# Advanced Level Data Analysis Training Course

**DATE I** Feb 18, 19, 20, 2025
**VENUE I** Feb 18: 2F IRIS , Feb 19: 2F Camellia,
Feb 20: 2F Lotus, Grand InterContinental Seoul Parnas Hotel

CAPTURA
Capturing data on Antimicrobial resistance Patterns and Trends in Use in Regions of Asia

The Fleming Fund
Regional Grants

UKaid
from the British people

MOTT MACDONALD

International Vaccine Institute

# DATA ANALYTICS: EXPLORING GEOSPATIAL VARIABILITY AND PREDICTIVE MODELING WITH MACHINE LEARNING

**Ramesh Kumar Lama**

Research professor

Chosun University, South Korea

# OUTLINE

- Data Analytics
- Types of Data Analytics
- Machine Learning Problems
- Types of Machine Learning Methods
- Case Studies
  - Study1: Predicting Antimicrobial Resistance (AMR) Using Machine Learning
  - Study2: Geospatial Analysis in Antibiotic Resistance Prediction

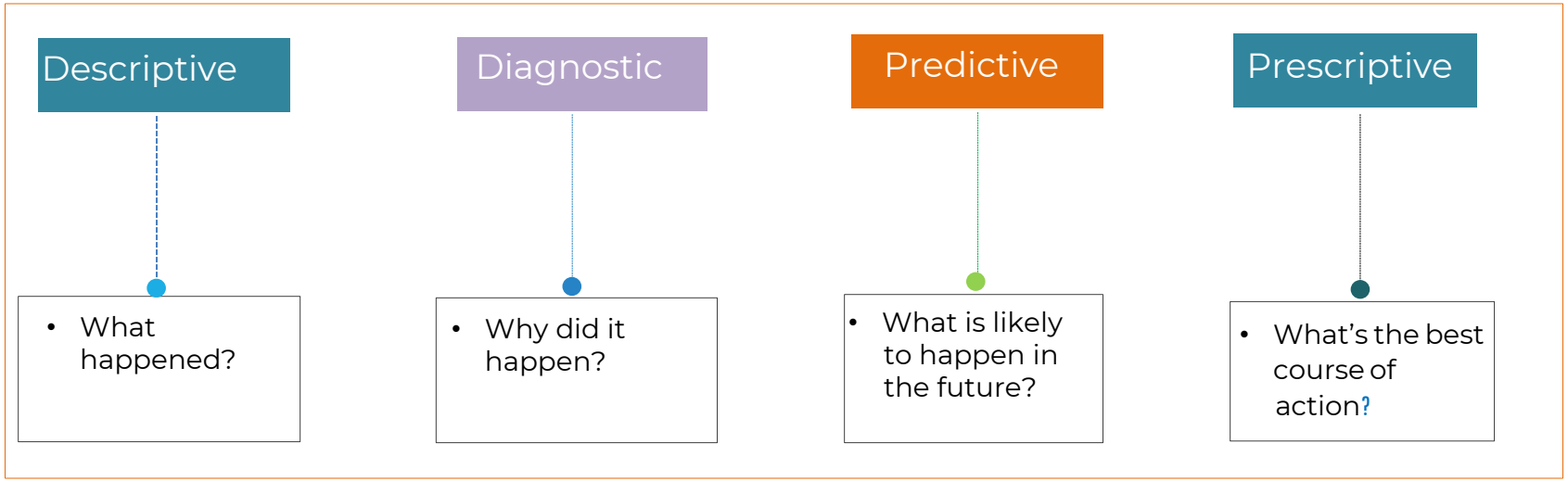
# DATA ANALYTICS

Science of examining raw data to reach certain conclusions

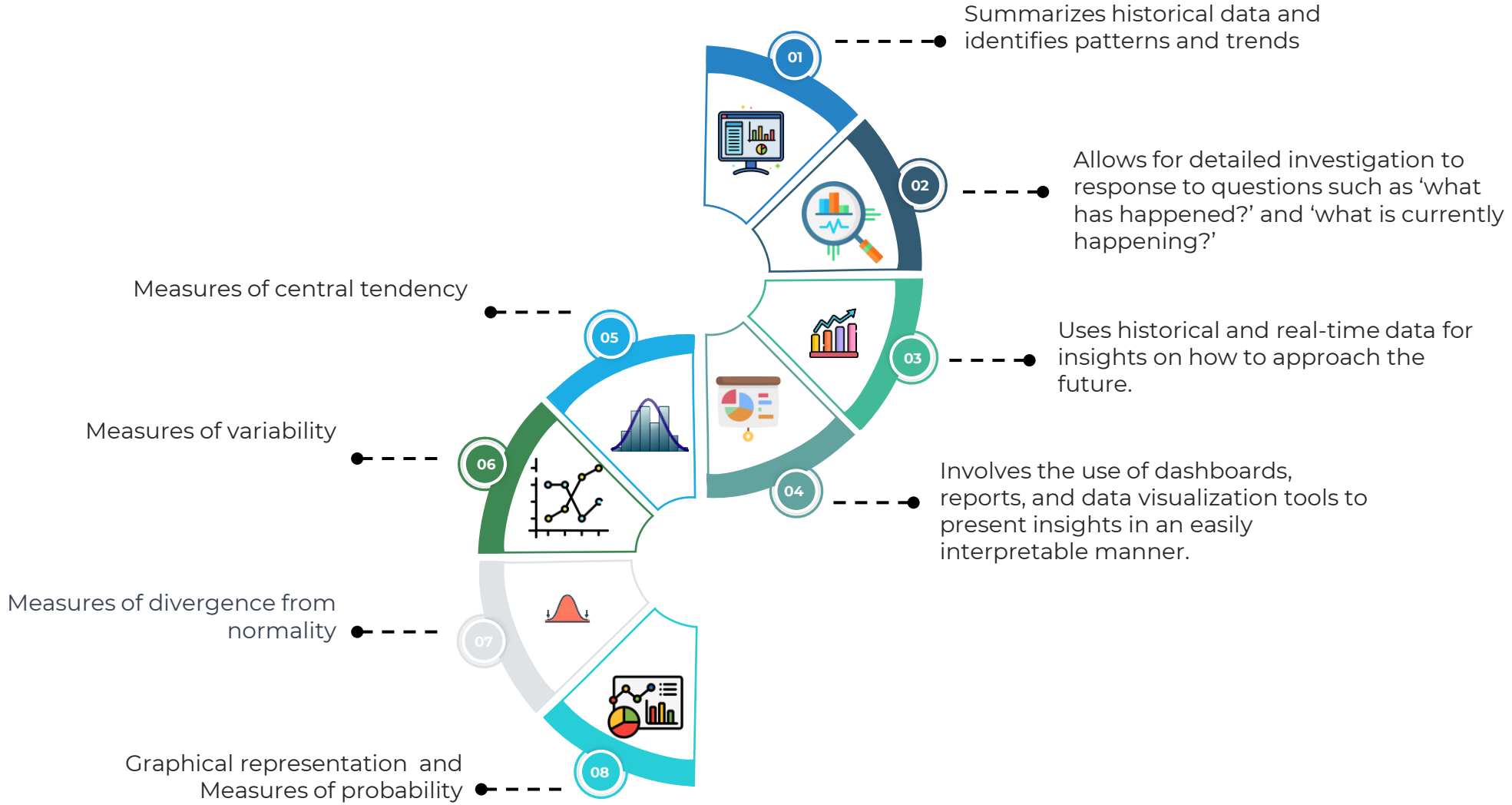
Involves examining data sets to gain insights or draw conclusions about what they contain, such as trends and predictions about future activity.

Involves examining the data to discover patterns, correlations, insights, and trends.

