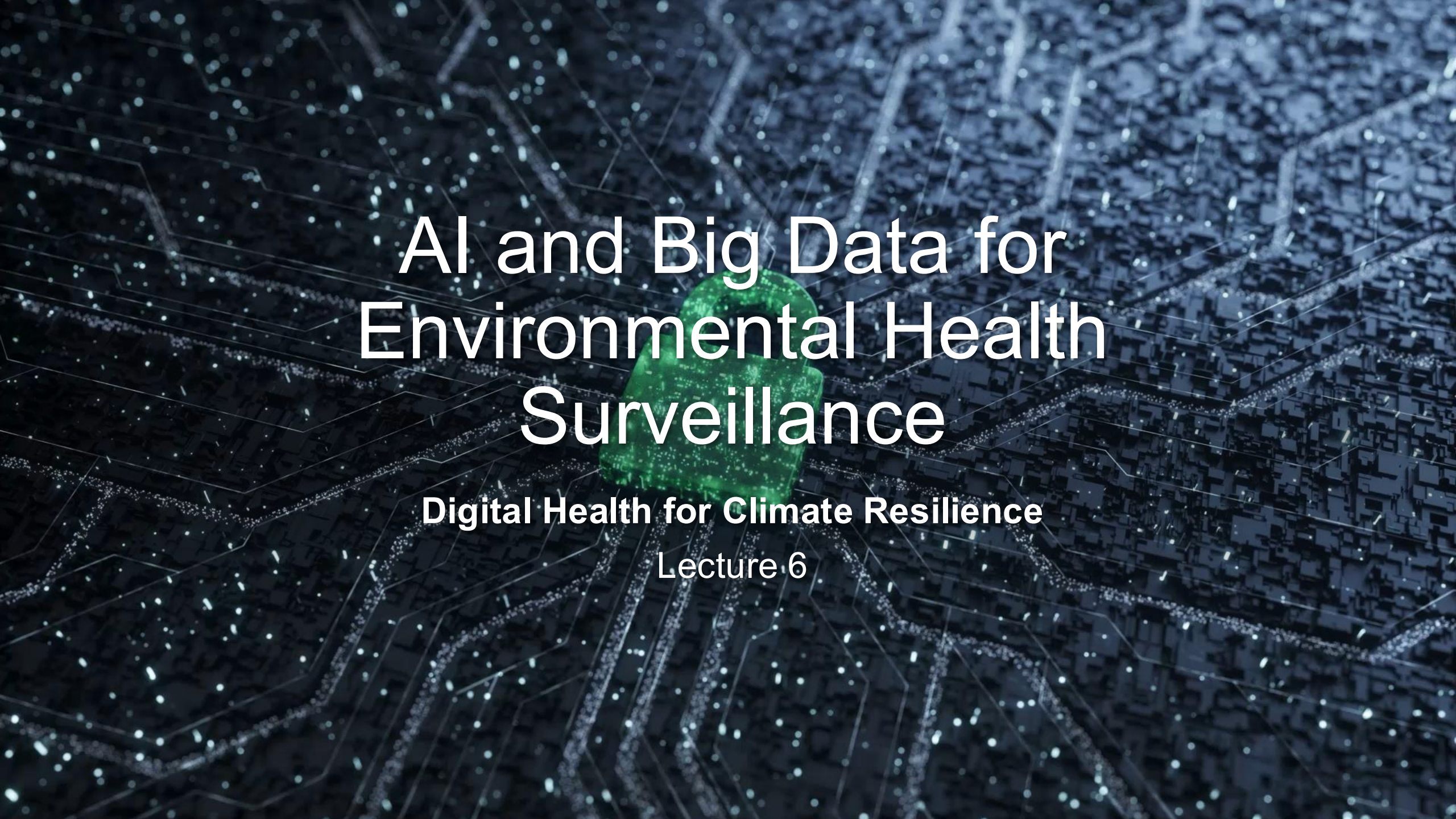

AI and Big Data for Environmental Health Surveillance

Digital Health for Climate Resilience

Lecture 6

Lecture Outline

- **Defining AI and Big Data in Environmental Health**
- **The Paradigm Shift:** From Reactive to Predictive Surveillance
- **The Data Revolution:** Sources of Big Data in Environmental Health
- **Foundational AI/ML Concepts for Surveillance**
- **Application 1:** Early Warning and Outbreak Prediction
- **Application 2:** Wastewater Surveillance Enhanced by AI
- **Application 3:** Exposomics and Integrating Health-Environment Data
- **Application 4:** Geospatial AI (GeoAI) for Exposure Assessment
- **Application 5:** Source Tracking and Attribution
- **Case Study 1:** Airborne Microbe Surveillance in Bangladesh
- **Case Study 2:** The MOSSAIC and EHRLICH Projects
- **Case Study 3:** AI_r Project in South Africa
- **Methodological and Ethical Challenges**
- **The Future of AI in Environmental Health**
- **Conclusion and Key Takeaways**

Defining AI and Big Data in Environmental Health

- **Big Data:** Extremely large, complex datasets that cannot be easily managed or analyzed with traditional data processing tools. Characterized by the "5 V's":
  - **Volume:** Massive amounts of data (terabytes to petabytes)
  - **Velocity:** Data generated and streaming in real-time
  - **Variety:** Diverse data types (structured, unstructured, text, images, sensor streams)
  - **Veracity:** Data quality and uncertainty
  - **Value:** The potential to extract meaningful insights
- **Artificial Intelligence (AI):** The simulation of human intelligence processes by computer systems
- **Machine Learning (ML):** A subset of AI involving algorithms that learn patterns from data without being explicitly programmed
- **Deep Learning:** A subset of ML using multi-layered neural networks, particularly powerful for complex data like images and text
- **Integration for Environmental Health Surveillance:** The use of AI/ML to analyze massive, diverse environmental and health datasets to detect patterns, predict risks, and enable proactive public health interventions

