Unclassified / Non classifié

AI Methods in Public Health: Session 2

Lisa Lix, PhD
Professor, College of Community and Global Health,
Rady Faculty of Health Sciences, University of Manitoba

Gemma Postill
MD/PhD Student, Clinical Epidemiology and Health Care Research,
Institute of Health Policy, Management and Evaluation, University of Toronto

Affan Shoukat, PhD
SGI Assistant Professor in Data Science, Department of Mathematics and Statistics
Faculty of Science, University of Regina

AI4PH AI FOR PUBLIC HEALTH

Public Health Agency of Canada | Agence de la santé publique du Canada

AI Methods in Public Health

01 **Section one: Lisa Lix**
Exploring health data with AI; Choices: statistical models vs. machine learning models

02 **Section two: Gemma Postill**
Predictive and risk modelling

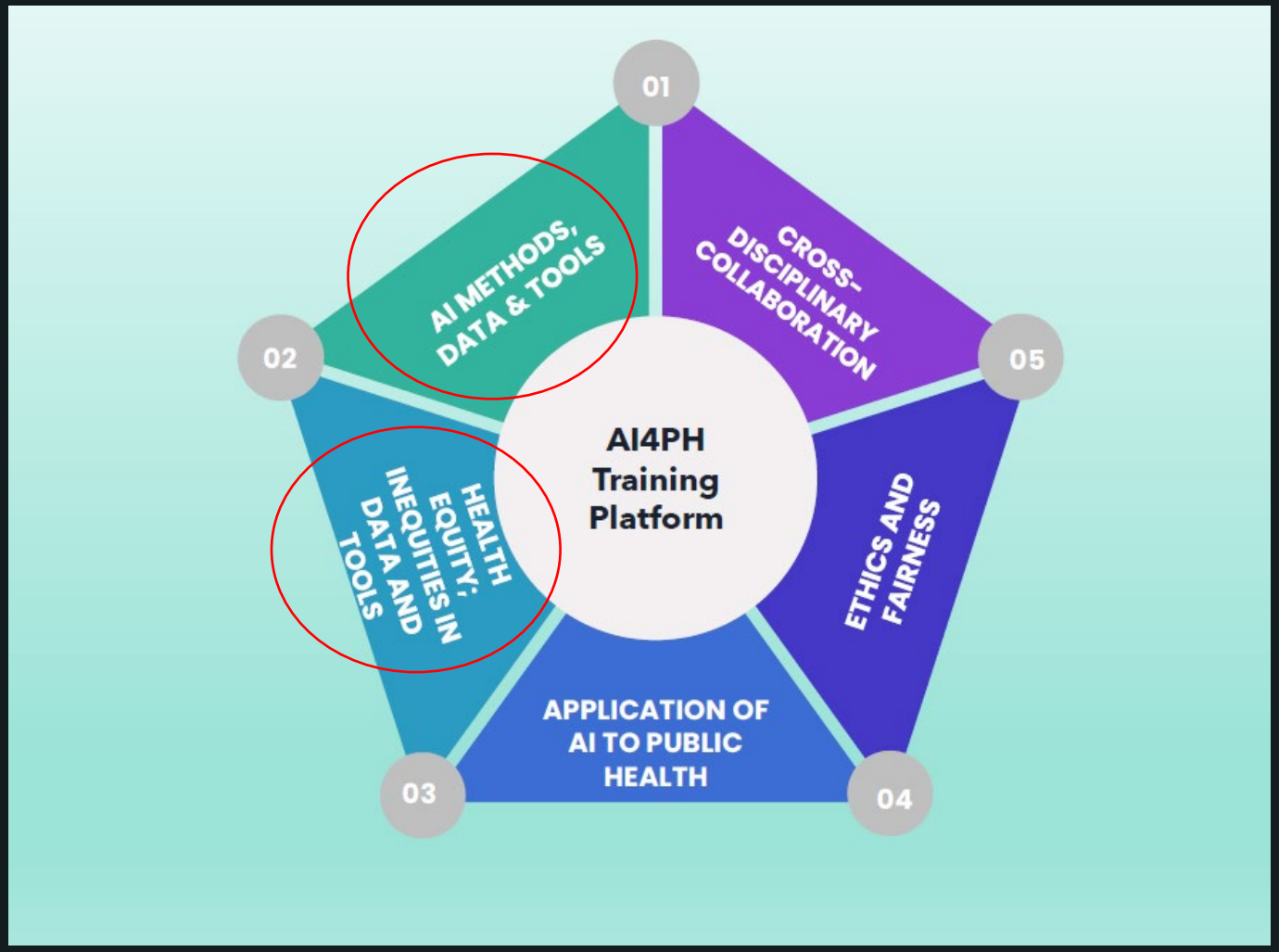
03 **Section three: Affan Shoukat**
Machine learning, forecasting, and explainability

