Ethics, Privacy, and Data Science

Leanne Wu (*she/her*)
Department of Computer Science
lewu@ucalgary.ca

DATA 601, Fall 2019

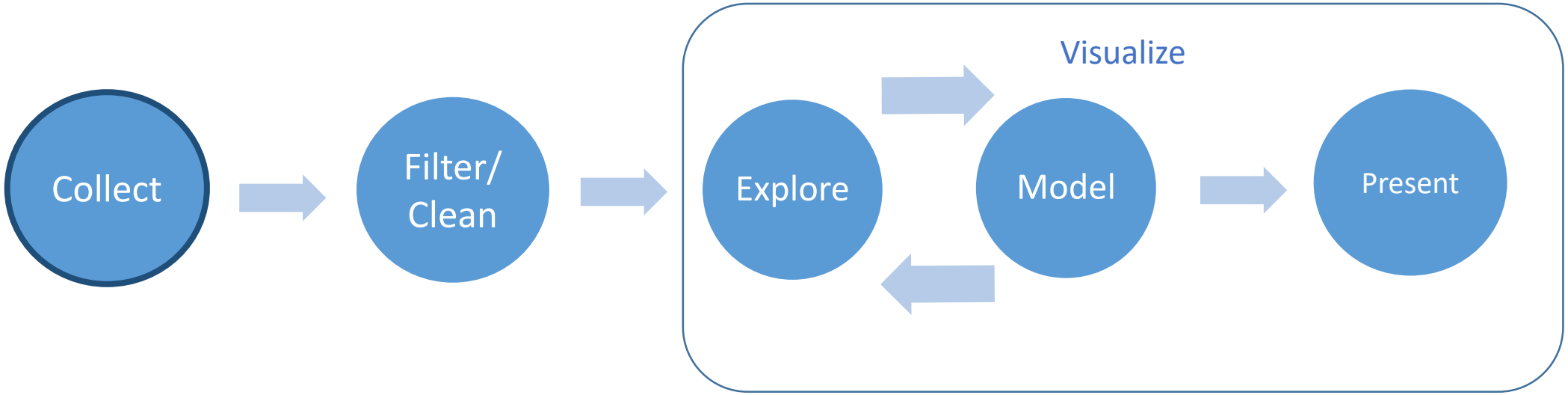
Ethics

What does it mean to be ethical?

What do you bring to your data science practice?

- Go here and take some time to take one (or more) of the tests: https://implicit.harvard.edu/implicit/takeatest.html
  - Did you get the result you expected?
- We all bring a particular point of view to our practice
  - Culture
  - History
  - Personal experience
  - Views and opinions of those close to us

