STAT:4740
Large Data Analysis Capstone

Intro – Big Data, Ethics, and R

Lecture 1
Jan. 16, 2018

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

Big Data

- From a scientific perspective, Big Data refers to extracting useful information from large, diverse, distributed, and heterogeneous data sets to accelerate scientific discovery and innovation (NSF Big Data Initiative, 2012).
- From a business perspective, it means using integrated data storage, analytics, and applications to help drive efficiency, quality, and personalized products and services, and to create new levels of business value (EMC Business Overview).
- Doug Laney (2001) – big data challenges
  - volume
  - velocity
  - variety

`http://magazine.amstat.org/blog/2014/05/01/statview-big-data/`

Skills needed for Big Data analysis

- the statistical skills to build and interpret appropriate models given the usually huge and complicated data
- the computing/engineering skills to carry out all necessary operations, including data retrieval, scalable optimization, and data visualization

`http://magazine.amstat.org/blog/2014/05/01/statview-big-data/`

Examples of Big Data

- scientific: `http://terra.nasa.gov/about/terra-instruments/modis`
  - transmits an average of 6.1 Mbits/sec 24/7
  - 250 m spatial resolution
  - 1-2 day temporal resolution
- business examples: `http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell`

This class – topics

- Overview of statistical and machine learning methods of learning from data using R
- Data visualization using R
- Storing and handling huge data using Hadoop and parallel computing
- Mathematical tools needed for Big Data
- Clustering algorithms for learning from data

This class – structure

- Blended format for first 8 weeks
  - Lectures/labs in classroom on Tuesdays
  - Videos, readings, and assignments for you to carry out on your own instead of scheduled Thursday class
- Group project work and presentations for last 7 weeks

Why do Big Data raise ethical issues?

- the OKCupid data release
  `http://fortune.com/2016/05/18/okcupid-data-research/`
- Can Facebook influence your emotions?
  `http://www.pnas.org/content/111/24/8788.full`
- determining the identity of Banksy
  `https://poseidon01.ssrn.com/delivery.php?ID=607074089082020071017121093006078100019000003105206`
- research ethics in the age of big data
  `http://www.forbes.com/sites/kalevleetaru/2016/06/17/are-research-ethics-obsolete-in-the-era-of-big-dat#2ea7c4af7aa3`

Ethical considerations in research and data analysis

- integrity of data and methods

  > The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may impact the integrity or reliability of the statistical analysis. Objective and valid interpretation of the results requires that the underlying analysis recognizes and acknowledges the degree of reliability and integrity

- reproducible results

`http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx`

Ethical considerations in research on human subjects

- ethics: norms for conduct that distingush between acceptable and unacceptable behavior `https://www.niehs.nih.gov/research/resources/bioethics/whatis/`
- prevent harm and suffering
- inform research subjects

  > Research subjects are to be given all the information they require to gain a reasonable understanding of the field of research in question, of the consequences of participating in the research project, and of the purpose of the research. Subjects shall also be informed about who is funding the research.

- obtain free and informed consent

  > As a general rule, research projects that include individuals can be initiated only after securing participants free and informed consent.4 The informants have the right to withdraw from participation at any time, without this entailing any negative consequences for them.

  - protection of vulnerable populations from coercion
- respect individuals' privacy and close relationships

  > Researchers shall show due respect for an individuals privacy. Informants are entitled to be able to check whether confidential information about them is accessible to others.

- respect confidentiality
- store information that can identify individuals securely

`http://graduateschool.nd.edu/assets/21765/guidelinesresearchethicsinthesocialscienceslawhumanit pdf`

U.S. regulations in human subjects research

- developed in response to egregiously unethical medical studies conducted in the 1940s and subsequently
- `https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html#46.102`
  - " this policy applies to all research involving human subjects conducted, supported or otherwise subject to regulation by any federal department or agency which takes appropriate administrative action to make the policy applicable to such research"
  - one of the exceptions "(4) Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects."
  - requires institutions with federally funded research to have Institutional Review Boards (IRBs) that approve and monitor procedures in human subjects research
- UI Human Subjects Office `https://hso.research.uiowa.edu/`

Ethical considerations specific to big data

- http://bdes.datasociety.net/
- http://sites.nationalacademies.org/DBASSE/BBCSS/Protection_of_Human_Subjects_in_Behavioral_and_Social_Sciences/index.htm

Why R

- free and open-source
  - anyone can view and modify source code
- the leading tool for statistics, data analysis, and machine learning
- capabilities of professional graphics
- full-fledged programming language
  - links to standard linear algebra libraries
  - has its own math library
  - object oriented
- platform-independent – runs on Windows, Mac, Linux, etc.
- allows integration with other programming languages and interaction with many data sources
- straightforward parallelization of computation
- large and active user community
- over 8000 add-on packages for specialized tasks

https://www.r-bloggers.com/why-use-r/
https://cran.r-project.org/