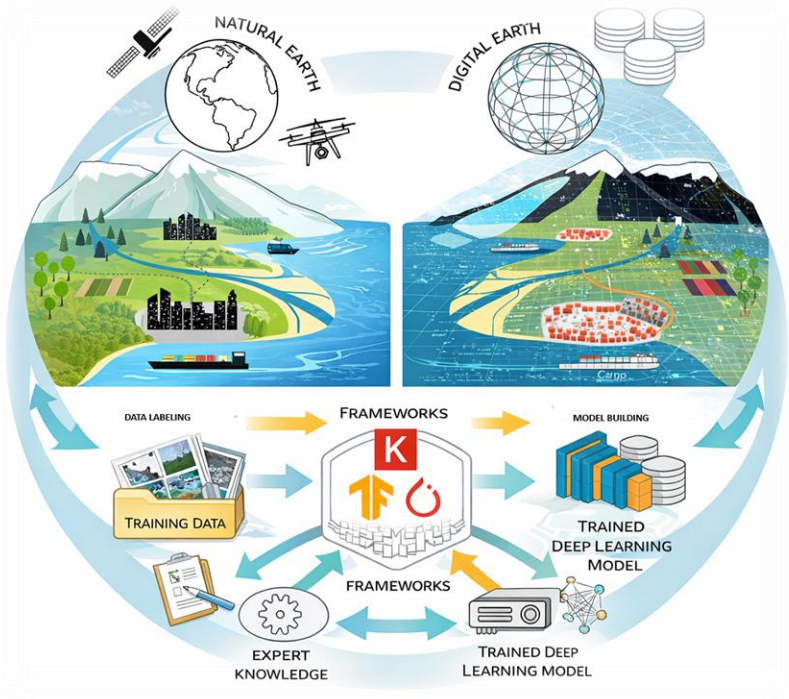
From Data to Insight: Applying Machine Learning to Environmental Challenges

Hafez Ahmad

Assistant Professor Candidate

MISSISSIPPI STATE UNIVERSITY™
DATA SCIENCE ACADEMIC INSTITUTE

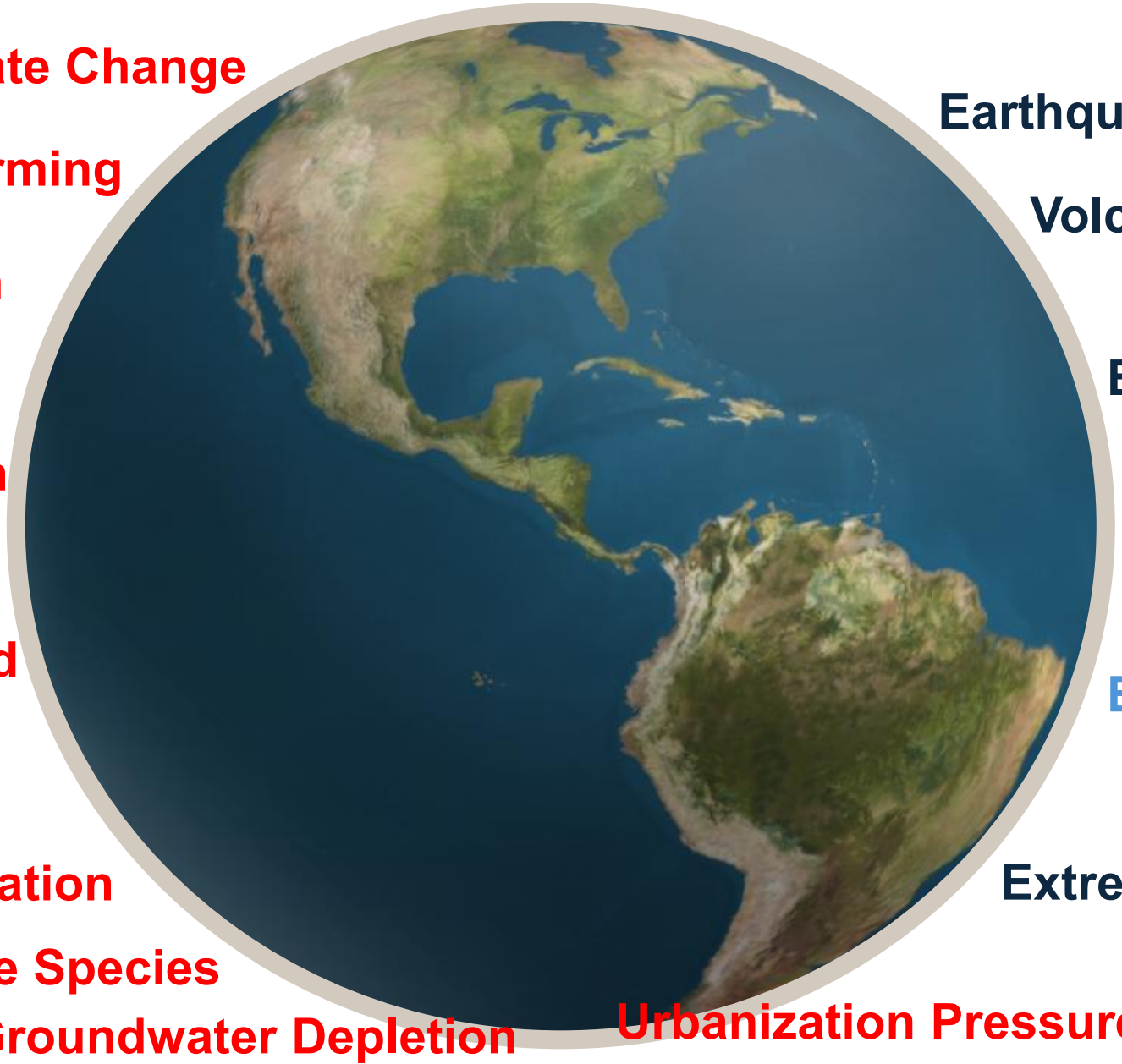


Syllabus

Learning Goals

1. Understand Environmental Challenges
2. Data Explosion
3. Machine Learning in Environmental Contexts
4. Machine Learning Techniques
5. Large Language Models
6. Data Handling & Machine Learning Workflow
7. Model Evaluation & Interpretation
8. Deployment & Scalability
9. Ethical & Equity Considerations
10. **A Case Study + Hands-on Lab**