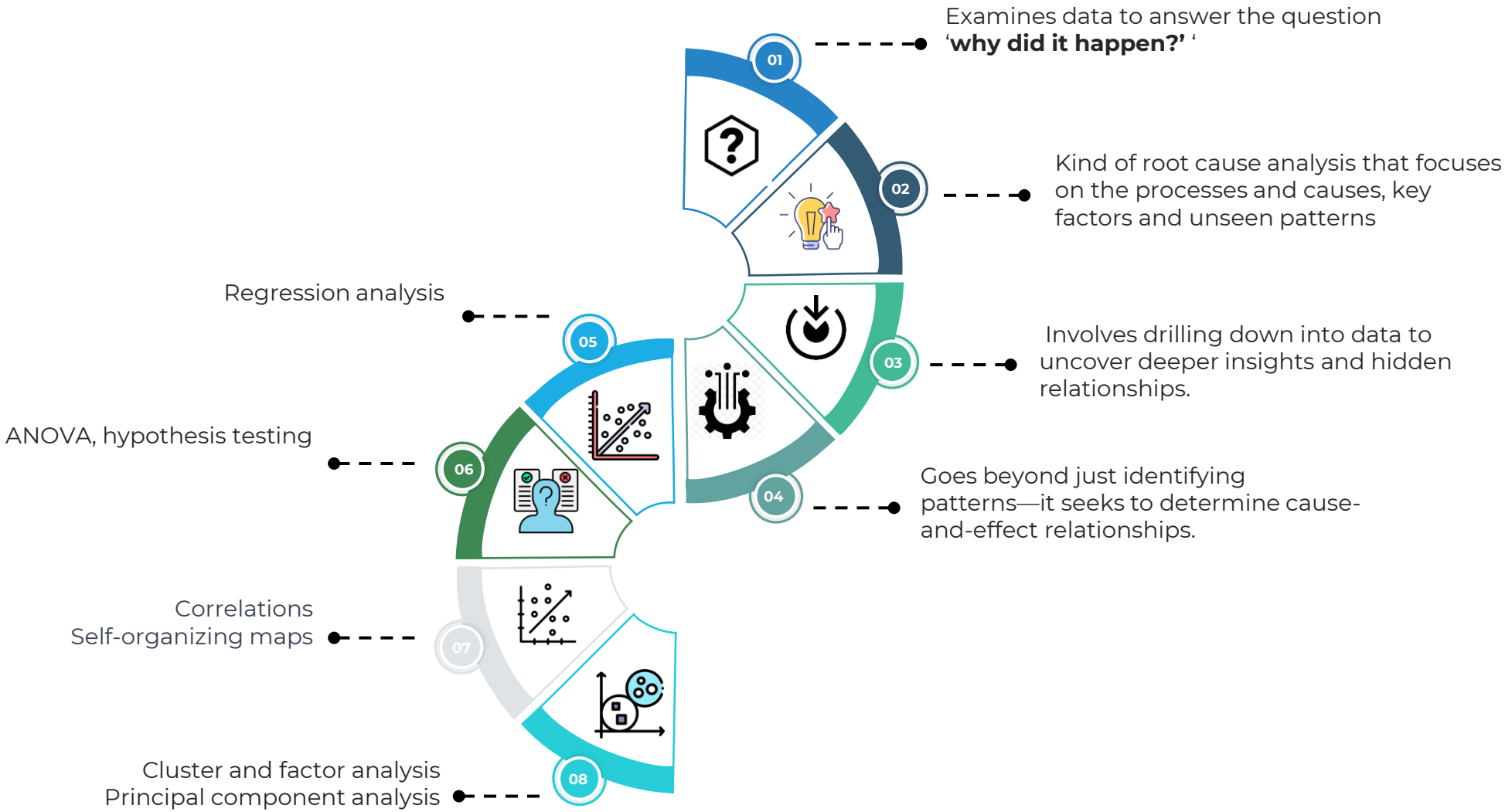
# TYPES OF DATA ANALYTICS



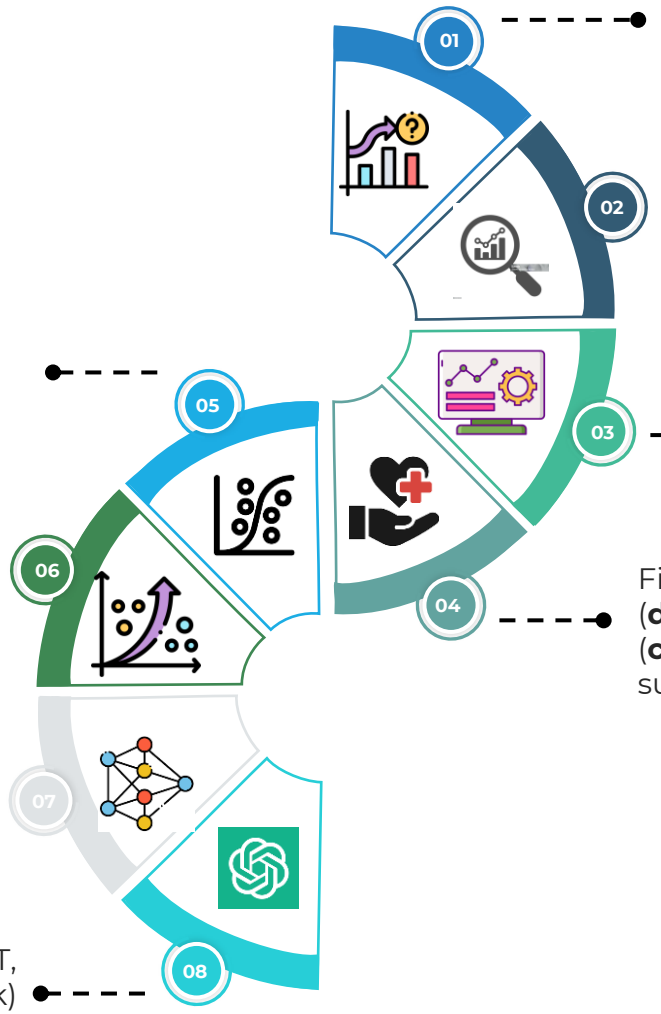
# DESCRIPTIVE ANALYTICS



# DIAGNOSTIC ANALYTICS



# PREDICTIVE ANALYTICS




Leverages historical and real-time data to make informed forecasts about future events, helping organizations anticipate trends and potential outcomes

Identifies patterns, correlations, and hidden relationships in data to estimate the likelihood of future scenarios.

Used to make data-driven decisions, optimize operations, and take proactive measures to mitigate risks or seize opportunities.

Finance (**fraud detection**), healthcare (**disease prediction**), marketing (**customer behavior analysis**), and supply chain (demand forecasting).

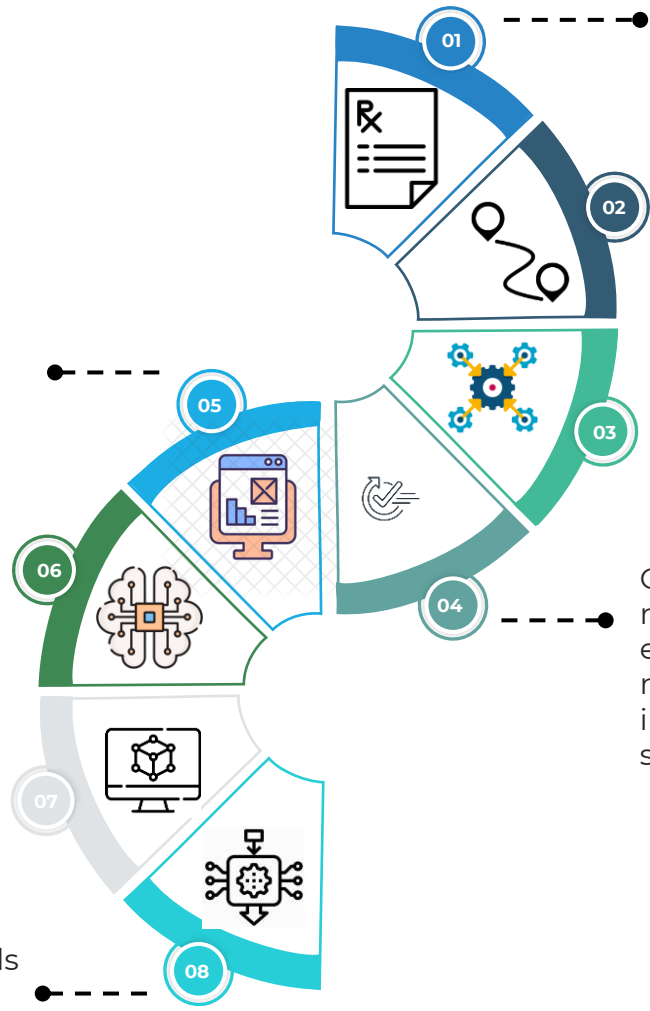
Logistic regression, Naive Bayes, Decision trees and random forest, collaborative filtering, clustering, and categorization.

Multivariate logistic regression, Kalman filtering, association rules, LASSO and ridge regression, conditional random fields (CRF)

Artificial Neural Networks (ANNs)

Transformers (e.g., BERT, GPT, DeepSeek)

# PRESCRIPTIVE ANALYTICS




Determines what actions to take to change undesirable trends.

Defined as deriving optimal planning decisions given the predicted future and addressing questions such as ‘what shall we do?’ and ‘why shall we do it?’

Integrates optimization techniques, simulation models, and reinforcement learning to recommend the best course of action, often automating decision-making processes..

Continuously adjusts recommendations based on new data, ensuring dynamic and adaptive responses to evolving conditions in industries like healthcare, finance, and supply chain management.

Optimization Algorithms.

Machine Learning

Simulation Techniques

Causal Inference Models

# KEY TAKEAWAYS

**Why machine Learning?**

### Traditional Methods for Descriptive & Diagnostic Analytics

- Traditional statistical methods are effective for descriptive and diagnostic analytics, helping summarize and understand past data.

### Machine Learning for Predictive & Prescriptive Analytics

- However, with the increasing volume and complexity of data, machine learning techniques offer greater accuracy and scalability for predictive and prescriptive analytics, enabling more informed decision-making and actionable insights

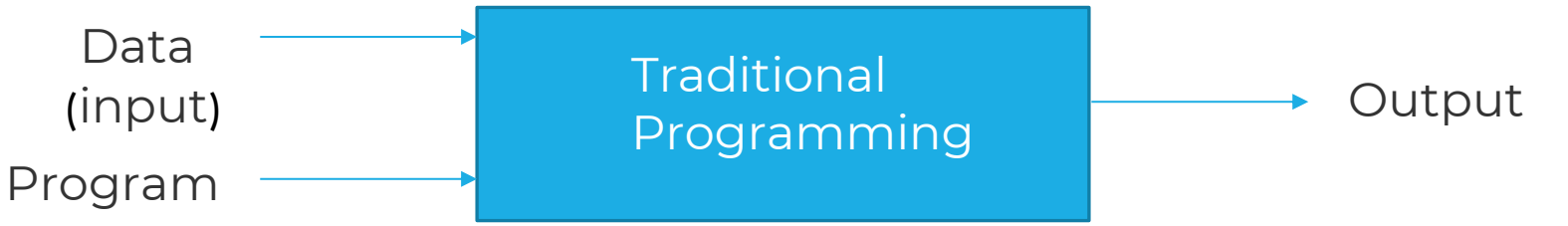
### Predictive Analytics for Future Insights

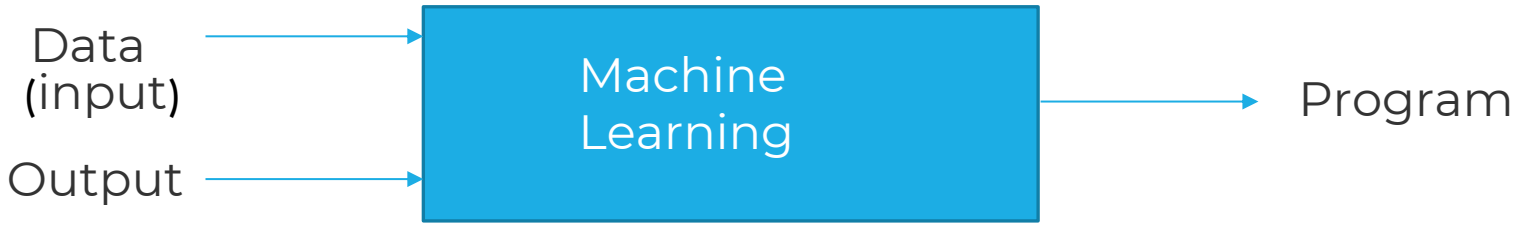
- Predictive analytics relies on machine learning techniques to forecast future trends, helping businesses anticipate customer behavior, market trends, and risks.