The Paradigm Shift: From Reactive to Predictive Surveillance

Traditional Surveillance

Reactive · Delayed · Limited sources

- Data from clinics & labs only — with significant time lags
- Descriptive statistics & simple analytical models
- Outbreaks detected after significant human impact
- Response is reactive — after the damage is done

VS

AI-Powered Surveillance

Proactive · Real-time · Multi-source

- Diverse real-time streams: satellites, sensors, EHRs, social media, search trends
- ML detects subtle patterns & anomalies invisible to humans
- Predicts outbreaks before they occur from precursor signals
- Enables proactive, targeted interventions — before harm spreads

★ **THE GOAL:** Shift the surveillance curve to the left — detect earlier, respond faster, prevent more.

EHR = Electronic Health Record | ML = Machine Learning | Proactive surveillance reduces outbreak impact by enabling earlier public health response

The Data Revolution: Sources of Big Data in Environmental Health

- **Satellite Remote Sensing:**
  - Aerosol Optical Depth (AOD) for PM2.5 estimation
  - Land Surface Temperature (LST) for heat mapping
  - Vegetation indices (NDVI) for vector habitat identification
  - Water quality (algal blooms, sediment)
- **Ground-Based Sensors:**
  - Reference-grade monitors (sparse, accurate) .
  - Low-cost sensor networks (dense, real-time, require calibration)
  - IoT-enabled weather stations
- **Mobile and Wearable Devices:**
  - GPS tracks for personal exposure assessment
  - Smartwatches for physiological responses (heart rate, activity)
  - Mobile apps for symptom reporting
- **Clinical and Health Data:**
  - Electronic Health Records (EHRs)
  - Disease registries (cancer, infectious diseases)
  - Syndromic surveillance (emergency department visits)
- **Digital Exhaust:**
  - Social media posts
  - Search engine queries (e.g., Google Flu Trends)
  - Mobility data from phones
- **Genomic and Exposomic Data:**
  - Pathogen sequencing
  - Biomarkers of exposure