Environmental Challenges



Climate Change

Global Warming

Pollution

Plastics

Deforestation

Acidification

Energy and Food Supply issues

Overfishing

Soil Degradation

Invasive Species

Groundwater Depletion

Urbanization Pressure

Earthquakes

Volcanoes

Epidemic

Fresh Water Scarcity

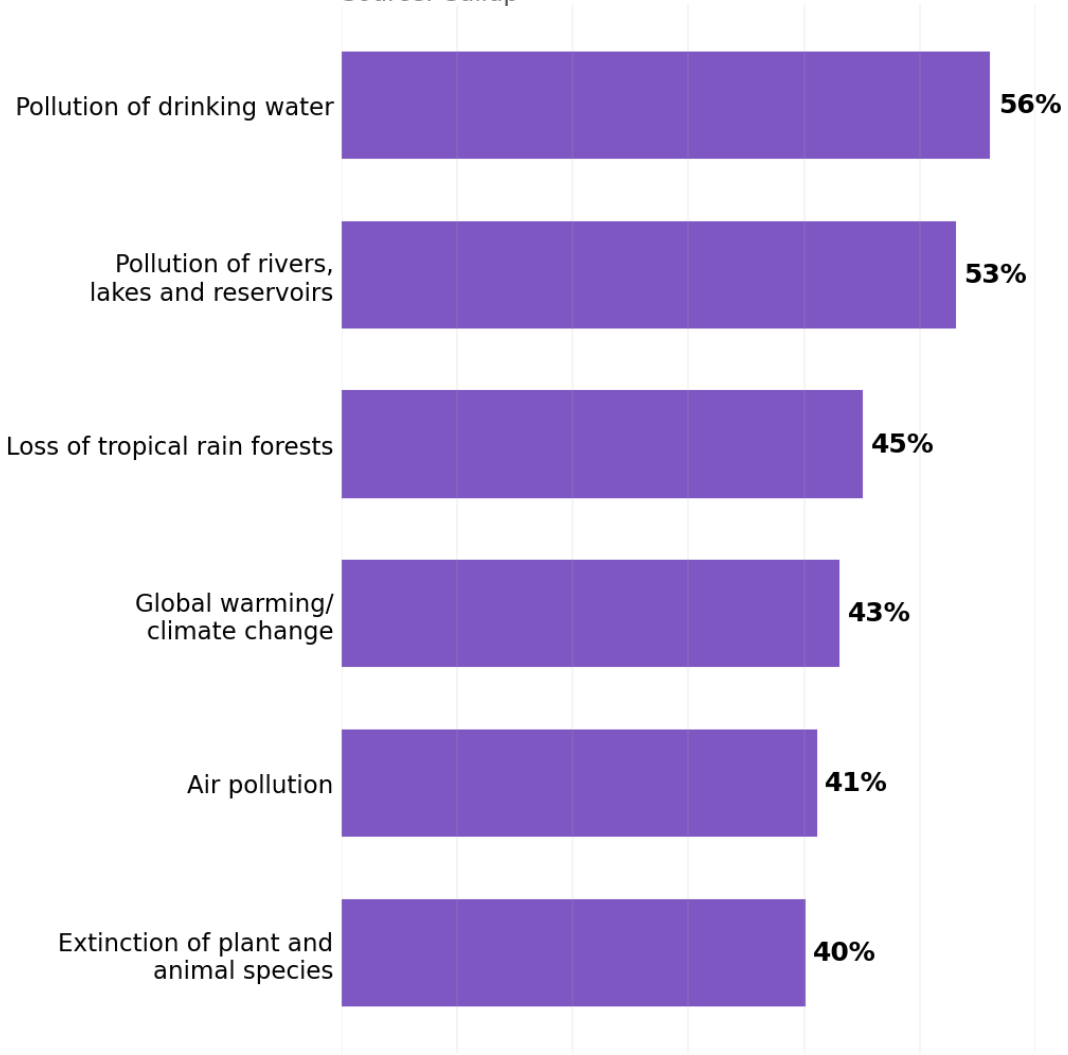
Biodiversity loss

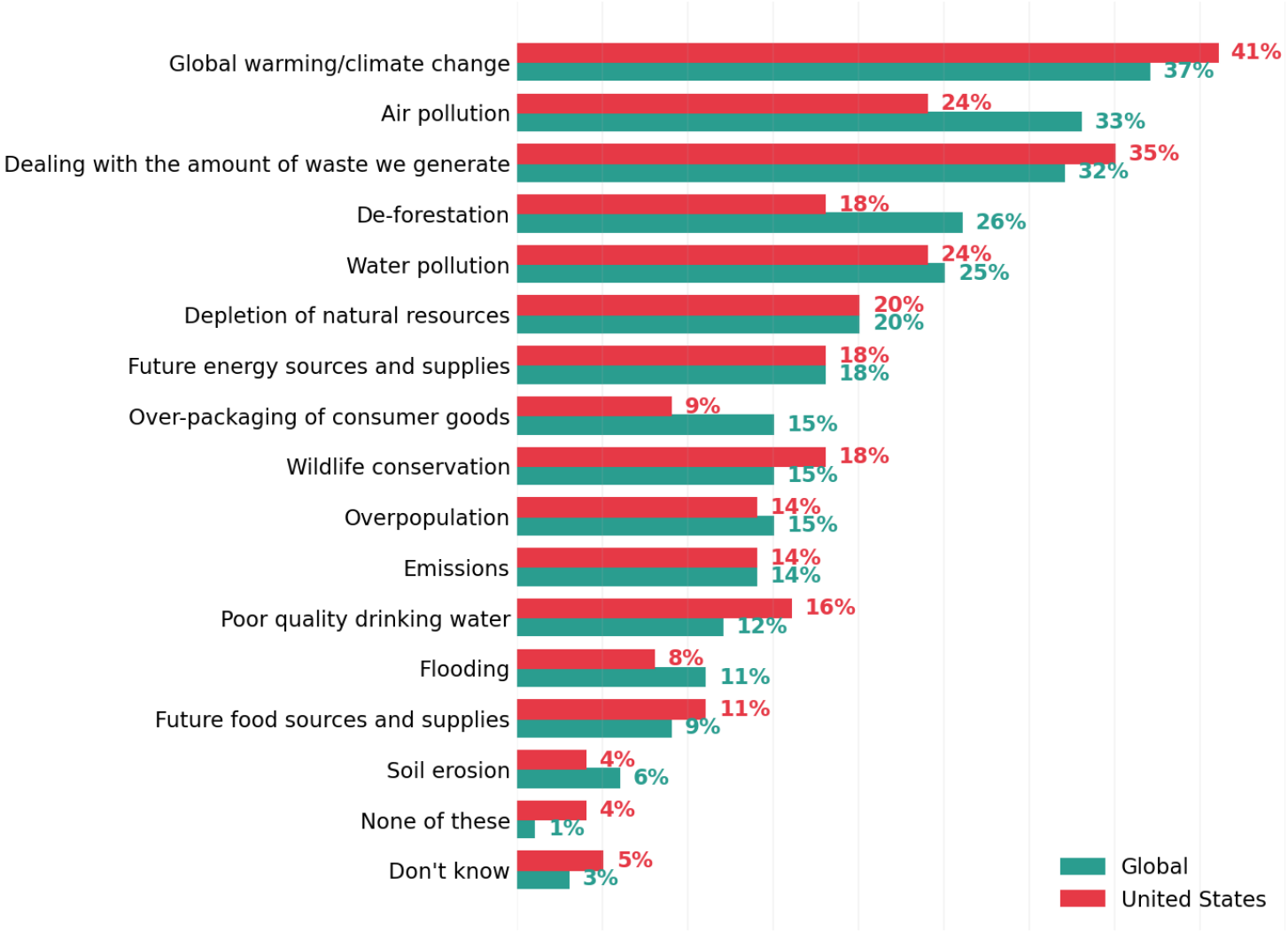
Extreme Weather Events

Environmental Challenges

Environmental Challenges: U.S. vs Global