### Prescriptive Analytics for Actionable Recommendations

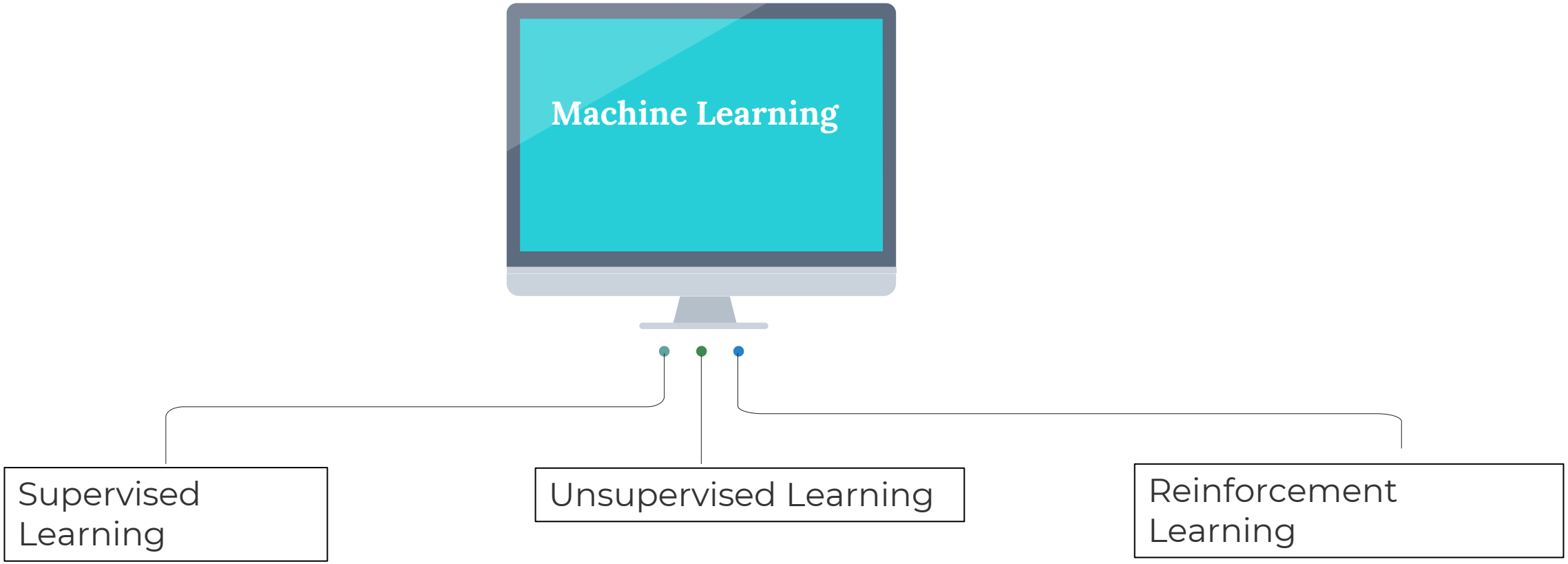
- Prescriptive analytics goes beyond prediction by recommending optimal actions using AI, optimization models, and reinforcement learning, enabling data-driven decision-making.`

# MACHINE LEARNING PROBLEMS



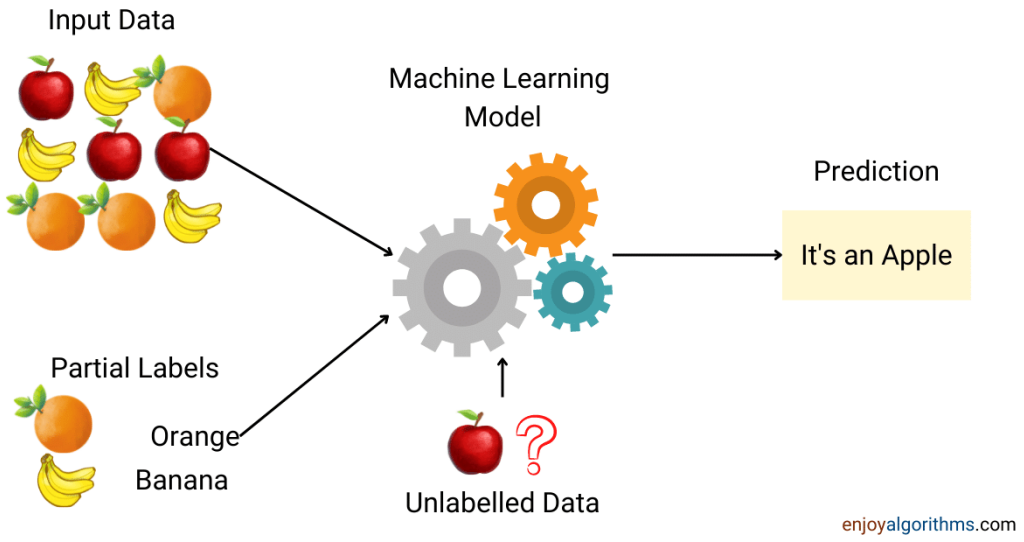



# MACHINE LEARNING PROBLEMS



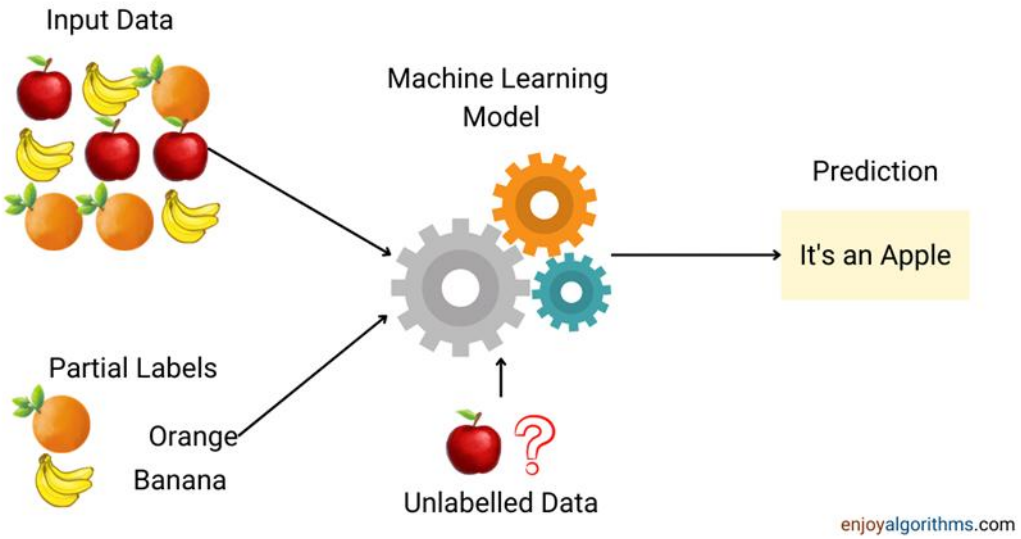
# SUPERVISED LEARNING

- Machines are trained using well “labeled” training data, and on the basis of that data, machines predict the output.
- The training data provided to the machines works as the supervisor that teaches the machines to predict the output correctly.
- It applies the same concept as a student learning under the supervision of the teacher.



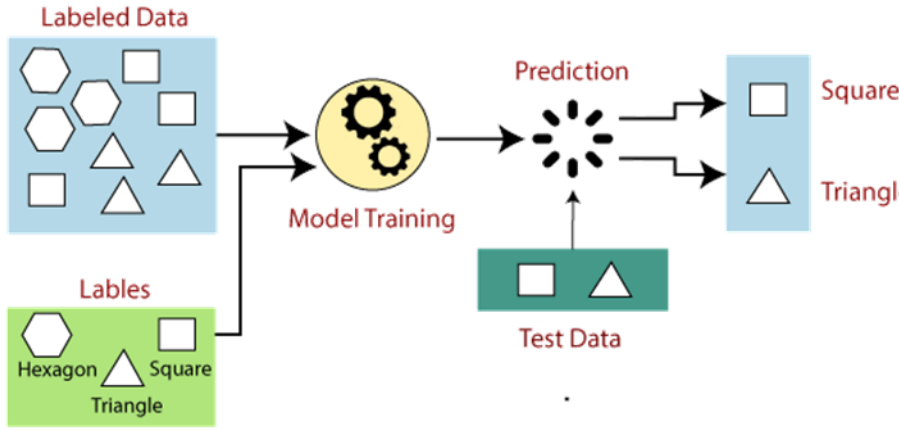
# FORMAL DEFINITION

- Supervised learning is a process of providing input data as well as correct output data to the machine learning model.
- A supervised learning algorithm aims to **find a mapping function to map the input variable(x) with the output variable(y)**.
- In the real world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.



# HOW SUPERVISED LEARNING WORKS ?

- Models are trained using labeled datasets, where the model learns about each type of data.
- Once the training process is completed, the model is tested based on test data
- Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon.
- Now the first step is that we need to train the model for each shape.
- If the given shape has four sides, and all the sides are equal, then it will be labeled as a **Square**.
- If the given shape has three sides, then it will be labeled as a triangle.
- If the given shape has six equal sides then it will be labeled as a **hexagon**.
- Now, after training, we test our model using the test set, and the task of the model is to identify the shape.
- The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the basis of a number of sides, and predicts the output.

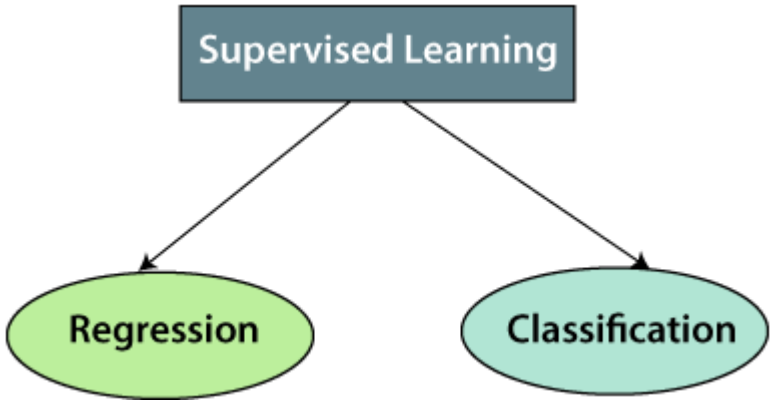


# STEPS INVOLVED IN SUPERVISED LEARNING

- Collect/Gather the labeled training data.
- Split the training dataset into a training dataset, test dataset, and validation dataset.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

# TYPES OF SUPERVISED LEARNING ALGORITHMS

Supervised learning can be further divided into two types of problems



# REGRESSION

- Regression algorithms are used if there is a relationship between the input variable and the output variable.
- It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.
    - Linear Regression
    - Regression Trees
    - Non-Linear Regression
    - Bayesian Linear Regression
    - Polynomial Regression

# CLASSIFICATION

- Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc. Spam Filtering,
  - •Random Forest
  - •Decision Trees
  - •Logistic Regression
  - •Support Vector Machines