Foundational AI/ML Concepts for Surveillance

- **Supervised Learning:** Learning a function that maps an input to an output based on labeled examples
  - *Application:* Predicting disease outbreaks (input: climate data; output: outbreak yes/no)
  - *Algorithms:* Random Forests, Gradient Boosting, Support Vector Machines, Neural Networks
- **Unsupervised Learning:** Finding patterns or structure in data without labeled outcomes
  - *Application:* Anomaly detection—identifying unusual patterns in sensor data that might indicate a chemical spill or emerging outbreak
  - *Algorithms:* Clustering (K-means), autoencoders
- **Natural Language Processing (NLP):** Enabling computers to understand and process human language
  - *Application:* Analyzing social media or news reports for early outbreak signals; extracting information from unstructured clinical notes
- **Computer Vision:** Enabling computers to interpret and understand visual information from images or videos
  - *Application:* Analyzing satellite or drone imagery to identify mosquito breeding habitats; counting people at risk during a disaster
- **Time Series Analysis:** Modeling data points indexed in time
  - *Application:* Forecasting air pollution levels or disease incidence based on historical patterns and current conditions

Application 1: Early Warning and Outbreak Prediction

- **Goal:** Predict infectious disease outbreaks before they occur, enabling proactive intervention
- **How It Works:**
  - Integrate diverse data streams: climate forecasts, satellite imagery (vegetation, water bodies), historical disease data, population mobility, and even social media trends
  - Train ML models to identify the complex, non-linear relationships between these drivers and subsequent disease incidence
  - Generate probabilistic forecasts of outbreak risk at fine spatial and temporal scales
- **Examples:**
  - **Dengue:** Models can predict outbreaks weeks to months in advance based on temperature, rainfall, and humidity, allowing vector control to be deployed preemptively
  - **Malaria:** Integrating rainfall, temperature, and vegetation data to predict high-risk areas and times for targeted interventions
  - **Cholera:** Using sea surface temperature and rainfall to predict cholera risk in coastal regions

Application 2: Wastewater Surveillance Enhanced by AI

- **Traditional Wastewater Surveillance:**
  - Requires prior knowledge of a pathogen's genetic sequence to design specific tests .
  - Can only detect what you are looking for
  - Typically confirms outbreaks after clinical cases have occurred
- **AI-Enhanced Wastewater Surveillance:**
  - Uses machine learning to analyze complex genomic data from wastewater samples
  - Can detect **emerging pathogens and novel variants without prior knowledge** of their genetic makeup
  - Identifies unique genomic signatures that signal the presence of a threat
  - Can potentially identify outbreaks **before the first patient enters a clinic**
- **Example:** A UNLV-led study published in *Nature Communications* used AI to detect influenza, RSV, mpox, and other pathogens in wastewater without prior knowledge of their sequences

Application 3: Exposomics and Integrating Health-Environment Data

- **The Challenge:** Understanding how environmental exposures across the life course affect health requires linking detailed exposure data with long-term health outcomes
- **Exposomics:** The comprehensive assessment of an individual's lifetime environmental exposures
- **AI-Powered Platforms:**
  - Integrate geocoded residential histories (where people have lived) with high-resolution environmental datasets (air pollution, water quality, land use, climate)
  - Link these exposure estimates to health records (cancer registries, EHRs)
  - Use ML to identify associations between specific exposures and health outcomes, accounting for complex mixtures and time lags
- **Example:** Dr. Heidi Hanson's **MOSSAIC and EHRLICH projects** at Oak Ridge National Laboratory
  - Created the **Centralized Health and Exposomic Resource (C-HER)**
  - Links geocoded residential histories from cancer registries with environmental exposure datasets
  - Enables large-scale investigations into how environmental factors affect cancer incidence, treatment response, and survival

Application 4: Geospatial AI (GeoAI) for Exposure Assessment

- **The Challenge:** Traditional exposure assessment often relies on residential address or a single monitor, missing the dynamic nature of where people actually spend their time
- **GeoAI:** The integration of geospatial technologies (GIS, remote sensing) with AI/ML
- **Key Advances:**
  - **High-Resolution Exposure Mapping:** ML models (e.g., Land Use Regression) can integrate satellite data, traffic patterns, land use, and sensor data to create detailed maps of pollution, noise, and heat at fine spatial scales (e.g., 100m)
  - **Personal Exposure Modeling:** Combining GPS tracks from mobile devices with high-resolution exposure maps to estimate individuals' dynamic exposures as they move through different microenvironments
  - **Activity Space Analysis:** Moving beyond residential address to consider the full range of locations where people spend time (home, work, commute, school)
- **Source:** A comprehensive narrative review in *Current Environmental Health Reports* identifies GeoAI and passive mobile data as two transformative shifts

