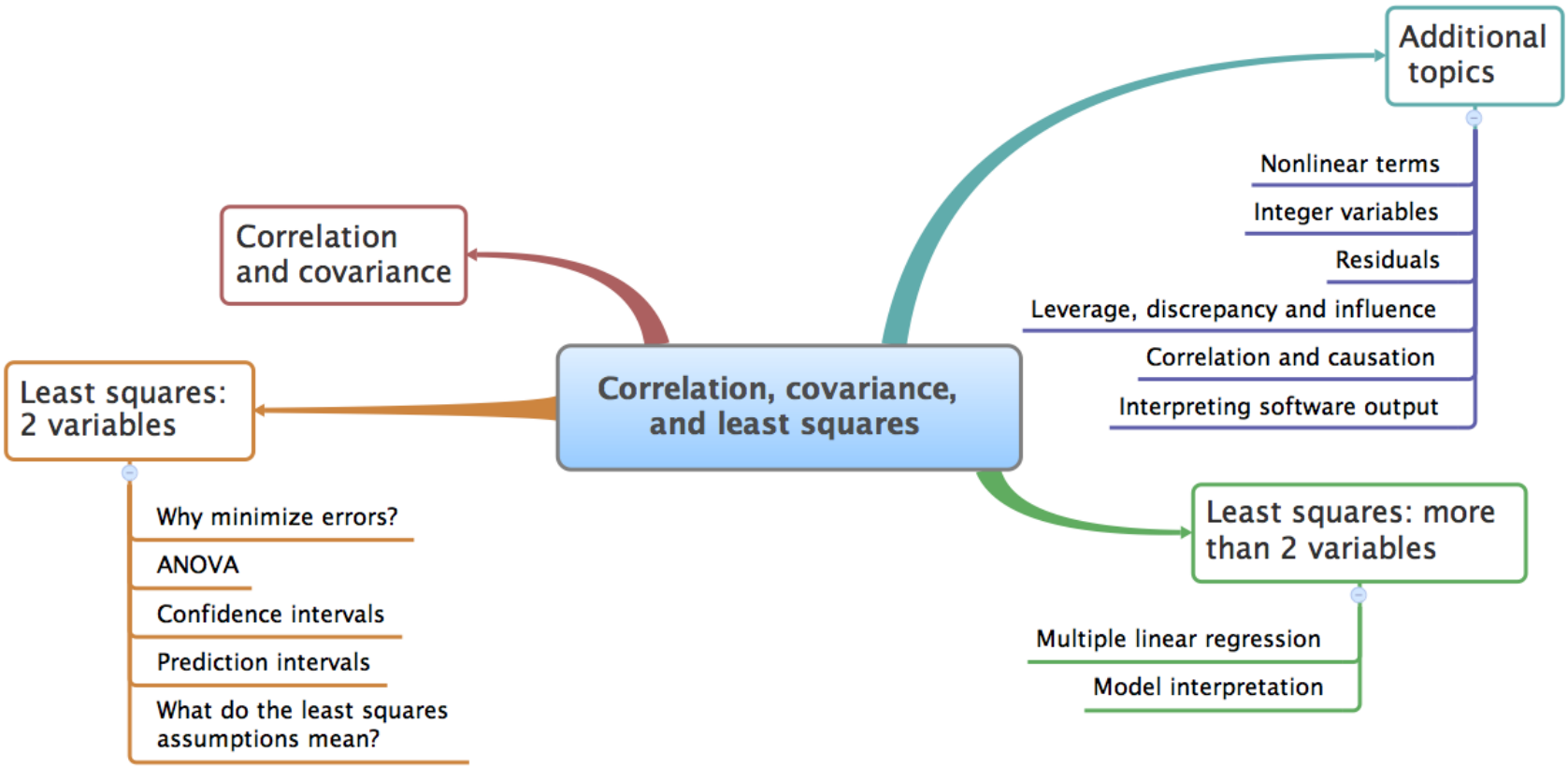
Section 2


Review of topics related to univariate statistics
All about variability
Histograms and probability
Distributions
Binary
Uniform
Normal
t
Poisson
Chi squared
Central limit theorem
Importance of independence
Testing for differences using confidence intervals
With reference data
Without reference data
Paired tests