n = 1,010 U.S. adults (March 01–15, 2021)
Source: Gallup

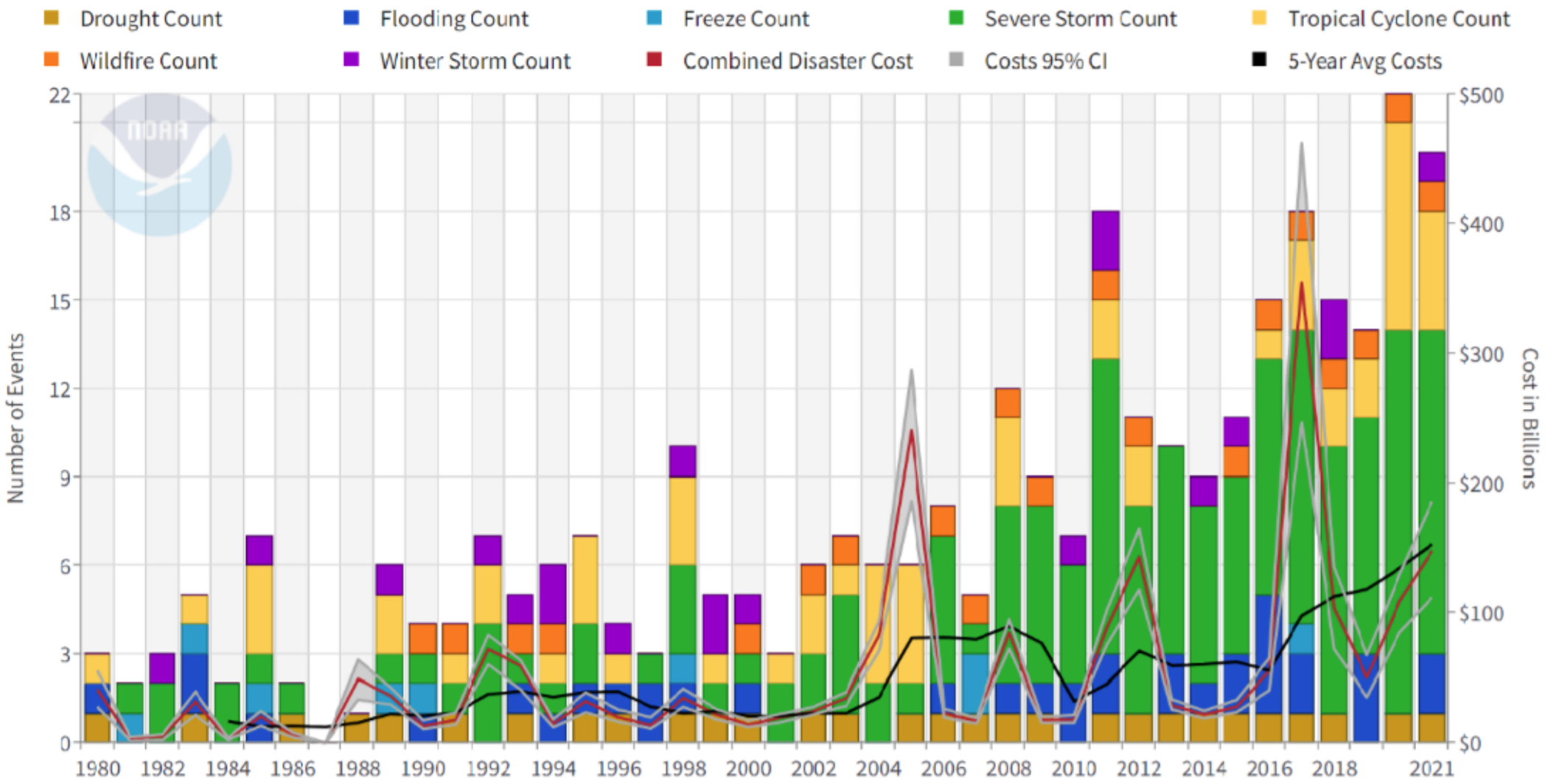




code

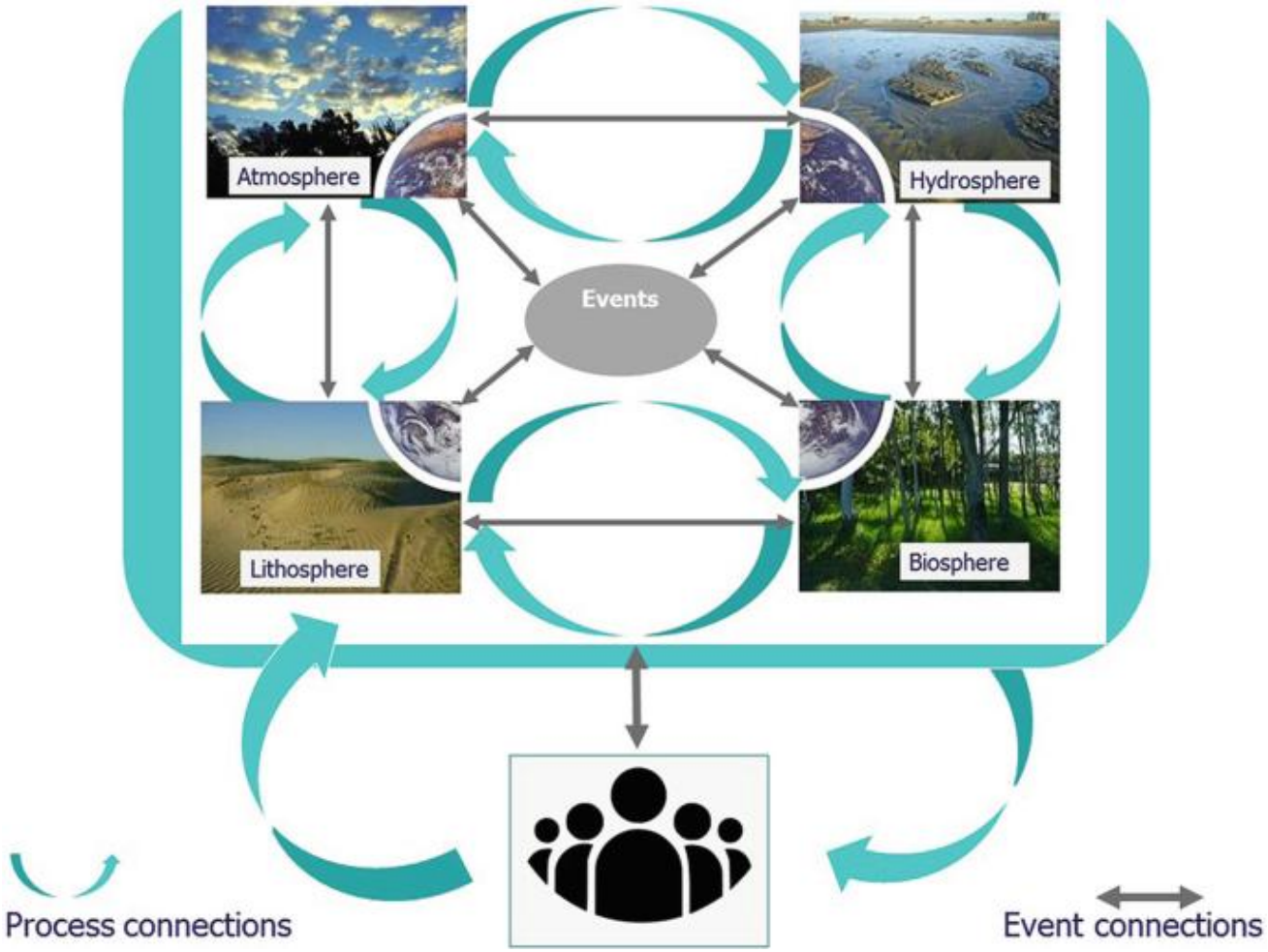
@Ipsos

Escalating Impacts: Billion-Dollar Disasters



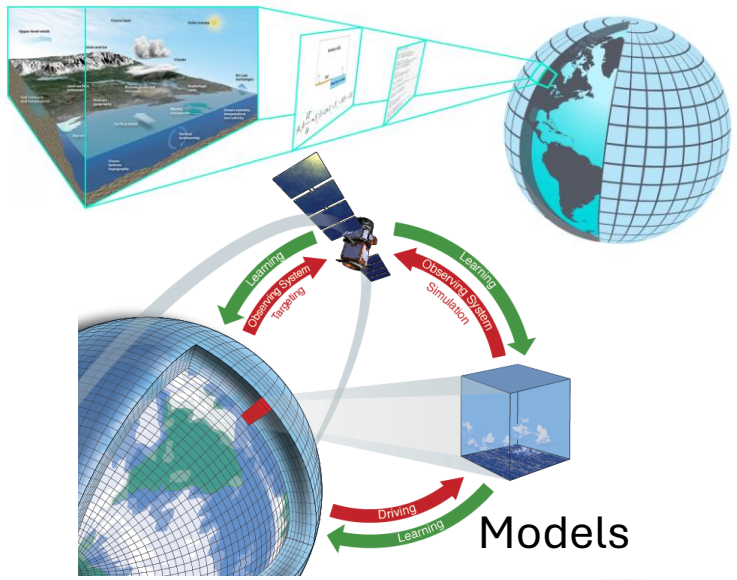
Environmental Systems

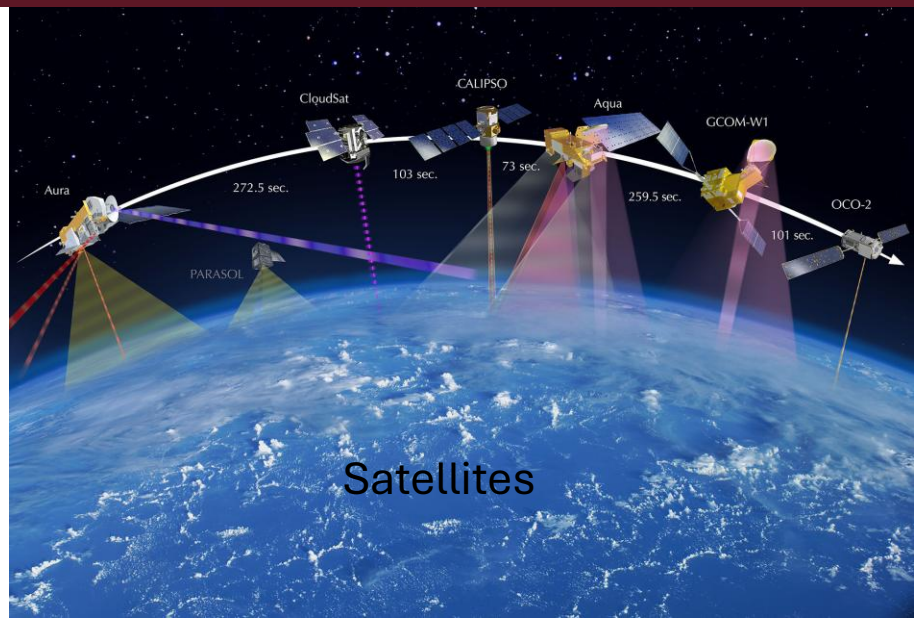
- Interconnected components: atmosphere, land, water, and human systems
- Processes span local, regional, and global scales
- Highly dynamic ,nonlinear, and feedbacks across scales
