Challenging Data

Prof. Matthew Roughan
matthew.roughan@adelaide.edu.au
http://www.maths.adelaide.edu.au/matthew.roughan/

Prof. Aurore Delaigle
aurored@unimelb.edu.au
University of Melbourne

ACEMS

Nov 1, 2017

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Section 1

Getting Started

Who is Involved in the Theme

Leaders: Aurore Delaigle and Matt Roughan
CI list

- Nigel Bean
- Peter Forrester
- Rob Hyndman
- Kerrie Mengesen
- Tony Pettitt
- Louise Ryan
- Scott Sisson
- Ian Turner
- Matt Wand

Very big and messy data

- How to handle massive datasets? Coarsen? Aggregate? How?
- How do you compute things efficiently?
- Ex: millions of time series. How to visualize/forecast them?
- Ex: ancient DNA data: huge but also very poor quality (don't have what you want) and no way to get missing information.
- Ex: Social media data (big, low-quality, unstructured, ...).
- Ex: traffic data (big, messy, non standard).
- Who is involved? Bean, Delaigle, Hyndman, Garoni, Ryan, Sisson, Wand. Between node/teams collaboration is happening.

Symbolic data

- Big, non-standard data stored in a less traditional form than usual.
- Ex: Continuous data are only observed with a limited accuracy $\Rightarrow$ very big data sets necessarily involve a lot of rounding/ties. Same as “interval data”. Traditional methods do not work.
- Ex: Very big datasets intentionally rounded or summarized in things like histograms. How do you analyze such data?
- Who is involved? Delaigle, Mengersen, Roughan, Ryan, Sisson. Between node collaboration is happening.

Streaming data

- Data keep coming all the time (not collected at once).
- Would be very inefficient to redo all calculations each time new observations arise $\Rightarrow$ need new iterative and efficient algorithms.
- Non stationarity: data process evolve with time: how to deal with this?
- Who is involved? Delaigle, Hyndman, Wand.

Partially observed functional or surface data

- Data are in the form of curves (ex: rainfall over year, growth curves etc) or surfaces (ex: one image per individual).
- Observe only fragments of the individual curves or surfaces.
- Who is involved? Delaigle, Kohn (enabling algorithm team), Mengersen, Ryan. Between node collaboration is happening

Does Challenging mean BIG?

Let's look at some definitions

- *Big data is extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.* OED
  - How big is "extremely" large?
  - And what about astronomy, physics, ... data?

Does Challenging mean BIG?

Let's look at some definitions

- *Big data is extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.* OED
  - How big is "extremely" large?
  - And what about astronomy, physics, ... data?
- *Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications.* WIKIPEDIA
  - What is traditional?
    - parallel processing was invented before digital computers existed
    - likewise sampling

Big Data and its Warts

- There is a misconception that “big” overcome issues (bias, representativeness, ...) and hence has truthiness
- There is a tendency to think Big Data == Machine Learning

https://xkcd.com/1683/

- And BIG doesn't capture what makes a lot of problems hard

A Definition of Challenging

A common definition of challenging data

- volume, velocity, variety, ...
  - volume = big
  - velocity = fast
  - variety = sources, formats, content, ...

But this doesn't really do it so ...

A Big Definition of Challenging

The 27 V's of challenging data

- volume, velocity, variety, ...
    - volume = big
    - velocity = fast
    - variety = sources, formats, content, ...
    - very long = e.g., data collected over decades: *e.g.,* SAX 45 and up
    - variability = inconsistent data rate, *e.g.,* event triggered
    - veracity = big doesn't mean correct, but it does make it hard to clean the data
    - vacuity = there's a lot of data, but no signal
    - priVacy = what it says
    - vagueness = data but no question?
    - volatility = noisy; all over the place
    - vaultification = data is kept locked up in separate boxes
    - victimised = the data has already be tortured
    - verbosity = wordy and unstructured
    - viscousity = hard to wade through
    - vagrancy = hard to pin down
    - vapidity = are we bored yet?
    - ...