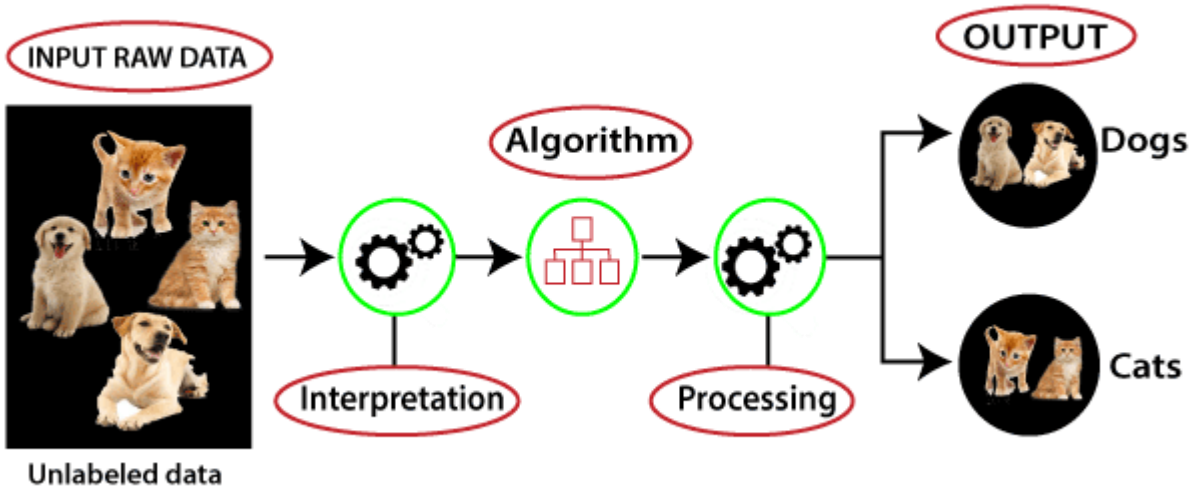
# UNSUPERVISED MACHINE LEARNING

- Models are trained using unlabeled dataset and are allowed to act on that data without any supervision.
- The goal of unsupervised learning is to find the underlying structure of the dataset, group that data according to similarities, and represent that dataset in a compressed format.

# WHY UNSUPERVISED LEARNING?

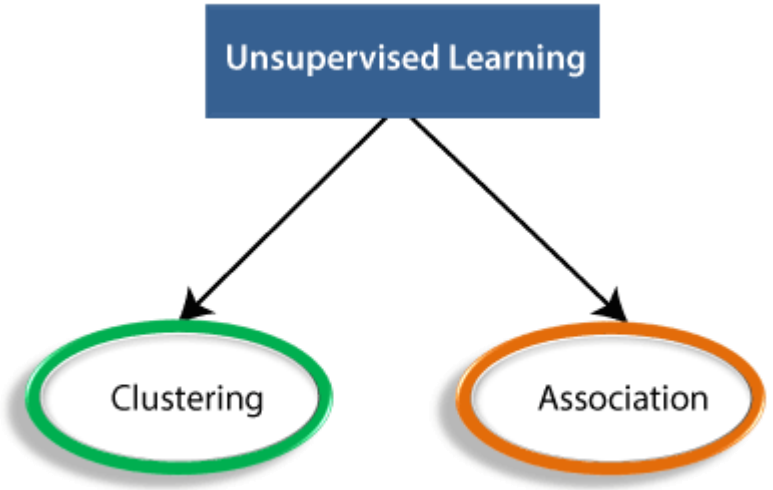
- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

# HOW UNSUPERVISED LEARNING WORKS?




- Here, we have taken unlabeled input data, which means it is not categorized and corresponding outputs are also not given.
- Now, this unlabeled input data is fed to the machine learning model to train it.
- Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.
- Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and differences between the objects.

# TYPES OF UNSUPERVISED LEARNING ALGORITHM




**Clustering:**

- Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and have less or no similarities with the objects of another group.
- Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

**Association:**

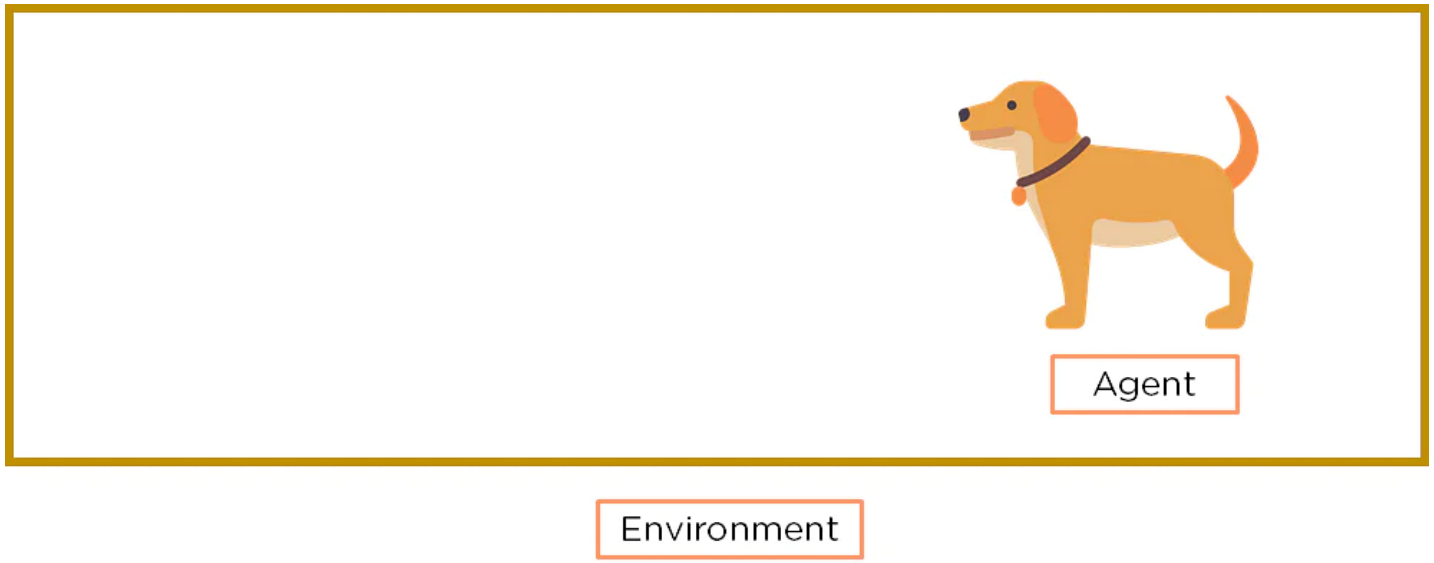
- An association rule is an unsupervised learning method that is used for finding the relationships between variables in a large database.
- It determines the set of items that occur together in the dataset.
- Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) items.
- A typical example of Association rule is Market Basket Analysis.

# REINFORCEMENT LEARNING

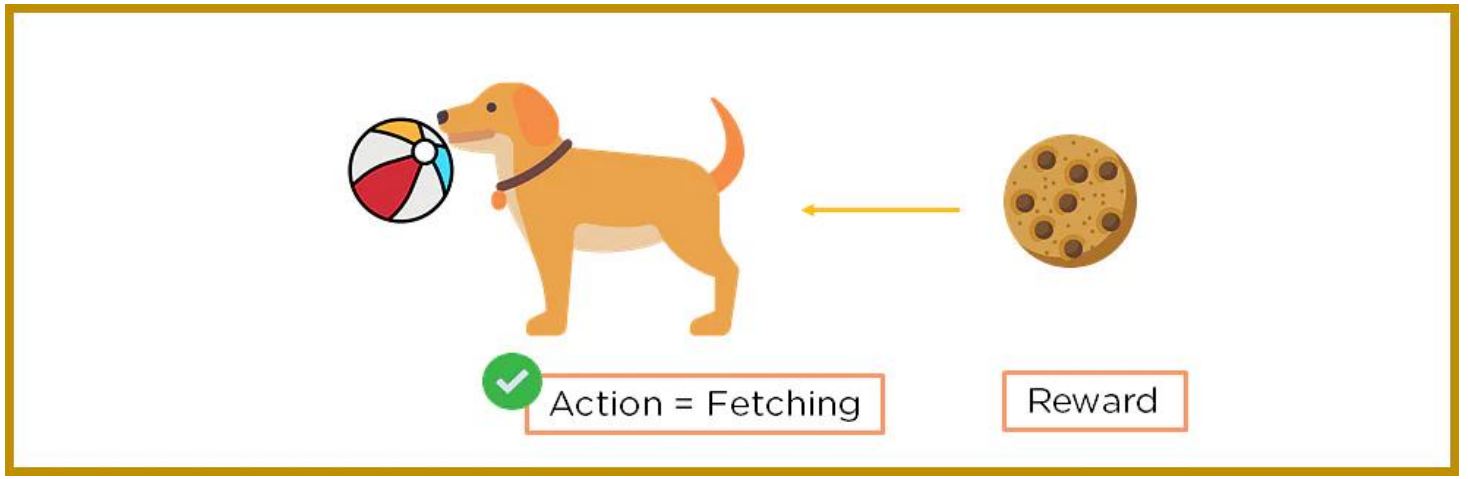
- More general than supervised/unsupervised learning
- It is about learning the optimal behavior in an environment to obtain maximum reward.
- This optimal behavior is learned through interactions with the environment and observations of how it responds
- We model an environment after the problem statement.
- The model interacts with this environment and comes up with solutions all on its own, without human interference.
- To push it in the right direction, we simply give it a positive reward if it performs an action that brings it closer to its goal or a negative reward if it goes away from its goal.
- Similar to children exploring the world around them and learning the actions that help them achieve a goal.

# REINFORCEMENT LEARNING

- To understand reinforcement learning better, consider a dog that we have to house-train.
- Here, the dog is the agent, and the house is the environment.



# REINFORCEMENT LEARNING




- We can get the dog to perform various actions by offering incentives such as dog biscuits as a reward.
- The dog will follow a policy to maximize its reward and hence will follow every command and might even learn a new action, like begging, all by itself.

# REINFORCEMENT LEARNING

# REINFORCEMENT LEARNING

- The dog will also want to run around and play and explore its environment.
- This quality of a model is called Exploration.
- The tendency of the dog to maximize rewards is called Exploitation.
- There is always a tradeoff between exploration and exploitation, as exploration actions may lead to lesser rewards.

# REINFORCEMENT LEARNING

- In the absence of a supervisor, the learner must independently discover the sequence of actions that maximize the reward.
- This discovery process is similar to a trial-and-error search.
- The quality of actions is measured by not just the immediate reward they return, but also the delayed reward they might fetch.
- As it can learn the actions that result in eventual success in an unseen environment without the help of a supervisor, reinforcement learning is a very powerful algorithm.