- Hard to keep pace with scale and variability

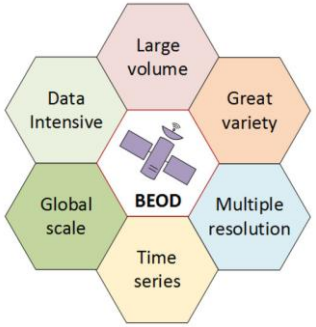


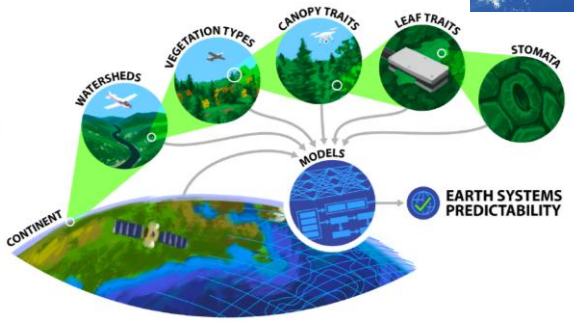
Dixon (2025)

World of exploding data



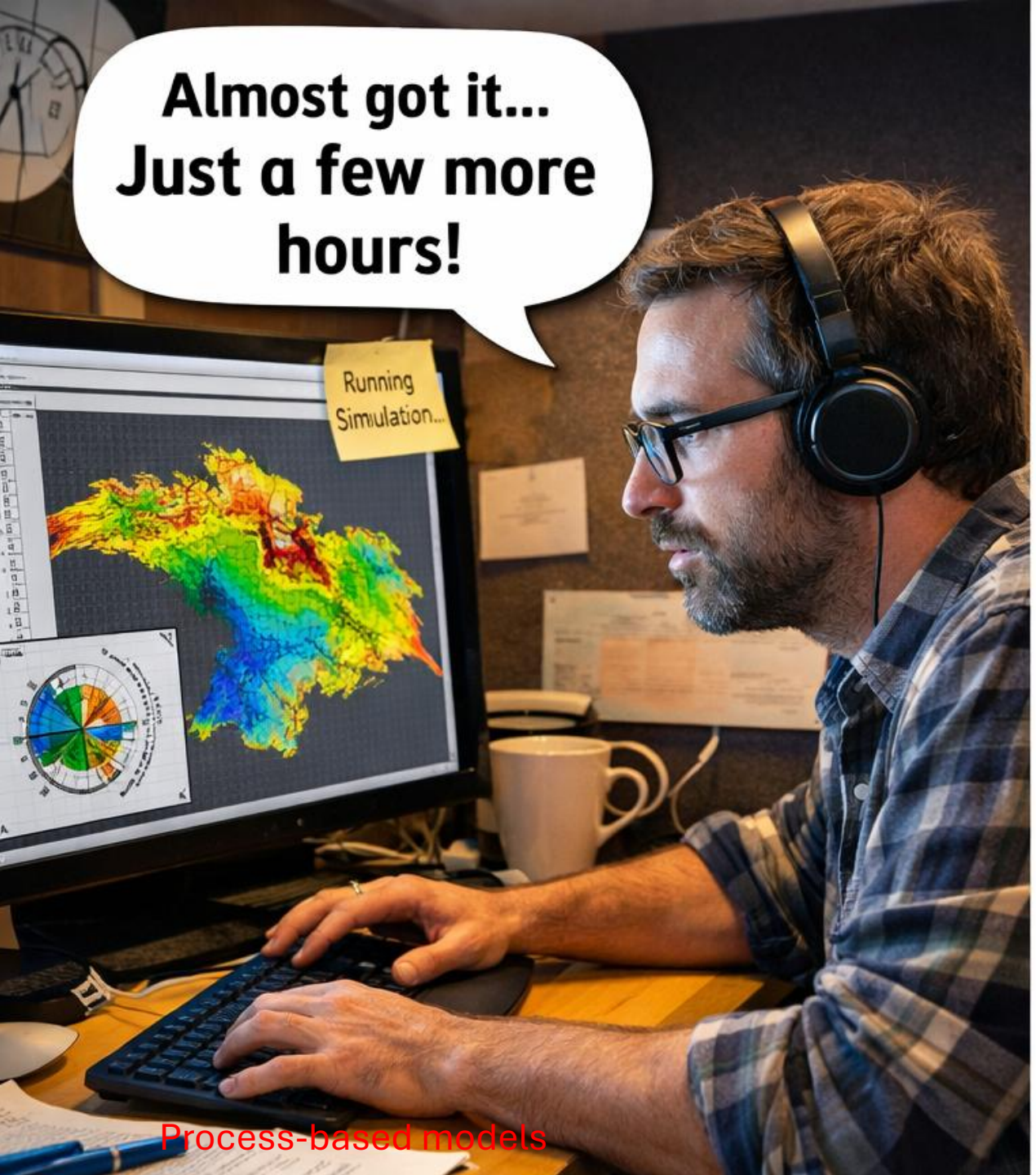






World of exploding environmental data

But MORE data ≠ MORE insight



Process-based models

Machine Learning

Traditional Programming

- Explicit rules written by humans
- Works well when logic is known and stable

Machine Learning (Training)