Behind the Numbers: Health Data & Statistics

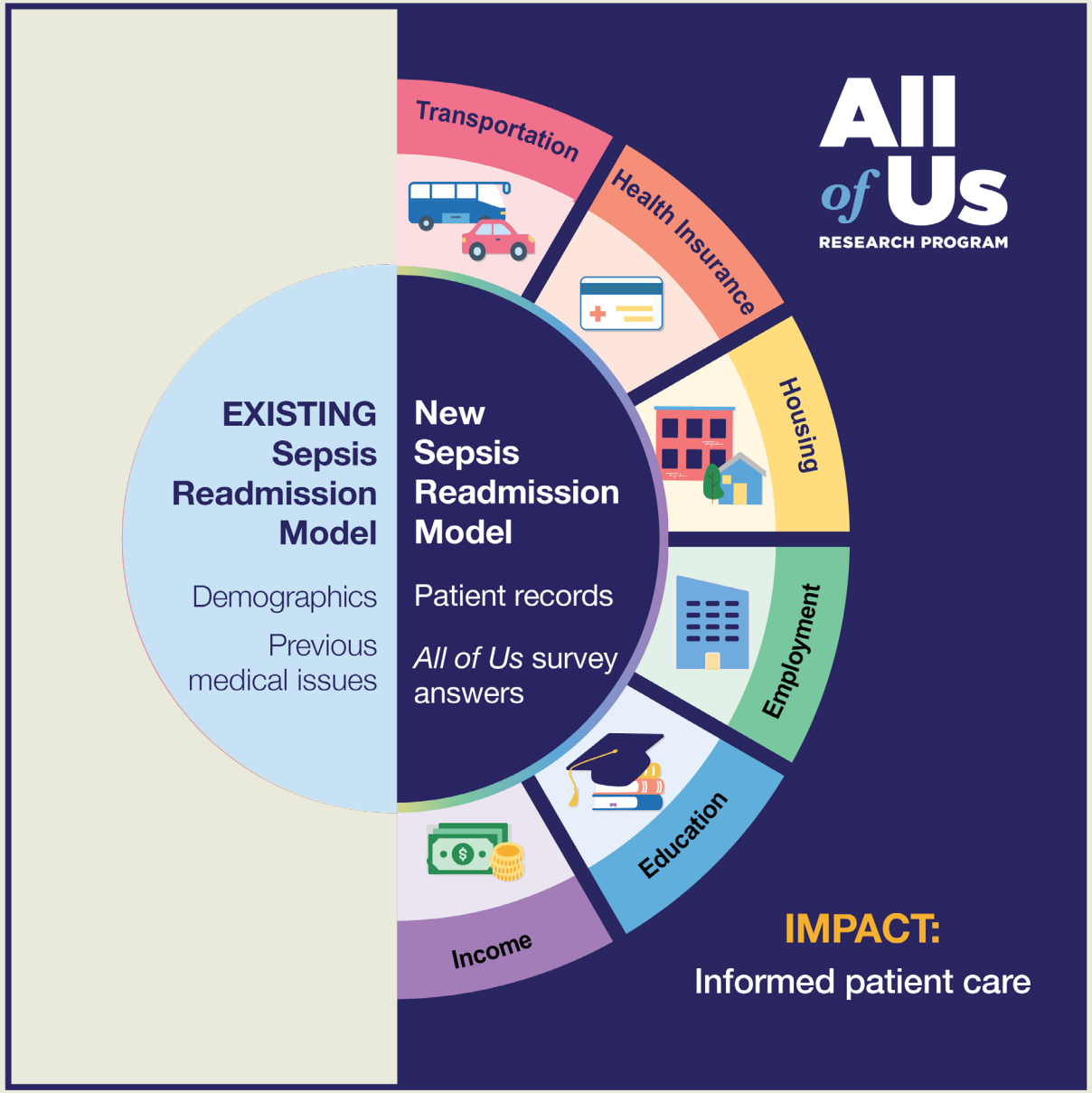
Lisa Lix
Professor & Canada Research Chair
University of Manitoba