# CASE STUDIES

**Study 1:** Predicting Antimicrobial Resistance (AMR) Using Machine Learning
**Study 2:** Geospatial Analysis in Antibiotic Resistance Prediction

# STUDY1 : PREDICTING ANTIMICROBIAL RESISTANCE (AMR) USING MACHINE LEARNING

**Karina-Doris Vihta** – University of Oxford

**Publication Details:**
•**Journal:** Communications Medicine
•**Publication Year:** 2024

# STUDY GOAL

- Addresses the critical global health issue of antimicrobial resistance (AMR).
- The primary goal is to enhance the prediction of AMR trends at a nationwide, aggregate hospital level, thereby facilitating targeted interventions.
- By leveraging machine learning techniques, the researchers aim to utilize historical data on AMR and antimicrobial usage to forecast future resistance patterns

# METHODS

**Data Collection:**

- The study gathered data on antimicrobial use and AMR prevalence from bloodstream infections across hospitals in England.
- This data was organized by hospital group (referred to as Trust) and financial year (April–March) for 22 specific pathogen–antibiotic combinations, covering the period from FY2016-2017 to FY2021-2022.

**Modeling Approach:**

- The researchers employed the Extreme Gradient Boosting (XGBoost) machine learning model to predict future AMR prevalence.
- They compared the performance of XGBoost predictions against traditional forecasting methods, including: Carrying forward the previous year's value.
- Projecting the difference between the two preceding years into the future.
- Applying linear trend forecasting (LTF).

**Feature Importance Analysis:**

- To enhance the interpretability of the model, the study calculated feature importances within the XGBoost framework.

# DATA ANALYTICS & AMR PREDICTION

- Data analytics helps in understanding AMR trends through:
- Descriptive Analytics
  - Summarizing historical AMR data.
- Diagnostic Analytics
  - Analyzing why resistance trends are changing.
- Predictive Analytics
  - Forecasting future AMR levels using ML models.
- Prescriptive Analytics
  - Recommending optimal antibiotic policies to control AMR.

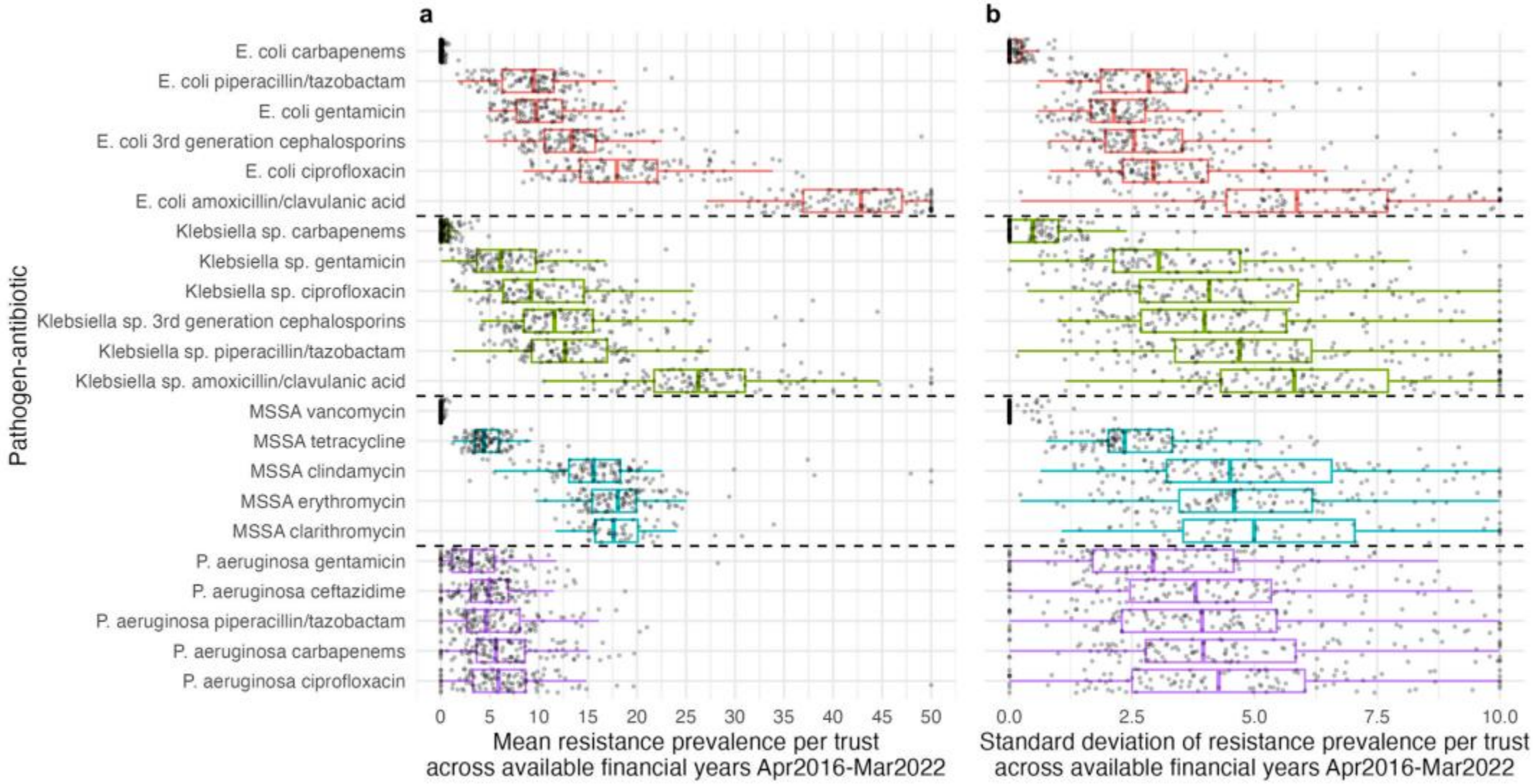
# MACHINE LEARNING METHODS USED

The study compares different forecasting methods:

- XGBoost (Extreme Gradient Boosting - Captures complex relationships.
- Previous Value Carried Forward - Uses last recorded AMR value as a prediction.
- Linear Trend Forecasting (LTF) - Uses historical trends for estimation.
- Difference of Previous Two Years Taken Forward - Uses recent trends for projection.

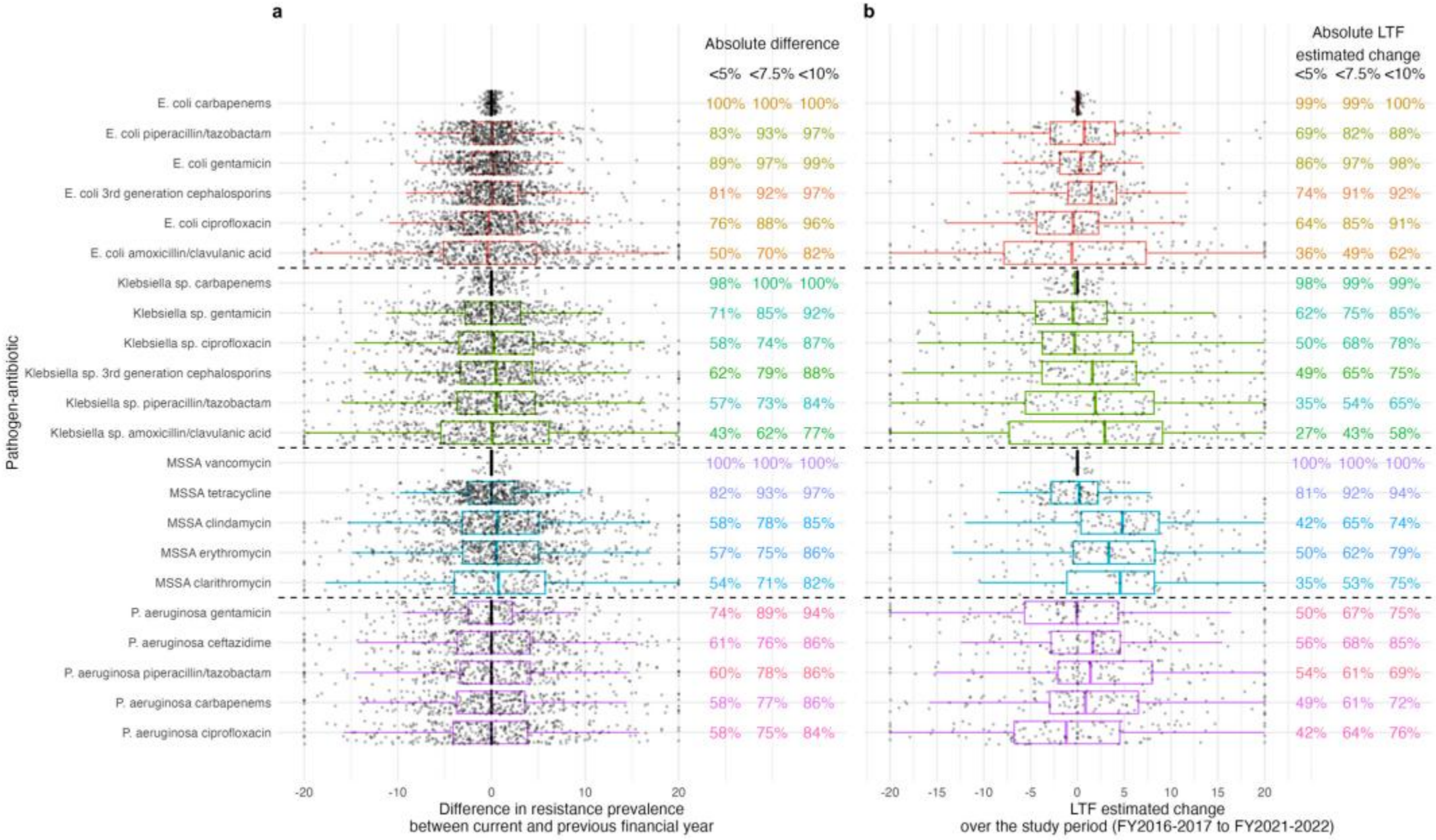
Result: XGBoost outperforms all methods, especially in hospitals with fluctuating AMR levels.

