VB21-Machine Learning with Big Data
Day 1

Akitaka Matsuo

Essex IADS

ecpr

About me

- Aki Matsuo
  - Research fellow in Institute for Analytics and Data Science (IADS) and Department of Government, University of Essex
  - Ph.D in Political Science, Rice University
  - Research Interest
    - Political methodology (text analysis, scaling)
    - Legislative politics
    - Social media analysis
      - UK Politics / Japan v Korea
  - Member of quanteda project
- TA: David Liao, Ph.D candidate in Gov Dept, Uni of Essex

Your turn

- Name
- School/Company
- Discipline
- Position
- Your motivation
  - Big Data or ML

Course schedule

- Day 1: (Big) Data Management
- Day 2: Database + Machine Learning Basics
- Day 3: Regression
- Day 4: Classification
- Day 5: Tree-based Method + Sparklyr

Schedule Day 1

- Course info
- Discussion
- Rstudio setting up
- tidyverse exercise
  - How it works
  - I will give you some time to work on coding
    - Group work
  - Then I will show the answer

Course information

- Each class:
  - A bit of review and coding exercises
- If you are taking this course for a credit, please attend all lectures (or let me know if you will miss a class)
- For additional credit, do the take home assignment
  - Find the data, then carry out the analysis
- We will have office hours
  - What's your timezone?
  - 20 min slots
  - Google spreadsheet
  - Send an email to David to put you on the list
  - One zoom room for office hour
- Social?

Discussion

- Three-V of big data
    - Volume, Velocity and Variability
  - Which-V do the data of your interest fit in?
- Grimmer (2014) suggests two direction
    - Better causal inference
    - Improved measurement
  - How do your research interests relate to this argument?
- Lazar and Radford (2017)
  - Nowcasting. Should we be interested? Especially Covid-19 related ones?
  - Big data + Field experiments. What should be allowed?
    - You don’t know about the boundary of the influence