AI4PH National Training Platform



- Short courses
- Scholarships
- Internships
- Community building
- Workshops





Rich Data can Yield Better Predictions

Reference: Amrollahi, F., et al. Inclusion of social determinants of health improves sepsis readmission prediction models. Journal of the American Medical Informatics Association, 2022; 29(7), 1263–1270.

Rich data often comes from linkage of diverse sources

Linkage allows information on an individual from one source to be linked to information on the same individual from another source

- Deterministic linkage
- Probabilistic linkage

Linkage is a cost-effective and efficient way to gather information



Source: https://www.popdata.bc.ca/datalinkage

Types of health data used in AI applications



Survey: cross-sectional or longitudinal self-report instruments that collects information related to health status, health care utilization and health determinants

> Examples: Canadian Community Health Survey, Canadian Health Measures Survey

Social media: contains social media users' behaviors; Includes content viewed, interactions, and content posted

> Examples: Twitter, Facebook, Reddit

Types of health data used in AI applications – 2



Administrative: for managing the healthcare system

Examples: hospital records, prescription medication records

Registry: for monitoring the care and treatment of individuals with specific health conditions

Examples: hip replacement registries, cancer registries

Electronic medical record (EMR): used by physicians to maintain health care information for inpatients or outpatients

Many of these data sources are secondary data, so the user cannot control how they are collected

Biases in secondary health data



- **Data-related biases**
  - Missing and incomplete values
  - Entry errors
- **Human-induced biases**
  - Implicit healthcare provider bias
  - Referral and admission bias: lack of representation of people who are medically underserved
  - Diagnosis bias
  - Risk disparities bias
- **Machine bias**
  - Equipment bias



Spot the biases

Social media data

- Population bias
- Methods bias (e.g., choice of keywords and hashtags)
- Analytic bias (e.g., choice of natural language processing methods)

Clinical registry data

- Population bias
- Completeness of data capture
- Machine-related bias in recordings of clinical measurements

ELSEVIER

International Journal of Medical Informatics

Volume 164, August 2022, 104804



Review article

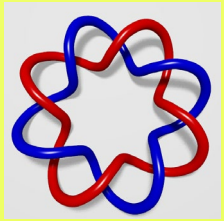
Biases in using social media data for public health surveillance: A scoping review

Yunpeng Zhao [a 1] ✉, Xing He [a 1] ✉, Zheng Feng [a] ✉, Sarah Bost [a], Mattia Prosperi [b] ✉, Yonghui Wu [a] ✉, Yi Guo [a] ✉, Jiang Bian [a] ✉