# RESULTS

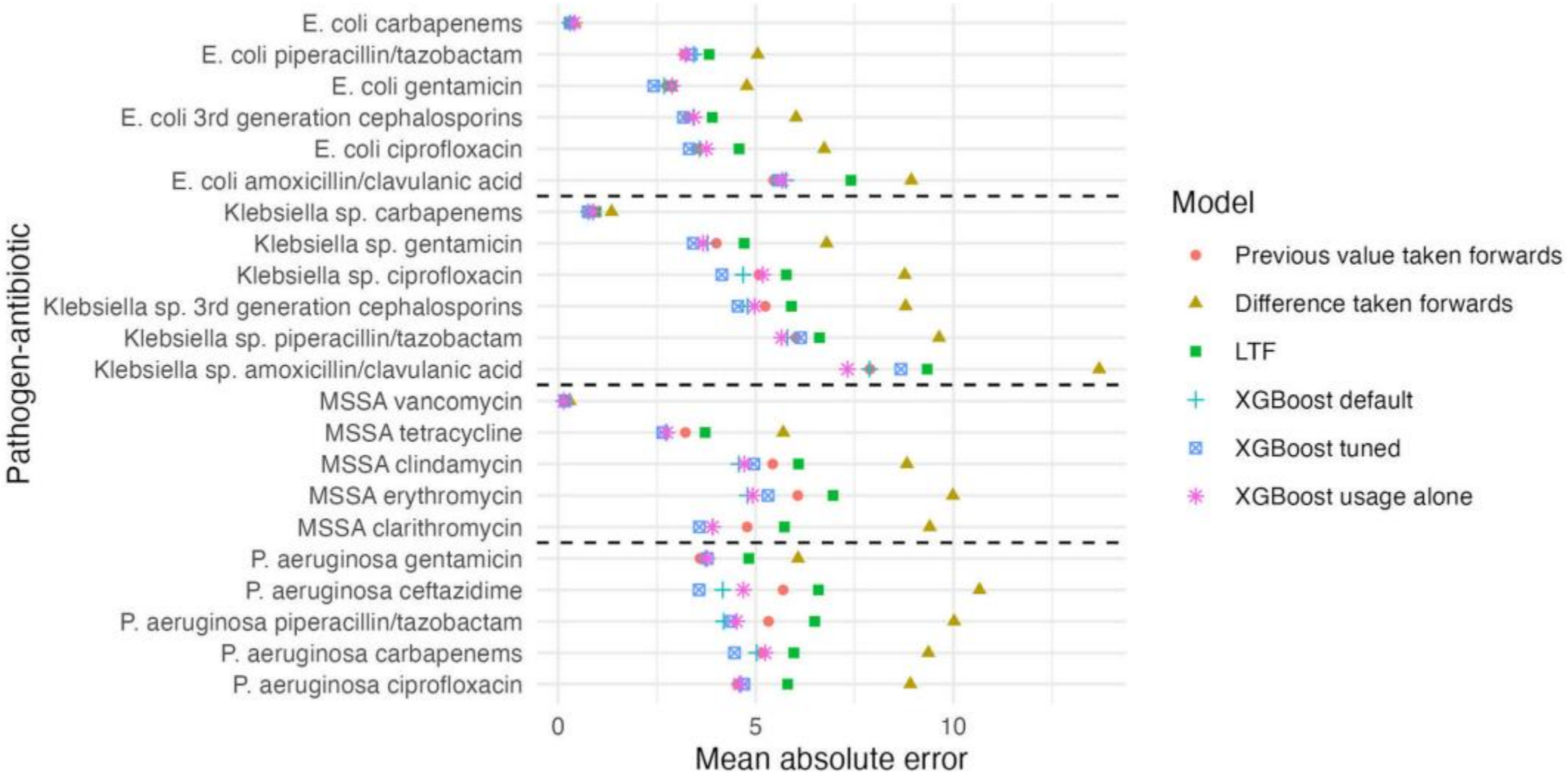



Mean resistance prevalence and standard deviation per Trust–pathogen–antibiotic

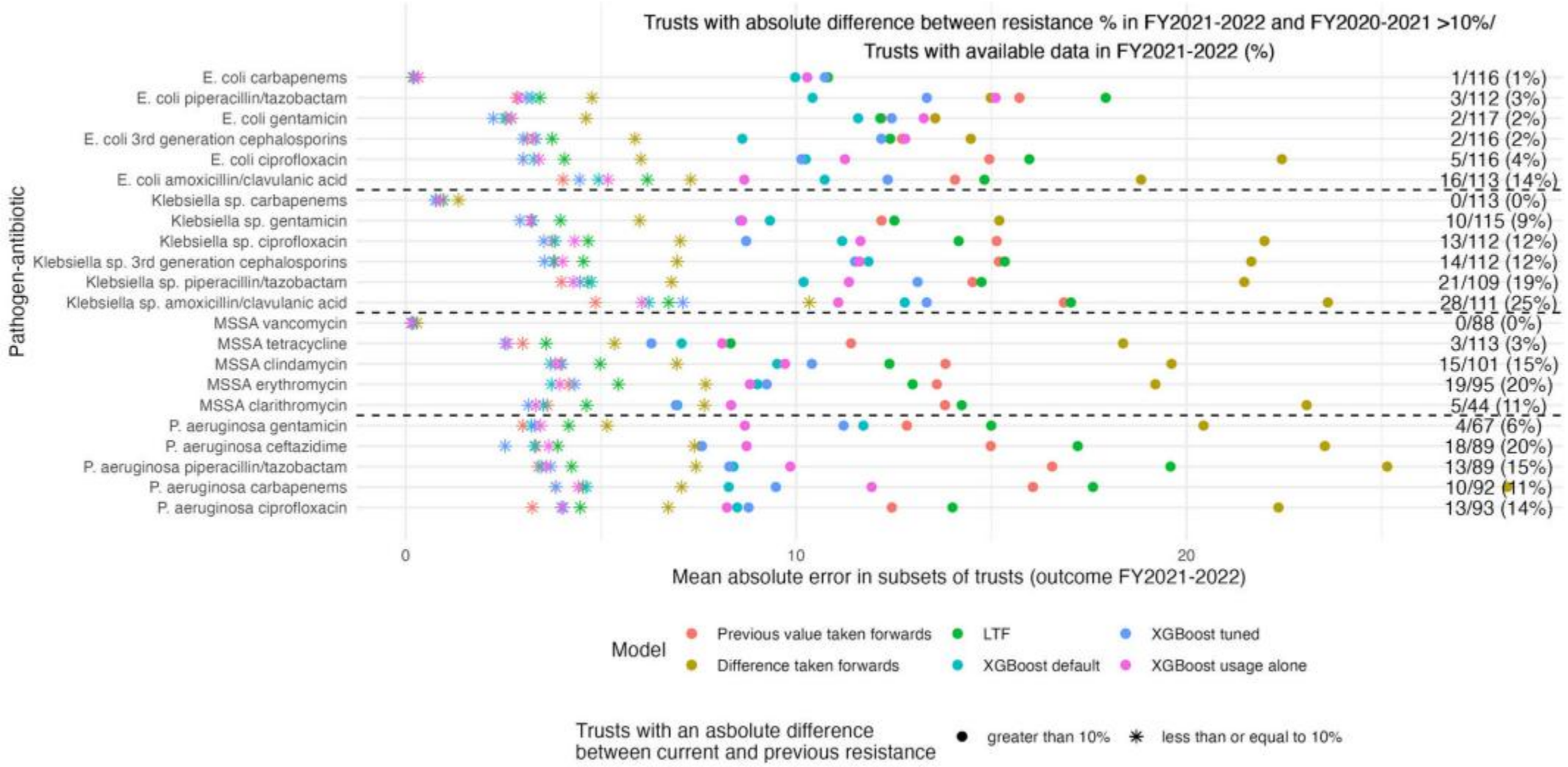
# RESULTS


a
b
Absolute difference
<5% <7.5% <10%
Absolute LTF estimated change
<5% <7.5% <10%
Pathogen-antibiotic
E. coli carbapenems 100% 100% 100% 99% 99% 100%
E. coli piperacillin/tazobactam 83% 93% 97% 69% 82% 88%
E. coli gentamicin 89% 97% 99% 86% 97% 98%
E. coli 3rd generation cephalosporins 81% 92% 97% 74% 91% 92%
E. coli ciprofloxacin 76% 88% 96% 64% 85% 91%
E. coli amoxicillin/clavulanic acid 50% 70% 82% 36% 49% 62%
Klebsiella sp. carbapenems 98% 100% 100% 98% 99% 99%
Klebsiella sp. gentamicin 71% 85% 92% 62% 75% 85%
Klebsiella sp. ciprofloxacin 58% 74% 87% 50% 68% 78%
Klebsiella sp. 3rd generation cephalosporins 62% 79% 88% 49% 65% 75%
Klebsiella sp. piperacillin/tazobactam 57% 73% 84% 35% 54% 65%
Klebsiella sp. amoxicillin/clavulanic acid 43% 62% 77% 27% 43% 58%
MSSA vancomycin 100% 100% 100% 100% 100% 100%
MSSA tetracycline 82% 93% 97% 81% 92% 94%
MSSA clindamycin 58% 78% 85% 42% 65% 74%
MSSA erythromycin 57% 75% 86% 50% 62% 79%
MSSA clarithromycin 54% 71% 82% 35% 53% 75%
P. aeruginosa gentamicin 74% 89% 94% 50% 67% 75%
P. aeruginosa ceftazidime 61% 76% 86% 56% 68% 85%
P. aeruginosa piperacillin/tazobactam 60% 78% 86% 54% 61% 69%
P. aeruginosa carbapenems 58% 77% 86% 49% 61% 72%
P. aeruginosa ciprofloxacin 58% 75% 84% 42% 64% 76%
-20 -10 0 10 20
Difference in resistance prevalence between current and previous financial year
-20 -15 -10 -5 0 5 10 15 20
LTF estimated change over the study period (FY2016-2017 to FY2021-2022)

# RESULTS


Pathogen-antibiotic
E. coli carbapenems
E. coli piperacillin/tazobactam
E. coli gentamicin
E. coli 3rd generation cephalosporins
E. coli ciprofloxacin
E. coli amoxicillin/clavulanic acid
Klebsiella sp. carbapenems
Klebsiella sp. gentamicin
Klebsiella sp. ciprofloxacin
Klebsiella sp. 3rd generation cephalosporins
Klebsiella sp. piperacillin/tazobactam
Klebsiella sp. amoxicillin/clavulanic acid
MSSA vancomycin
MSSA tetracycline
MSSA clindamycin
MSSA erythromycin
MSSA clarithromycin
P. aeruginosa gentamicin
P. aeruginosa ceftazidime
P. aeruginosa piperacillin/tazobactam
P. aeruginosa carbapenems
P. aeruginosa ciprofloxacin
0
5
10
Mean absolute error
Model
Previous value taken forwards
Difference taken forwards
LTF
XGBoost default
XGBoost tuned
XGBoost usage alone

