A Few Words on Big Data

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- Big data are, quite literally perhaps, everywhere.

- But what are big data and how can we deal with them?

Big Data

- Often defined in terms of three (or more) Vs:
  - volume (high-dimensional, i.e., lots of variables)
  - variety (variables of mixed type)
  - velocity (data keep coming)

- Much interesting work, including the short article by Puts et al. (2015), focuses on these three Vs.

- “Veracity” is another V-word that is often relevant. . . however, it may not be easy (possible?) to assess.

Big Data vs. Administrative Data

- With advances in computing, big data are now commonplace.

- Puts et al. (2015)[a] draw an important distinction between big data and administrative data:

  > Having gained such experience editing large administrative data sets, we felt ready to process Big Data. However, we soon found out we were unprepared for the task.

- This interesting observation aside, the precise meaning of the term big data is less important than using, and where necessary, developing appropriate approaches to tackle emerging data types.

[a] Puts, M., Dass, P. and de Waal, T. (2015). ‘Finding errors in Big Data’. *Significance* **12**(3), 26–29.

Comments

- Interestingly, in the mixture model-based context, there are approaches for data with Vs.

- For further reading, I would start with Puts et al. (2015).

- However they are defined, I think it is fair to say that big data may well be paradigm altering. . .