Show more

Considerations when using secondary health data

Linkage



Ownership



Quality



Access

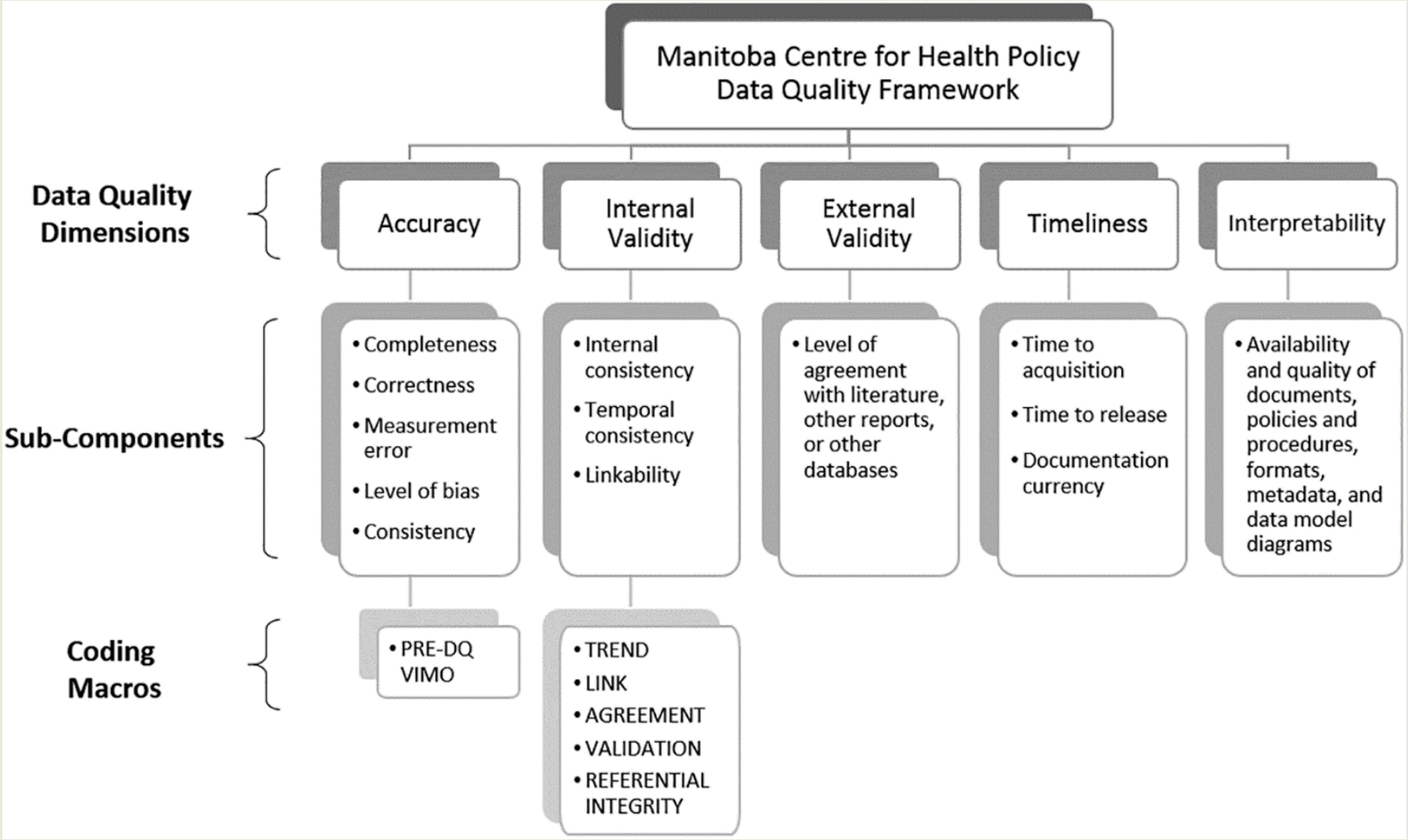


Consent



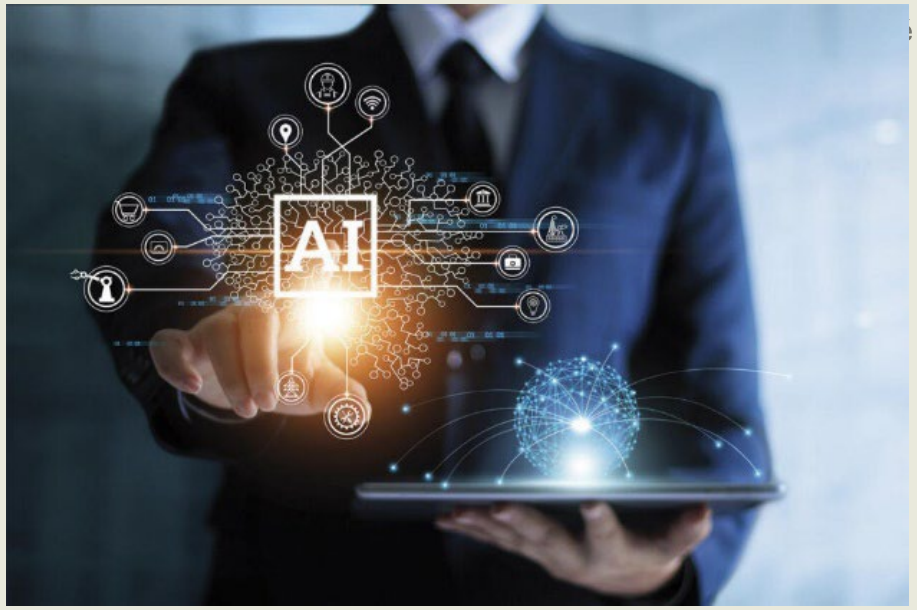
Security & Privacy





Source: Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. J Am Med Inform Assoc. 2017;25(3):224-229. doi:10.1093/jamia/ocx078

Synthetic data for privacy preservation (and more)



Defining synthetic data (Alan Turing Institute, UK):

- Data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)

Many types of synthetic data can be created:

- Tabular
- Time-series
- Text
- Images
- Video
- Audio

What are synthetic data used for?

- Preserving Data Privacy
- Developing and Training Models
- Augmenting Existing Datasets
- Supporting Training Activities

Benefits of synthetic data



Privacy preservation

- Reduce the risk of exposing real personal health information
- Synthetic data do not represent real people, only their characteristics

Better data availability and access

- Synthetic datasets can be freely shared and used without privacy restrictions that apply to real health data

Improved data for model training

- Allows for the creation of large-scale datasets that reflect diverse populations, rare diseases, or specific scenarios, that may be scarce in real-world data

Benefits of synthetic data – 2



Improve model training and validation

- A resource for training and validating machine learning models, especially for rare conditions or complex scenarios where real-world data is limited