- Computer learns rules from examples
- Patterns are inferred, not hand-coded

Machine Learning (Prediction)

Traditional programming uses predefined rules to process data. Machine learning learns rules from examples and applies them to new

What Machine Learning Enables

Machine Learning helps us

1. Detect **patterns** we cannot see manually
2. **Fuse** diverse datasets
3. Handle nonlinear relationships
4. **Predict** future conditions
5. Quantify **uncertainty** and drivers
6. Produce **decision-ready** information.

Machine Learning does **not** replace process-based models ; it complements physics and domain knowledge.

Example Applications

Precision Agriculture : Crop stress mapping, irrigation optimization

Coastal Zone : HAB detection, hypoxia prediction

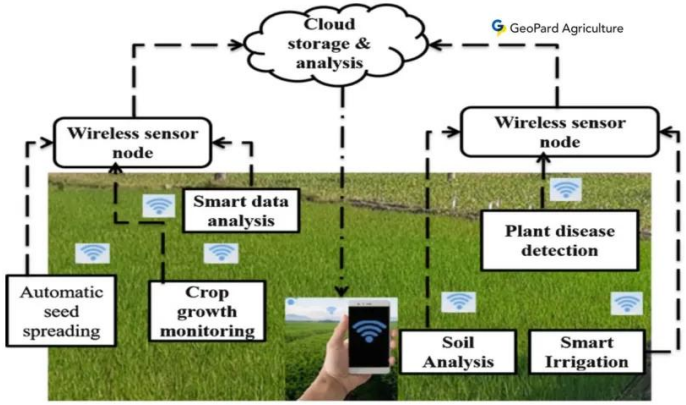
Nutrient & Soil: Predictive models for nutrient runoff, leaching, and soil condition

Hydrology: Flood risk modeling, sediment tracking

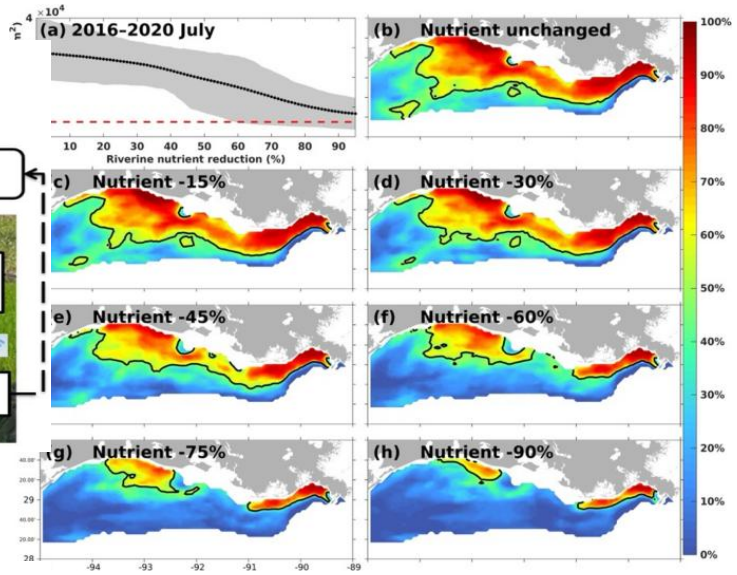
Ecosystem management: Wetland monitoring, biodiversity mapping

Climate Risk & Decision Support: Early warning for drought, heat stress, and extreme events

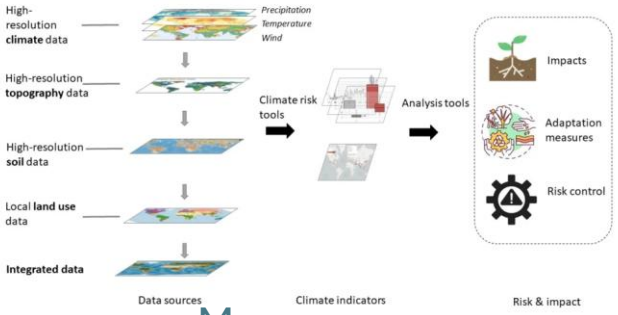
Policy interface: Outputs into management tools



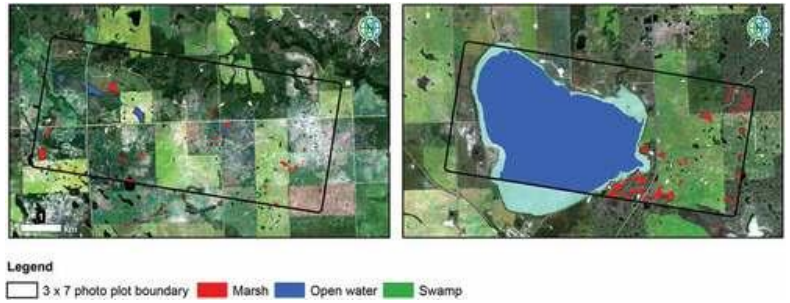
More



Ou et al. 2025



More

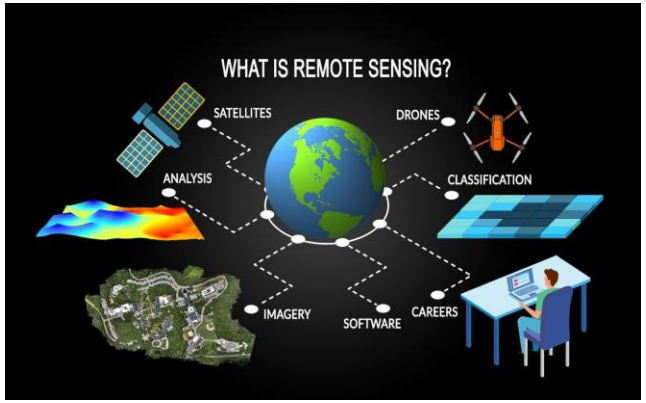


Example Applications

Atmosphere	Urban Heat Island, Air Quality, Extreme Weather, Pollution, Solar Radiation
Hydrology	Flood Risk, Groundwater, River Discharge, Snowpack, Water Quality
Ecology	Habitat Suitability, Carbon Sequestration, Biodiversity, Forest Health
Agriculture	Crop Yield, Precision Irrigation, Pest Forecasting, Soil Nutrients
Coastal	Harmful Algal Bloom (HAB), Hypoxia, Sediment Transport
Climate	Climate Risk, Carbon Flux, Drought Prediction
Urban	Stormwater Management, Green Infrastructure, Urban Cooling