Application 5: Source Tracking and Attribution

- **The Challenge:** When an environmental health threat is detected (e.g., a pathogen in wastewater, a chemical spill, an air pollution episode), we often need to identify the source to stop it
- **AI-Powered Source Tracking:**
  - In **sewer networks**, AI models combined with hydraulic modeling can trace a pathogen signal back to its likely origin (e.g., a specific neighborhood, building, or even hospital ward) based on flow patterns and sampling times
  - For **air pollution**, AI can perform source apportionment, analyzing the chemical composition of particulate matter to identify contributions from different sources (traffic, industry, biomass burning)
  - For **foodborne outbreaks**, AI can analyze genomic data from pathogens and integrate it with supply chain data to trace contaminated food back to its source
- **Example:** A study at Vrije Universiteit Amsterdam demonstrated that ML could back-trace pathogen sources in sewers using only outlet sampling, though it requires high-frequency data

Case Study 1: Airborne Microbe Surveillance in Bangladesh

- **Context:** Dhaka, Bangladesh, is a densely populated city with high burdens of respiratory infections and limited surveillance capacity
- **The Project:** Researchers created a combined dataset (2000-2023) integrating:
  - **Microbial air burdens:** Influenza A, SARS-CoV-2, Dengue virus RNA from air samples
  - **Environmental factors:** PM2.5, humidity, rainfall, temperature
  - **Health markers:** Hospitalizations for respiratory and febrile illness
  - **Socio-economic drivers:** Poultry outbreaks, rice harvest seasons, mobility shifts
- **AI Methods:** Deep learning architectures including autoencoders and convolutional neural networks (CNNs)
- **Key Findings:**
  - Elevated dengue RNA titers in air samples **preceded hospitalization peaks**
  - COVID-19 waves were associated with mobility shifts and PM2.5 exposure
  - The AI system compressed outbreak detection cycles from **weeks to days** with >85% predictive accuracy

Case Study 2: The MOSSAIC and EHRLICH Projects

- **Institution:** Oak Ridge National Laboratory (DOE), in collaboration with NIH and cancer centers
- **The Challenge:** Understanding how environmental factors affect cancer incidence, treatment, and survival requires linking massive, disparate datasets
- **The Solution:**
  - **MOSSAIC (Modeling Spatial and Spatiotemporal Analytics for Cancer)** : A platform for integrating and analyzing geospatial and environmental data
  - **EHRLICH (Environmental Health Research and Landscape Integrated Characterization)** : Focuses on linking environmental exposures to health outcomes
  - **C-HER (Centralized Health and Exposomic Resource)** : An AI-ready platform that unifies area-based and environmental exposure datasets with population health data
- **Data Integration:**
  - Geocoded residential histories from SEER cancer registries
  - High-resolution environmental datasets (air pollution, water quality, land use, climate)
  - Linked to clinical data on cancer treatment and survival
- **Impact:** Enables investigations into how environmental factors affect cancer across the life course, at scales previously impossible

Case Study 3: AI_r Project in South Africa

- **Context:** Gauteng province, South Africa—rapidly industrializing with significant air quality challenges and environmental justice concerns
- **The Problem:** Limited regulatory monitoring, high costs of reference-grade equipment, and communities disproportionately exposed to pollution but lacking data to advocate for change
- **The AI_r Solution:**
  - **Low-cost, locally manufactured air quality monitors** (cost reduced by factor of 2.5 to ~USD $100)
  - AI-powered models to predict pollution spikes and fill spatial gaps
  - Real-time public data dashboards accessible to communities
- **Dual Breakthrough:**
  - **Technological:** Dramatically reduced cost of monitoring, enabling dense community networks
  - **Social:** Empowered communities with real-time, actionable data for advocacy and personal protection (e.g., wearing masks during predicted spikes)
- **Impact:** Communities can now hold polluters accountable and make informed decisions to protect their health