The data science lifecycle



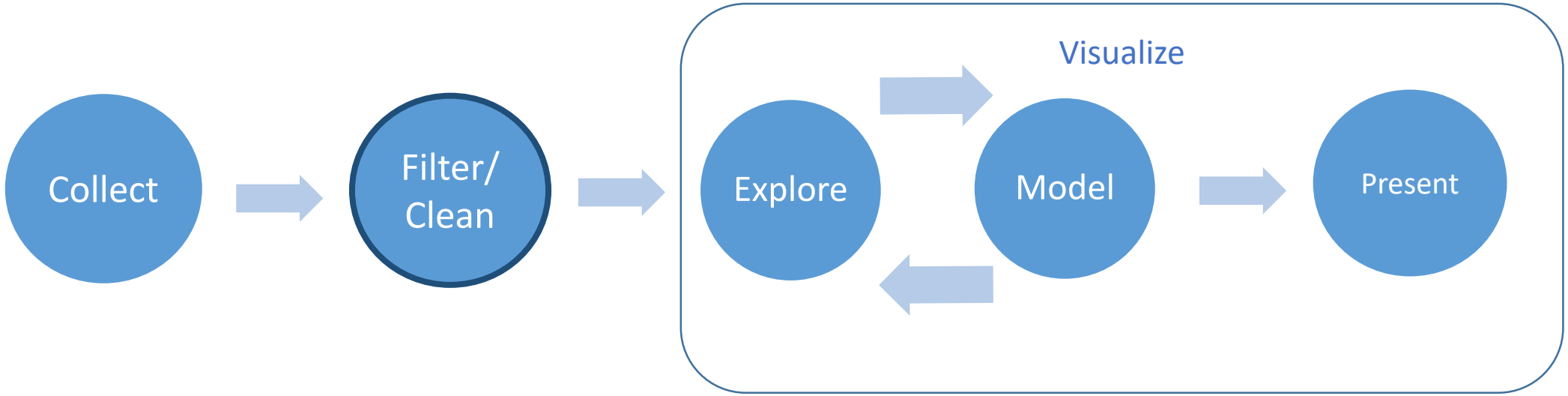
Data Collection

- How are you collecting your data?
  - New datasets:
    - Privacy
    - Representation, inclusivity, fairness
    - Ethical study design
  - Pre-existing datasets:
    - Copyright and Intellectual Property
- Designing the use of your data
  - Security
    - Where is your data located/who has access to it/how is it protected?
  - Do you intend to reuse your data?
    - Metadata
    - Descriptions of dataset, attributes

Privacy

- Article 12 of Universal Declaration of Human Rights (1948): “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.”
- OECD Principles of Privacy (1981)
- In Canada: Personal Information and Electronic Documents Act (PIPEDA), Privacy Act
  - In Alberta: Freedom of Information and Privacy Protection Act
- In European Union: General Data Protection Regulation (2018)

The data science lifecycle



Data Cleaning and Filtering

- Is it necessary to retain all of your data?
  - microdata vs anonymized data
  - does the data wrangling process change the meaning of your data?
- Are outliers important?
- Why have you picked the methods you have to clean or filter your data?
  - What does this affect?
  - “sanity check”: how representative of your population is your data?

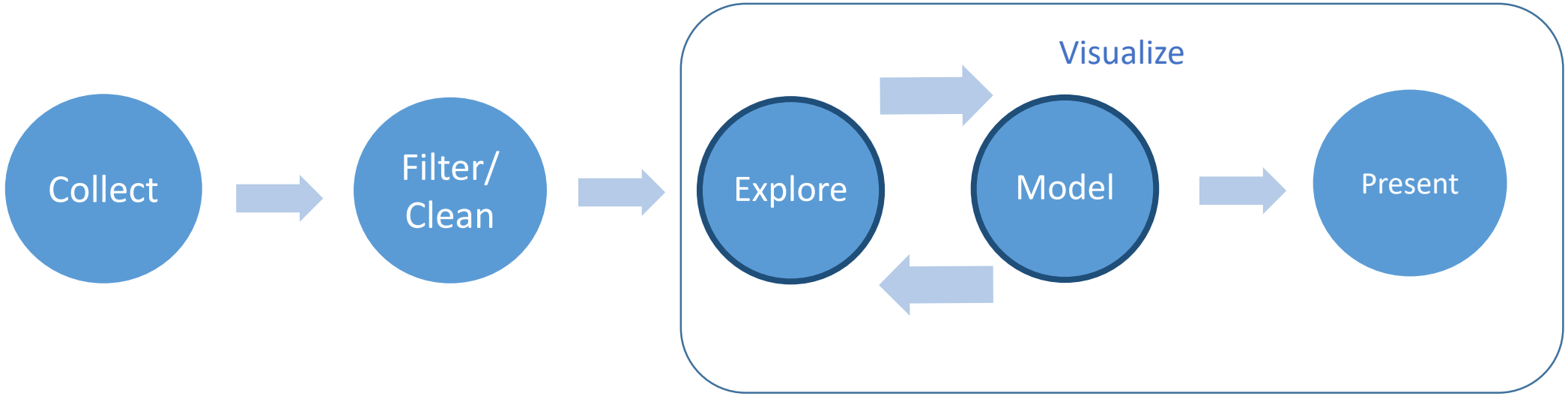
Privacy with large datasets: Anonymization and Differential Privacy