A Better Definition

The 27 V's is ludicrous extension of 3 V's to their logical conclusion

- Challenging data is exactly what its name says
- It's a good definition because
  - challenges are defined by exception, not inclusion
    - *e.g.,* we say "it's not possible to do X with data Y"
  - it encompasses the variety of what we do without endlessly extending the definition
  - **what is challenging for you might not be for someone else, so this definition encourages collaboration**

What Are the Challenges

Often the focus is on "hardware"

- data capture
- data storage
- (raw) data processing

We care more about the other end

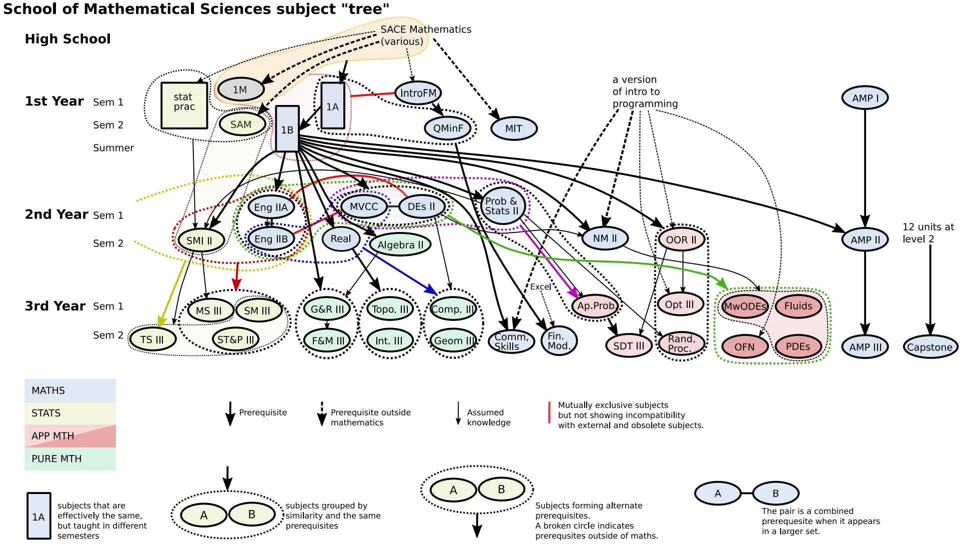
- data analysis and algorithms
- information distillation

In between

- data representation
  - tendency of data frames, JSON, ... is to push us towards flat data representations
  - I care about networks
  - representation affects
    - storage
    - algorithm performance
    - visualisation

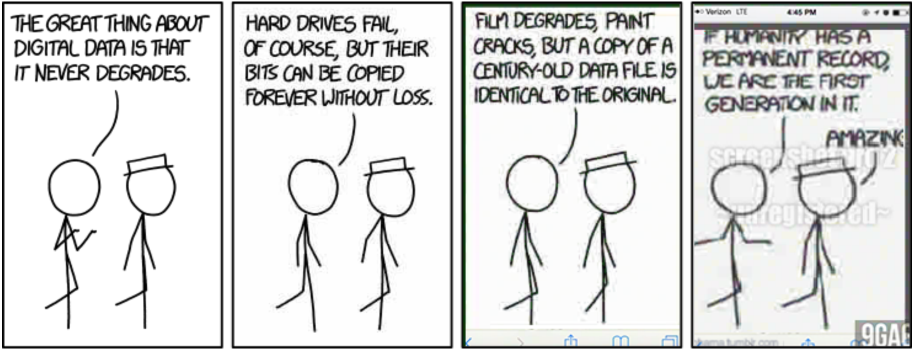
A Data Representation Example

Course Hierarchy as a **Metagraph**



Conclusion

I don’t like endings, so here let’s go with this:



https://xkcd.com/1683/

And now for a couple of more interesting talks on particular Challenging Data problems.