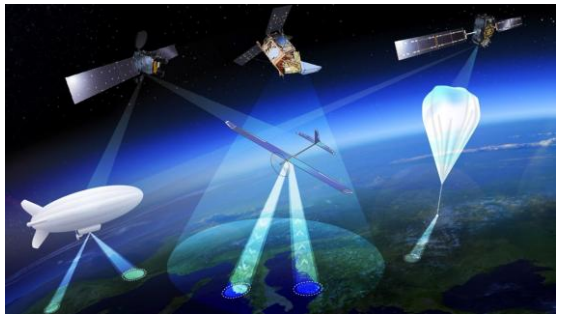
Data Sources and Formats

Remote sensing: Satellites (Sentinel-2, Landsat 8/9, MODIS, VIIRS), Drone, LIDAR & weather balloons

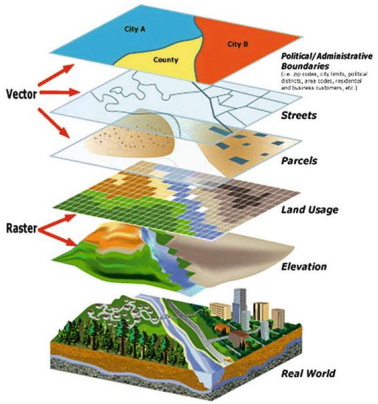


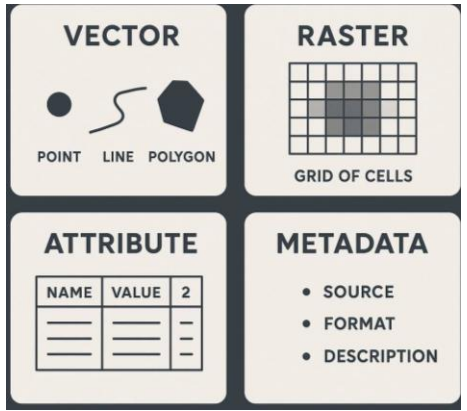
In-situ sensors: Nutrients, Soil, Water, temperature, DO, meteorological networks

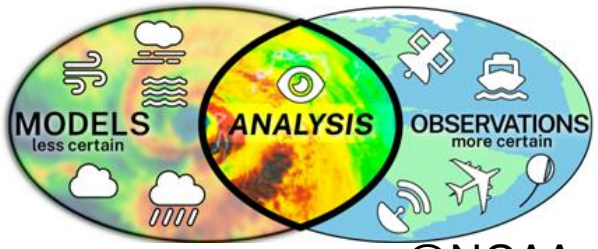
Simulation Model: Climate reanalysis, hydrodynamic and watershed models



Common formats: Txt, NetCDF, GeoTIFF, CSV, JSON,GRIB shapefile/GeoPackage, HDF, Zarr







@NOAA

Book 2

Data Handling and Preprocessing

Clean, well-structured data is the foundation of reliable ML models.

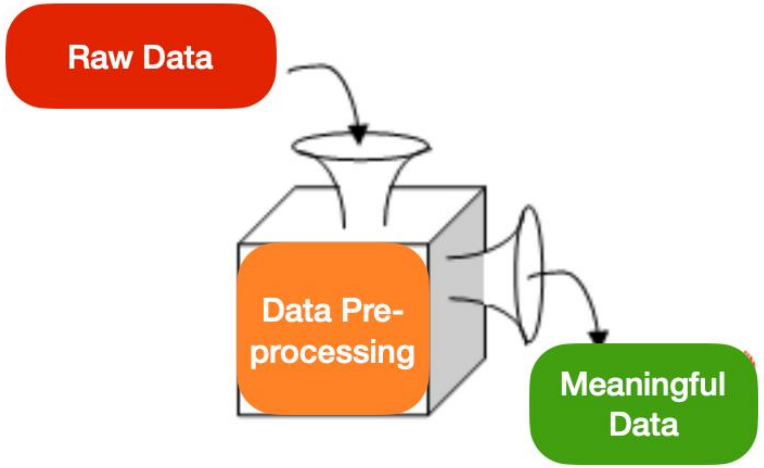
- **Data Quality Checks**:
  Remove outliers and handle missing values.

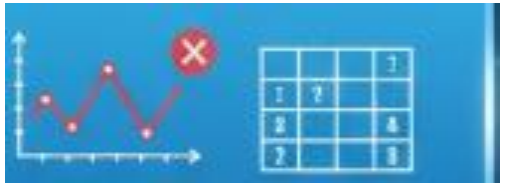
- **Spatial and temporal alignment**:
  Consistent projections and coordinate systems

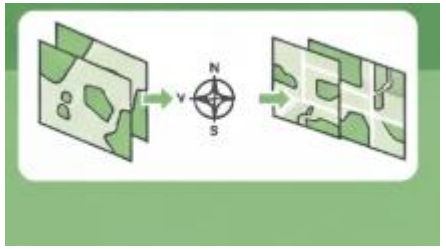
- **Normalization & Scaling**:
  Techniques like z-score or min-max scaling

- **Automation & Reproducibility**:
  Scripted workflows

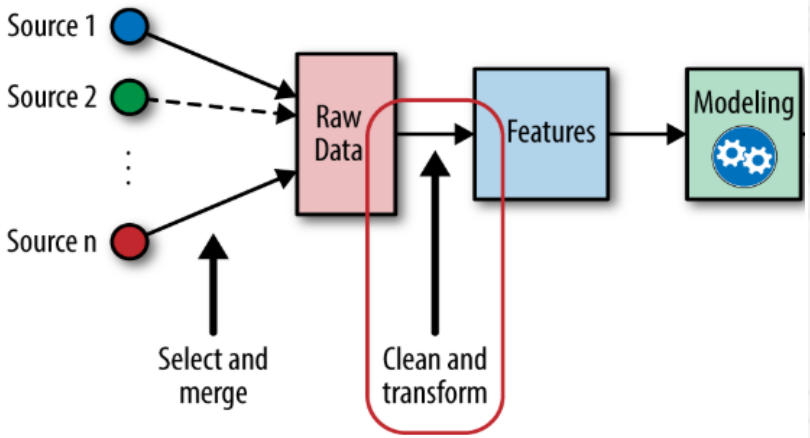
- **Documentation:**
  Steps for transparency and reproducibility