There are two main approaches to protecting the privacy of individuals in large datasets

- Remove detail (Anonymization)
  - Deletion of some attributes, or generalization of others (removing characters in a postal code, for instance)
  - Resulting dataset should meet certain statistical criteria to guarantee privacy
  - *k*-anonymization (Sweeney, 2002), ℓ-diversity (Machanavajjhala, 2007), others

- Add noise (Differential Privacy)
  - Perturb values for different members of the dataset so that statistical measures of the dataset are the same, but harder to guess the original values of each attribute
  - Differential Privacy (Dwork, 2006)

The data science lifecycle

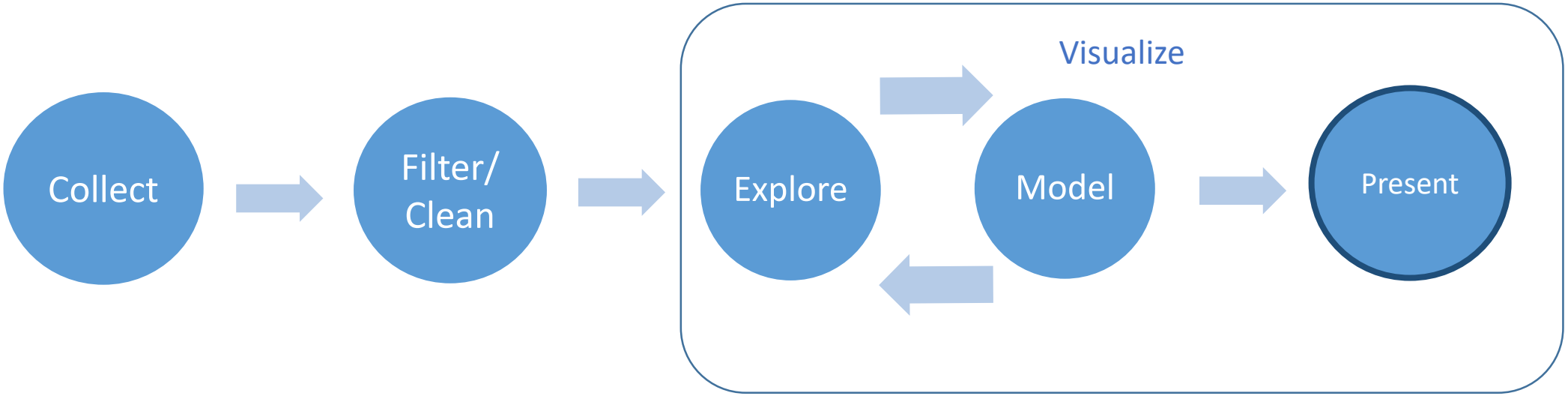


Data Modelling and Exploration

- Consider the data processing tasks used for HW3
  - How did you decide whether a day was rainy? Would this change if you lived in Seattle or Vancouver?

- Analytics frequently try to predict patterns based upon pre-existing data
  - “Garbage in, garbage out”
  - Outliers
  - Problems with “AI” systems (*Weapons of Math Destruction*, O’Neil, 2016)

The data science lifecycle



Visualization (Presentation)

- Besides questions of inclusion, importance, you may also consider the impact of good design
  - Mapping: What happens when you move the centre of the map? How does GapMinder use visualization to challenge static views of social development?
- Principles of Universal Design
  - Many resources out there! (For example: https://www.microsoft.com/design/inclusive/)

Ethics is integral to good data science

- The field is just beginning to understand the impact of analytics-driven decision-making on society
  - Politics (influencing voting patterns)
  - Law enforcement (profiling, predicting crime, identifying persons of interest, sentencing)
  - Health care (We have more data on some demographics than others)
  - Automation and self-driving vehicles
  - Education (Course selection, course design, personalized learning)

Doing better data science

- Acknowledge biases (conscious or unconscious)
- Include as many different points of view as possible
- Design data analysis with participation of stakeholders where possible
  - Who are your subjects/users/customers?
- Be mindful of your legal obligations
  - Protect your data