# RESULTS


Trusts with absolute difference between resistance % in FY2021-2022 and FY2020-2021 >10%/
Trusts with available data in FY2021-2022 (%)
Pathogen-antibiotic
E. coli carbapenems
E. coli piperacillin/tazobactam
E. coli gentamicin
E. coli 3rd generation cephalosporins
E. coli ciprofloxacin
E. coli amoxicillin/clavulanic acid
Klebsiella sp. carbapenems
Klebsiella sp. gentamicin
Klebsiella sp. ciprofloxacin
Klebsiella sp. 3rd generation cephalosporins
Klebsiella sp. piperacillin/tazobactam
Klebsiella sp. amoxicillin/clavulanic acid
MSSA vancomycin
MSSA tetracycline
MSSA clindamycin
MSSA erythromycin
MSSA clarithromycin
P. aeruginosa gentamicin
P. aeruginosa ceftazidime
P. aeruginosa piperacillin/tazobactam
P. aeruginosa carbapenems
P. aeruginosa ciprofloxacin
1/116 (1%)
3/112 (3%)
2/117 (2%)
2/116 (2%)
5/116 (4%)
16/113 (14%)
0/113 (0%)
10/115 (9%)
13/112 (12%)
14/112 (12%)
21/109 (19%)
28/111 (25%)
0/88 (0%)
3/113 (3%)
15/101 (15%)
19/95 (20%)
5/44 (11%)
4/67 (6%)
18/89 (20%)
13/89 (15%)
10/92 (11%)
13/93 (14%)
0
10
20
Mean absolute error in subsets of trusts (outcome FY2021-2022)
Model
Previous value taken forwards
LTF
XGBoost tuned
Difference taken forwards
XGBoost default
XGBoost usage alone
Trusts with an asbolute difference between current and previous resistance
greater than 10%
less than or equal to 10%

# KEY FINDINGS OF THE STUDY

- Minimal year-to-Year AMR Variability in most hospitals.
- XGBoost excels in cases with significant AMR changes.
- Feature Importance Analysis (SHAP values) reveals:
- Historical resistance to the same pathogen-antibiotic combination is the strongest predictor.
- Resistance patterns in different pathogens using the same antibiotics are significant.
- Antibiotic consumption patterns play a crucial role in future resistance.

# STUDY 2: GEOSPATIAL ANALYSIS IN ANTIBIOTIC RESISTANCE PREDICTION

**Laurel Legenza**, PharmD, PhD : University of Wisconsin–Madison, School of Pharmacy

**Publication Details:**
•**Journal:** Scientific Report
•**Publication Year:** 2023

# STUDY GOAL

**Assess Feasibility:**

- Determine whether geospatial analysis and data visualization methods can effectively detect clinically and statistically significant variations in antibiotic susceptibility rates at a neighborhood level.

**Identify Spatial Patterns:**

- Analyze patient-level antibiotic susceptibility data and residential addresses to uncover spatial patterns in antimicrobial resistance (AMR) across different neighborhoods

**Inform Public Health Interventions:**

- Evaluate the potential of geospatial methods to inform targeted public health interventions by identifying areas with higher rates of antibiotic resistance

# DATA ANALYTICS AND GEOSPATIAL AMR PATTERNS

- Data analytics enables understanding AMR trends through:
- Descriptive Analytics
  - Mapping AMR susceptibility patterns using geospatial data.
- Diagnostic Analytics
  - Identifying spatial clusters of AMR using statistical techniques.
- Predictive Analytics
  - Forecasting resistance trends with machine learning.
- Prescriptive Analytics
  - Recommending antibiotic policies using predictive insights.

# METHODS

**Data Collection:**

**Patient-Level Antibiotic Susceptibility Data:**

- The researchers collected 10 years of patient-level antibiotic susceptibility data from three regionally distinct Wisconsin health systems:
  - UW Health
  - Fort HealthCare
  - Marshfield Clinic Health System (MCHS)

**Geocoding Patient Addresses:**

- Patient addresses were geocoded to obtain spatial coordinates, which were then linked to U.S. Census Block Groups to facilitate neighborhood-level analysis.

**Inclusion Criteria:**

- The study included the initial Escherichia coli (E. coli) isolate per patient per year per sample source, provided the patient had a residential address in Wisconsin.
- Isolates from U.S. Census Block Groups with fewer than 30 isolates were excluded to ensure statistical reliability, resulting in a final dataset of 86,467 E. coli isolates.

# METHODS

**Spatial Autocorrelation Analysis:**

- The study employed Moran's I statistic to assess spatial autocorrelation of antibiotic susceptibility rates.
- Moran's I values range from -1 to +1, indicating whether the data are dispersed, randomly distributed, or clustered.

**Hot Spot Analysis:**

- The researchers conducted hot spot analysis to identify statistically significant local clusters of high (hot spots) and low (cold spots) antibiotic susceptibility within U.S. Census Block Groups.

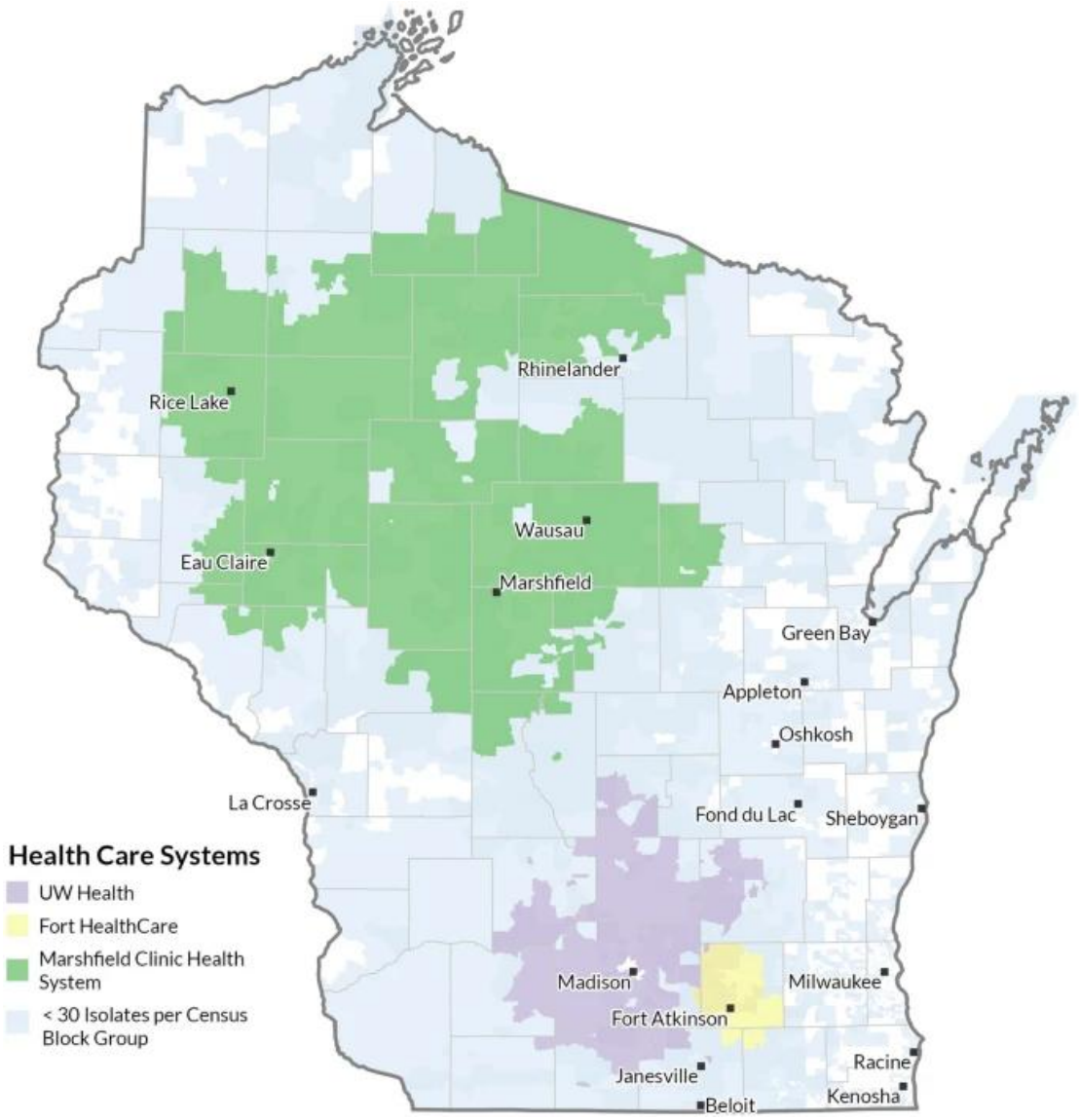
**Data Visualization:**

- Choropleth maps were created to visually represent spatial variations in antimicrobial resistance (AMR) across different neighborhoods

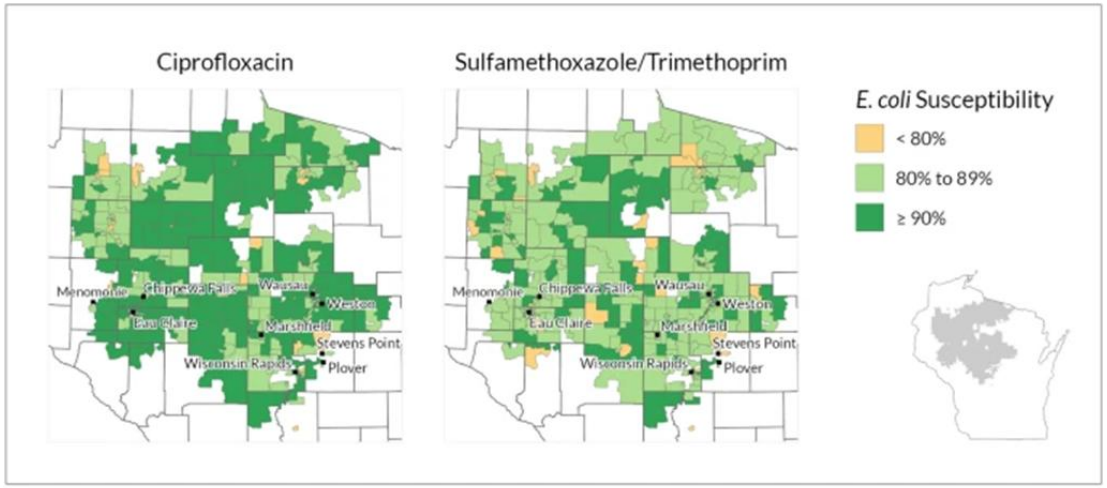
# MACHINE LEARNING METHODS IN AMR PREDICTION