Methodological and Ethical Challenges

- **Data Quality and Bias:**
  - ML models are only as good as their training data. Biased data (e.g., sensors concentrated in wealthier areas) leads to biased predictions
  - Low-cost sensors require rigorous calibration; uncalibrated data can be misleading
- **Privacy and Surveillance:**
  - Personal exposure monitoring and GPS tracking raise significant privacy concerns
  - Aggregated data can sometimes be re-identified
  - Need for robust data governance and ethical oversight
- **The "Black Box" Problem:**
  - Complex deep learning models can be difficult to interpret. Why did the model predict an outbreak? What features were most important?
  - Lack of interpretability can undermine trust and hinder public health action
- **Generalizability and Transferability:**
  - A model trained in one city or context may not work in another due to different environmental conditions, population characteristics, or health systems
- **The Digital Divide:**
  - The benefits of AI-powered surveillance may accrue primarily to wealthier, more connected populations and countries, exacerbating health inequities
- **Validation and Overfitting:**
  - ML models can easily overfit to training data, performing well in retrospect but failing in real-time deployment. Rigorous prospective validation is essential

The Future of AI in Environmental Health

- **Fusion of Data Streams:** Increasing integration of diverse data—genomic, exposomic, clinical, environmental, social—into unified platforms
- **Edge AI and Real-Time Processing:** Moving AI algorithms directly onto sensors and devices for real-time analysis and alerting, reducing data transmission need
- **Digital Twins:** Creating virtual replicas of cities or regions that simulate environmental processes and health impacts, allowing policymakers to test interventions before implementing them in the real world
- **Explainable AI (XAI):** Development of AI models that can explain their predictions in human-understandable terms, building trust and enabling action
- **Participatory AI:** Engaging communities in the design, deployment, and governance of AI systems to ensure they meet local needs and respect local values
- **Federated Learning:** Training AI models across multiple institutions without sharing raw data, addressing privacy concerns

Conclusion and Key Takeaways

- **AI and big data are transforming environmental health surveillance**, enabling a shift from reactive to proactive, predictive, and personalized approaches
- **Diverse data streams**—satellites, sensors, wearables, EHRs, genomics, digital exhaust—can be integrated to create a rich picture of environmental exposures and health outcomes
- **Key applications include** early outbreak prediction, AI-enhanced wastewater surveillance, exposomics, GeoAI for exposure assessment, and source tracking
- **Case studies from Bangladesh, the US, and South Africa** demonstrate the power and potential of these approaches in diverse settings
- **Significant challenges remain**—data quality, bias, privacy, interpretability, generalizability, and the digital divide—and must be addressed proactively
- **The future holds promise** with digital twins, explainable AI, federated learning, and participatory approaches
- **For public health professionals**, understanding these tools is no longer optional. We must become critical consumers and ethical stewards of AI and big data to ensure they serve the goal of health equity for all

Q&A / Discussion

Thank you.

Questions for Discussion:

- How can we ensure that AI-powered surveillance benefits low-resource communities and does not exacerbate the digital divide?
- What governance frameworks are needed to protect privacy while enabling public health benefits?
- How do we balance the power of predictive algorithms with the need for human judgment and community knowledge?
- In your context, what environmental health problem would you most want to tackle with AI and big data?

References

- Wilk-Jakubowski, J. L., Kuchcinski, A., Wilk-Jakubowski, G. K., Palej, A., & Pawlik, L. (2026). Digital Technologies and Machine Learning in Environmental Hazard Monitoring: A Synthesis of Evidence for Floods, Air Pollution, Earthquakes, and Fires. *Sensors*, *26*(3), 893. https://doi.org/10.3390/s26030893
- Rahman, M. T., Akib, A., Hasan, T., et al. (2025). AI Enhanced Real Time Monitoring of Airborne Microbes for Early Detection of Pandemics. *BURS 1st National Youth Research Summit 2025*, University of Barishal. http://www.aiub.edu/Files/student-research/AI_Enhanced_Real_Time_Monitoring_of_Airborne_Microbes_for_Early_Detection_of_Pandemics.html
- Hanson, H. A., et al. (2025). Centralized Health and Exposomic Resource (C-HER): Analytic and AI-Ready Data for External Exposomic Research. *arXiv:2511.03750*. https://arxiv.org/abs/2511.03750
- International Development Research Centre (IDRC). (2025). South African researchers tackle the global problem of poor air quality. https://idrc-crdi.ca/en/stories/south-african-researchers-tackle-global-problem-poor-air-quality