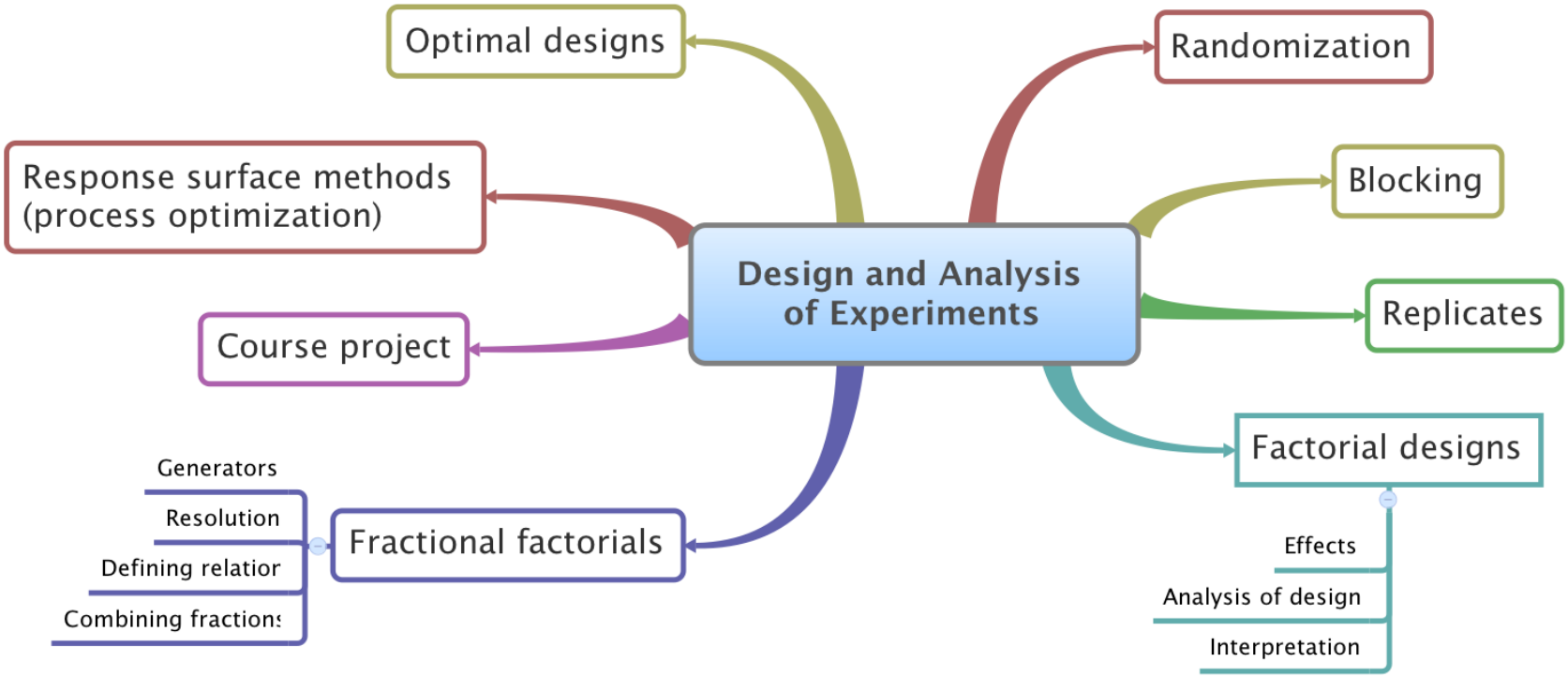
Section 3


Process monitoring using control charts
Definitions and concepts
Shewhart charts
CUSUM charts
Exponentially weighted moving average (EWMA)
Process capability
Industrial examples