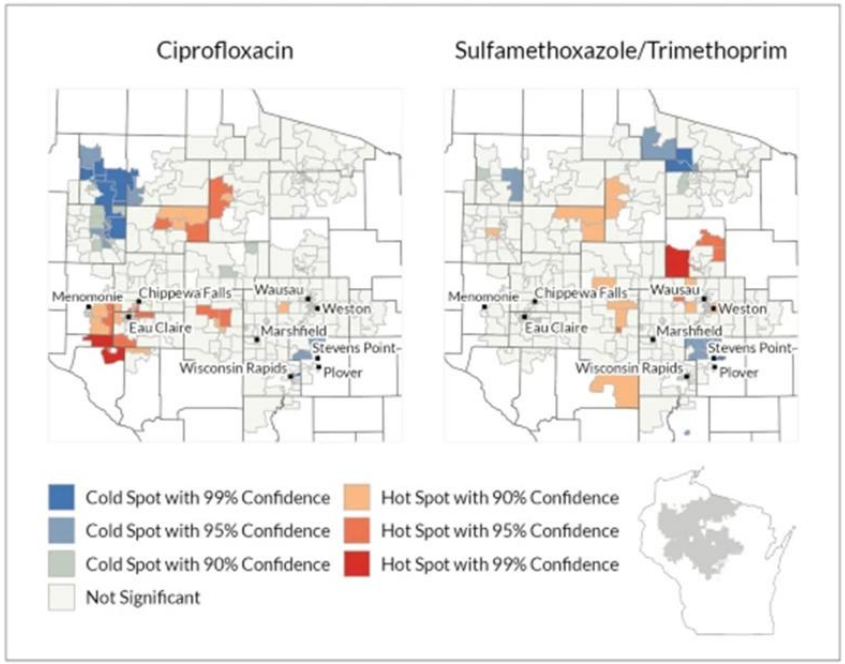
- The study integrates geospatial analysis with machine learning for AMR prediction:
- Spatial Clustering (Moran's I, Hot Spot Analysis) - Identifies resistance clusters.
- Choropleth Maps - Visualizes AMR hotspots at the neighborhood level.
- Regression Models & XGBoost - Predicts resistance trends.
- Geospatial Feature Engineering - Uses spatial relationships for prediction.
- Result: Geospatial ML enhances AMR prediction accuracy by incorporating local resistance patterns.

# DATA VISUALIZATION AND ANALYSIS

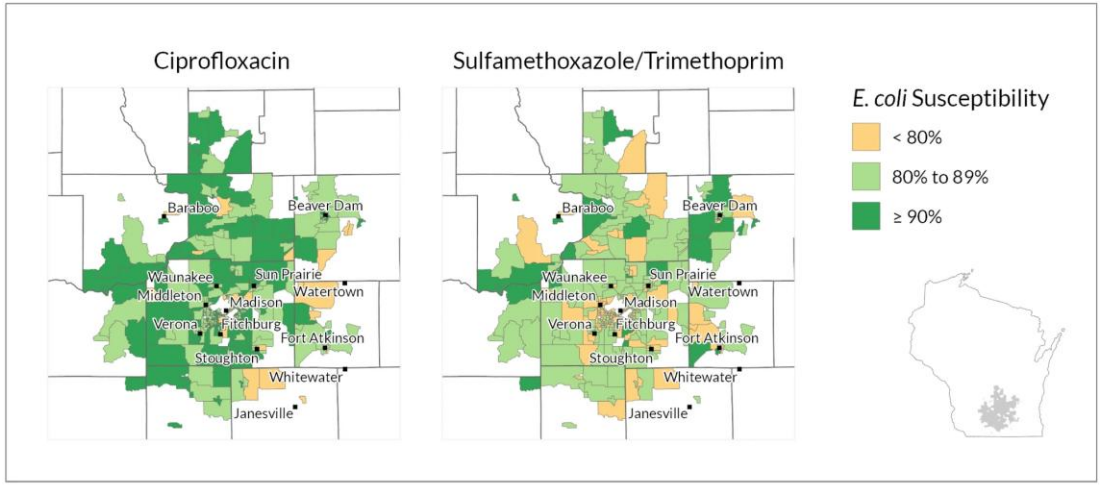


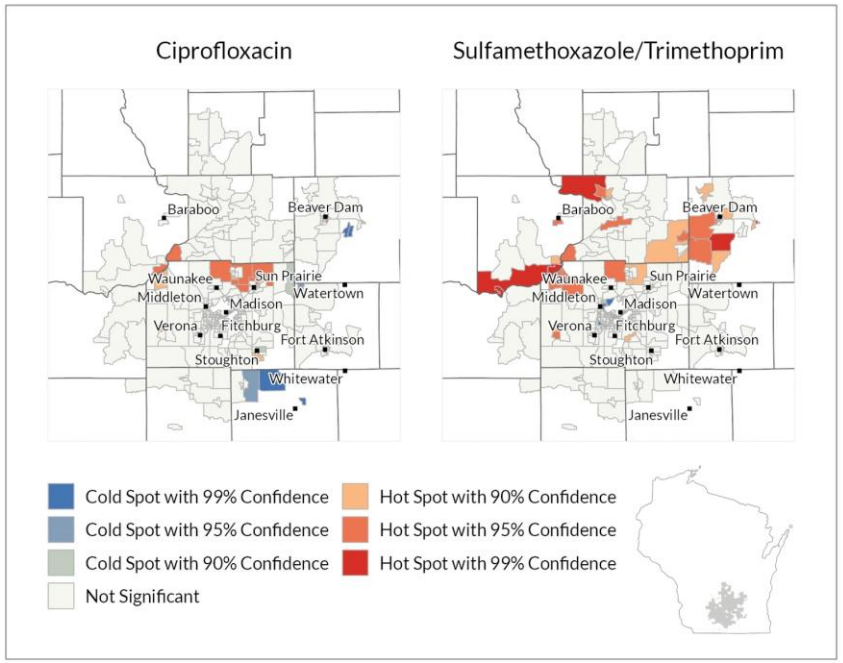
# DATA VISUALIZATION AND ANALYSIS



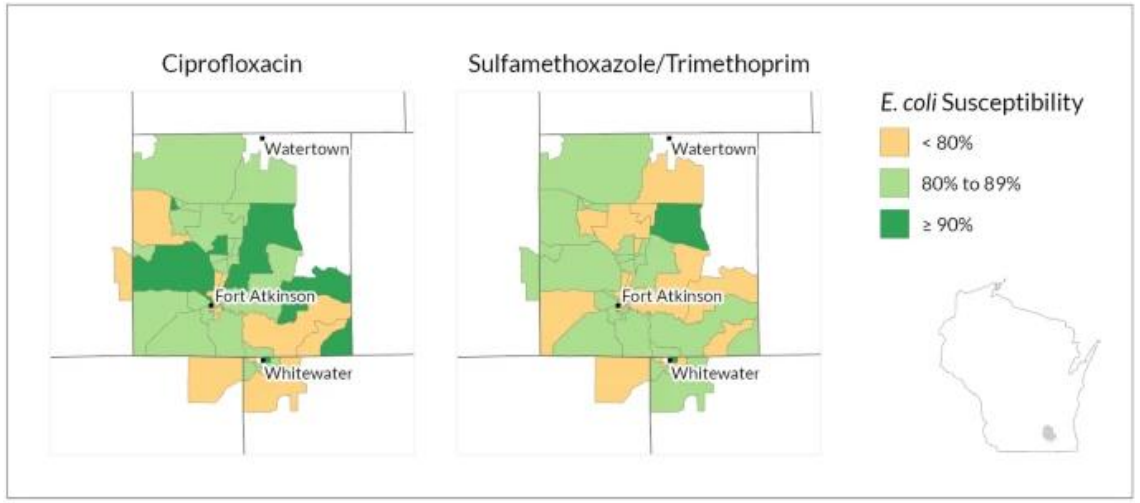



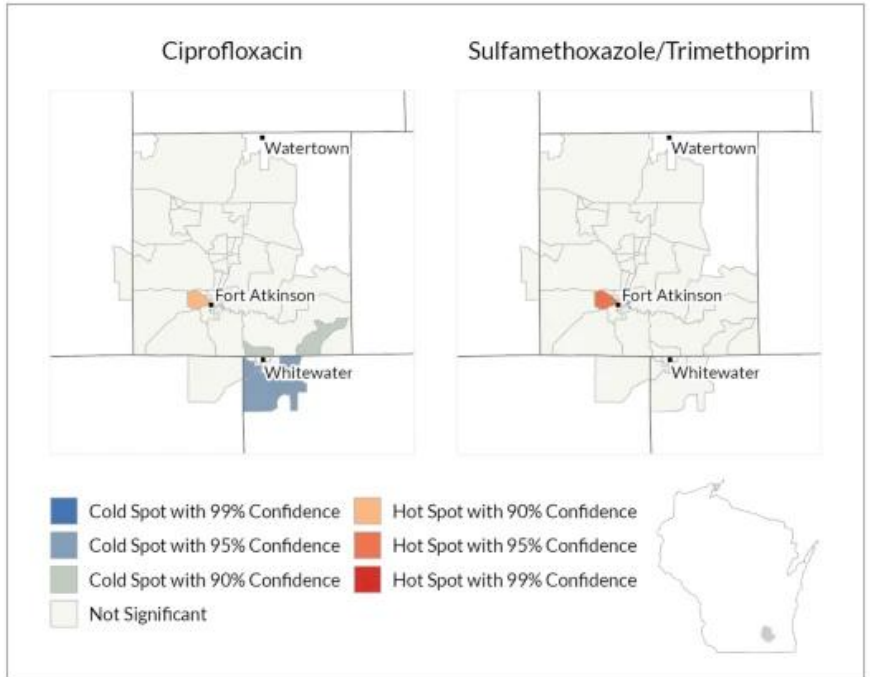
# DATA VISUALIZATION AND ANALYSIS



# DATA VISUALIZATION AND ANALYSIS



# KEY FINDINGS OF THE STUDY

- AMR Hotspots Identified: High-resistance areas detected using spatial clustering.
- Urban vs. Rural Differences: Resistance patterns varied by location.
- Geospatial Models Improve Prediction: Spatial autocorrelation methods enhanced forecasting.
- Public Health Relevance: Findings can inform clinical decision-making and policy adjustments.

# IMPLICATIONS FOR AMR MANAGEMENT

- Geospatial AMR data can improve localized treatment guidelines.
- ML models enable better early detection of resistance trends.
- Spatial insights help in targeted public health interventions.
- Future research should explore real-time geospatial AMR monitoring.

# CONCLUSION & FUTURE DIRECTIONS

- Geospatial Analysis enhances AMR surveillance and prediction.
- ML-based models improve the accuracy of AMR forecasting.
- Integration with clinical decision tools can guide antibiotic prescriptions.
- Further research needed on real-time AMR geospatial monitoring.

# THANK YOU