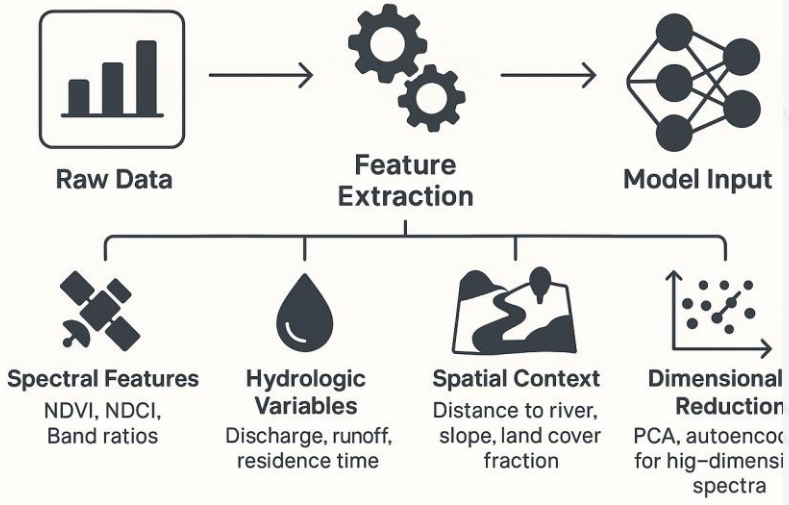


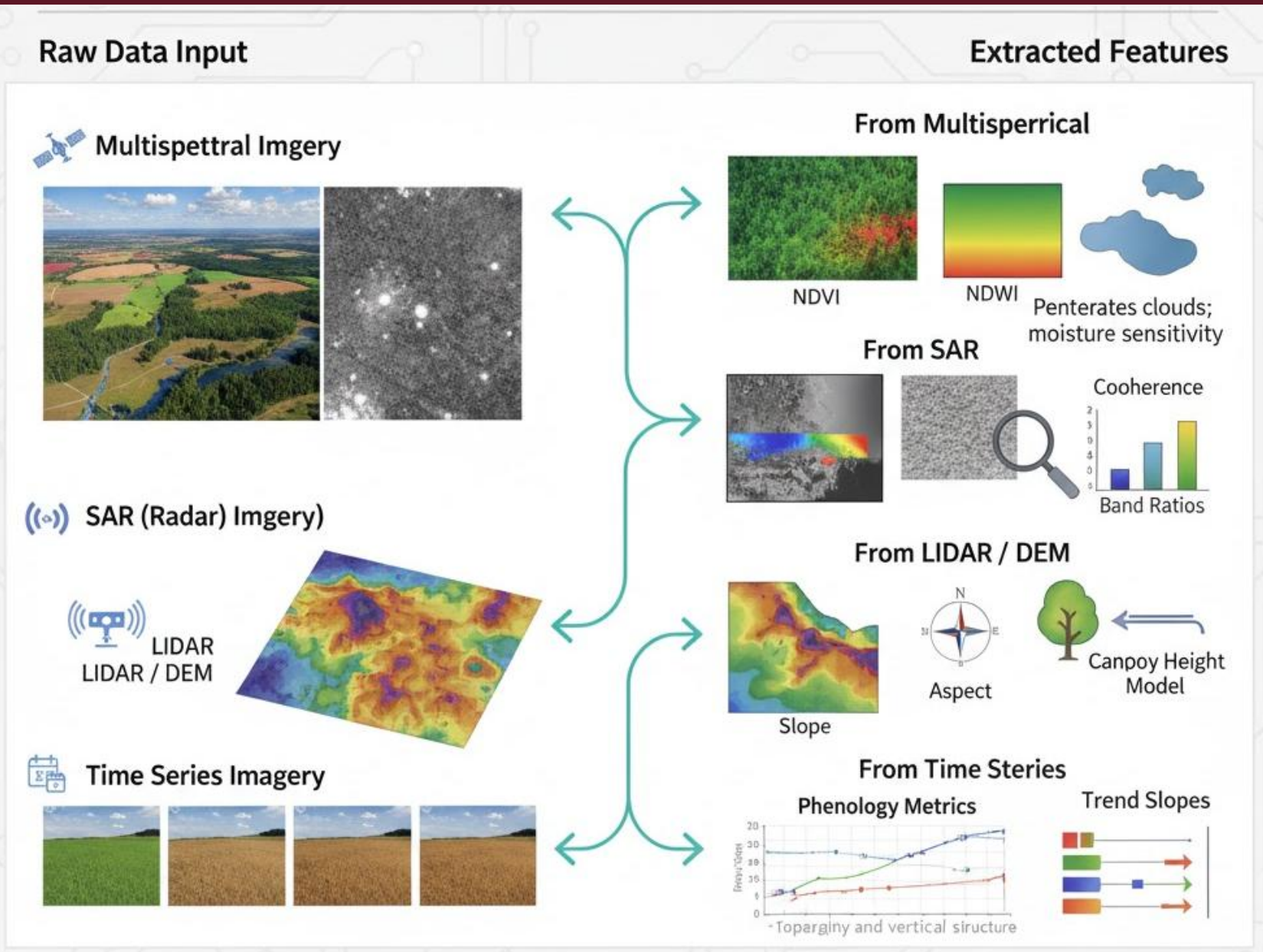




From Raw Data to Features







Machine Learning Techniques

Supervised Learning

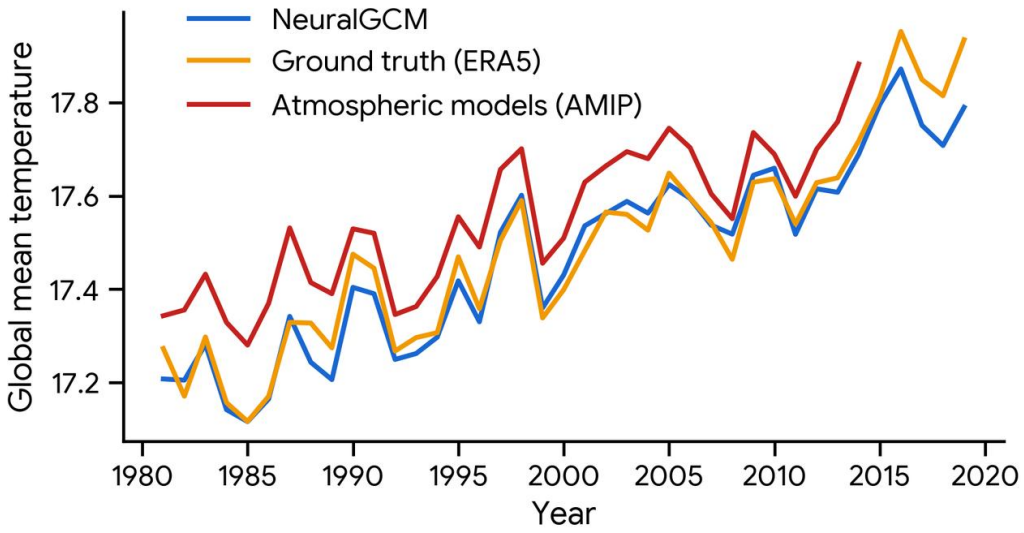
- Regression: Continuous numerical value
- Classification: Category or class label