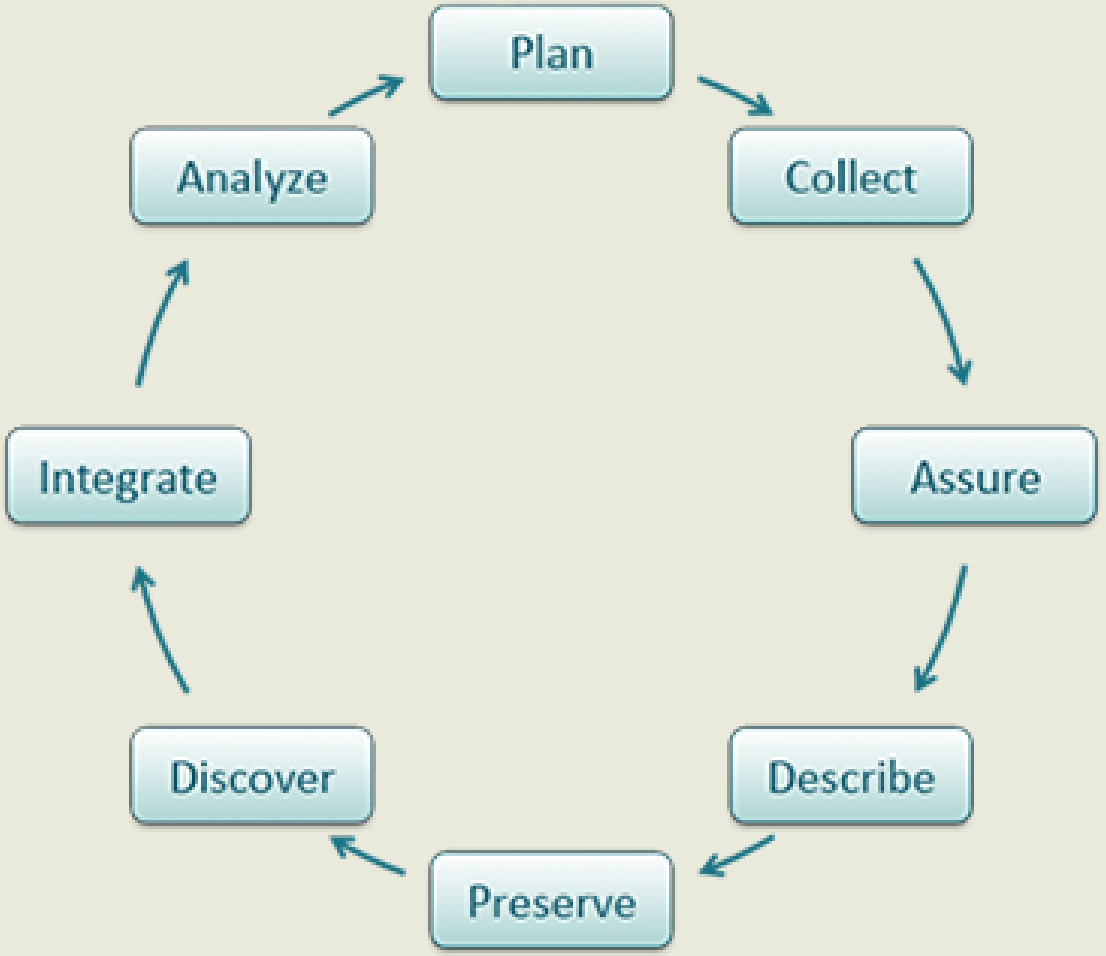
Address data biases

- Quality synthetic data can be created to mitigate biases in real-world data, which can lead to fairer and more equitable algorithms and applications

Accelerated innovation

- Faster data access can accelerate research, software creation, and testing new AI models

Data life cycle: Managing public health data



This framework describes the sequence of steps for

- collecting new data or extracting data from an existing repository,
- evaluating data quality,
- ensuring meta data and documentation are captured, and
- preparing data for analysis (e.g., pre-processing steps)

Source: https://www.dataone.org/data-life-cycle

Statistical models vs. machine learning: how to choose?

CONSIDER

Data complexity

- Unstructured or high-dimensional data may be better suited for machine learning models
- Traditional statistical models may suit simpler data

Data size and relatedness

- Lasso regression and tree-based machine-learning models are helpful for identifying the most “important” variables in a large set of potentially redundant variables

Data distribution

- Statistical models assume the data have specific characteristics, as a normal distribution, independence, constant variance, and linear relationships
- Machine learning models make minimal data distribution assumptions, but require large amounts of data to learn data patterns and relationships

Model choice: Key features

Feature	Statistical Models	Machine Learning Models (Unsupervised)
Core question	"Why does this happen?" (Inference)	“What will happen next?" (Prediction)
Underlying root	Applied mathematics / probability theory	Computer science / artificial intelligence
Interpretability	High (coefficients, p-values)	Variable or Low ("black box")
Data requirements	Smaller, cleaner datasets	Large, complex, or unstructured data
Common tools	R, SAS, classical stats packages	Python, TensorFlow, scikit-learn

KEY MESSAGES

Biases in secondary health data can influence the accuracy and generalizability of machine learning models, and ultimately the value of these models for decision making

Machine learning models do not make stringent assumptions about the data, but may not always be the optimal (or needed) choice. Consider the research question, as well as data complexity and distribution

Thank you for attending

KEEP IN TOUCH


ai4ph.dlsph@utoronto.ca


@ai4ph.bsky.social


/ai4ph-health-research-training-platform


AI4PH Slack workspace



AI4PH is supported by the Canadian Institutes of Health Research (CIHR).