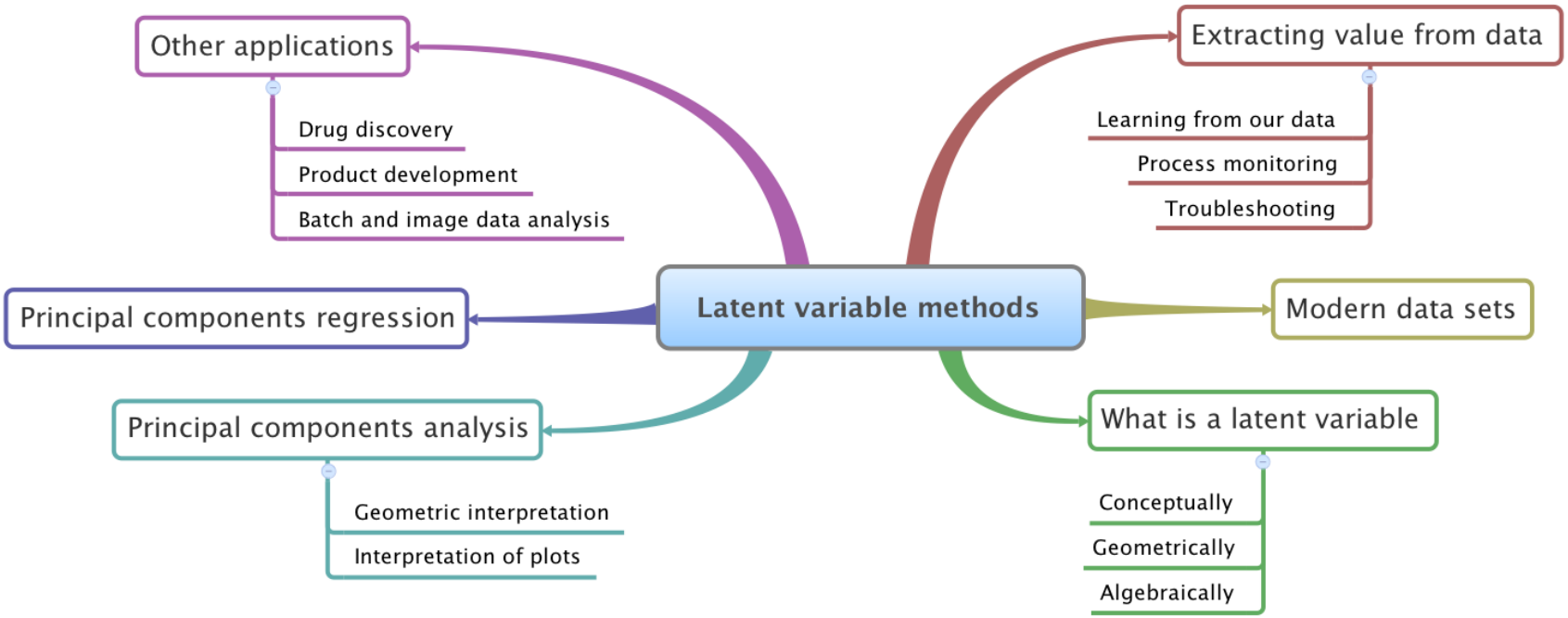
Section 4


Correlation, covariance, and least squares
Correlation and covariance
Additional topics
Nonlinear terms
Integer variables
Residuals
Leverage, discrepancy and influence
Correlation and causation
Interpreting software output
Least squares: 2 variables
Why minimize errors?
ANOVA
Confidence intervals
Prediction intervals
What do the least squares assumptions mean?
Least squares: more than 2 variables
Multiple linear regression
Model interpretation

Section 5


Design and Analysis of Experiments
Optimal designs
Randomization
Response surface methods (process optimization)
Blocking
Course project
Replicates
Fractional factorials
Factorial designs
Generators
Resolution
Defining relatior
Combining fractions
Effects
Analysis of design
Interpretation

Section 6


Latent variable methods
Other applications
Drug discovery
Product development
Batch and image data analysis
Extracting value from data
Learning from our data
Process monitoring
Troubleshooting
Principal components regression
Modern data sets
Principal components analysis
Geometric interpretation
Interpretation of plots
What is a latent variable
Conceptually
Geometrically
Algebraically

Enrichment topics

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