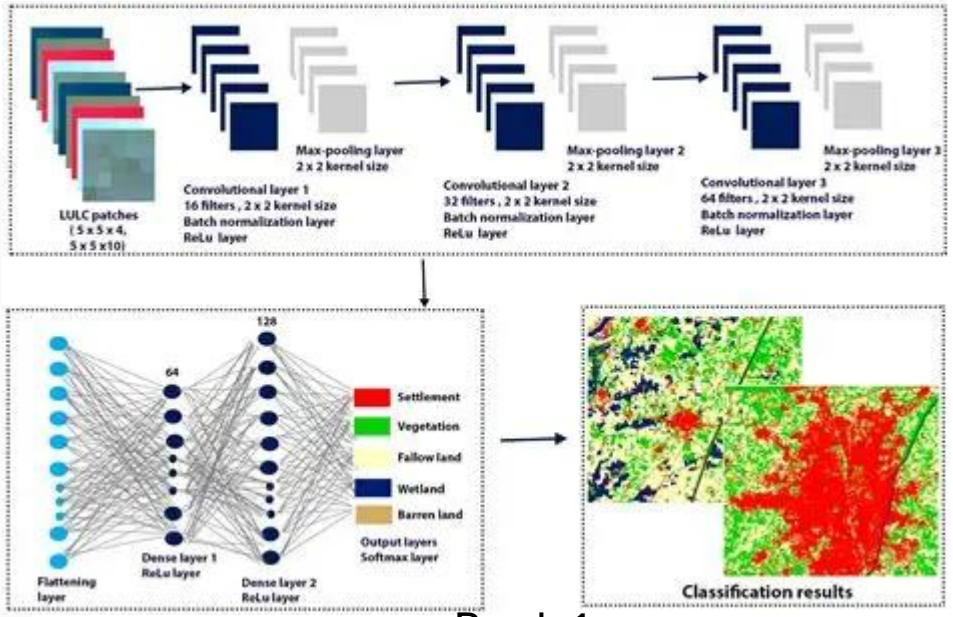
Unsupervised Learning

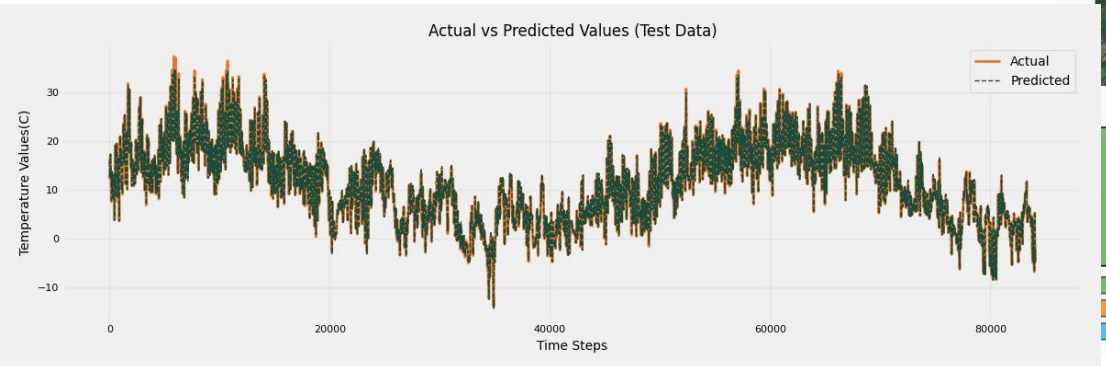
- Group similar regimes
- Identify unusual patterns

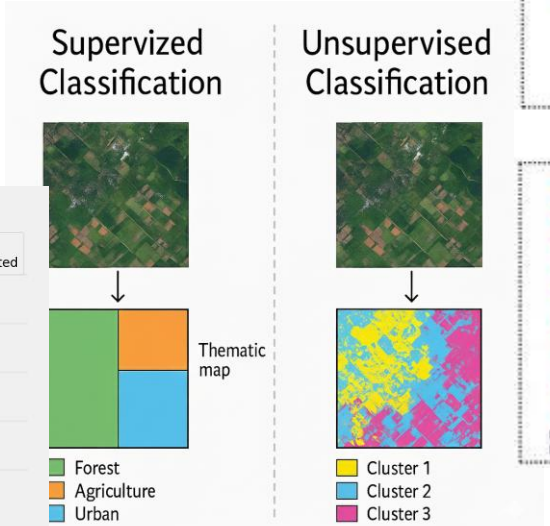
Deep Learning

- CNN
- LSTM









Book 1

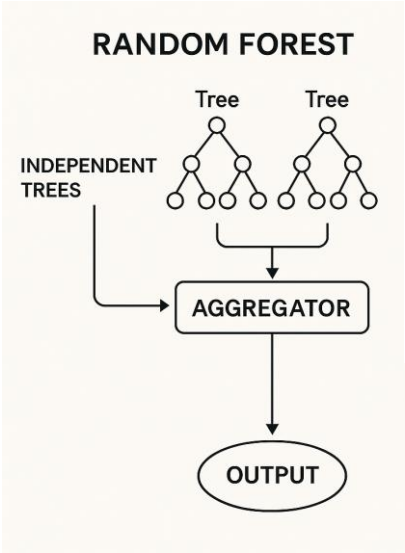
Algorithm Spotlight

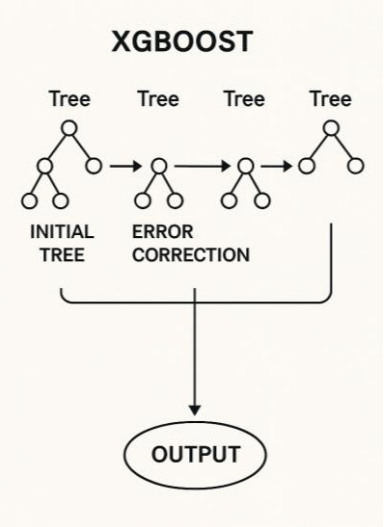
Random Forest:

- ❑ Ensemble of many independent decision trees
- ❑ Stable, robust baseline
- ❑ Handles nonlinear relationships automatically
- ❑ Provides simple feature importance

XGBoost / Gradient Boosting:

- ❑ Ensemble
- ❑ Trees built sequentially to correct previous errors.
- ❑ Effective when many drivers interact
- ❑ Often delivers higher predictive accuracy.

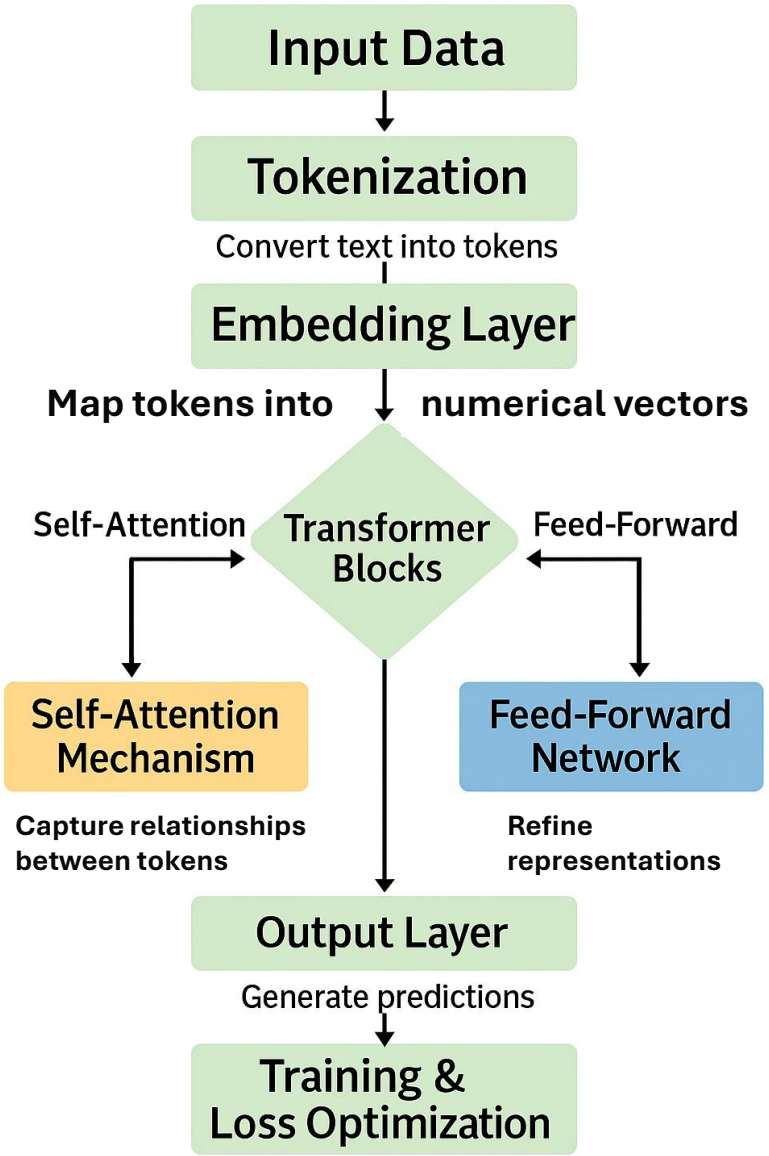


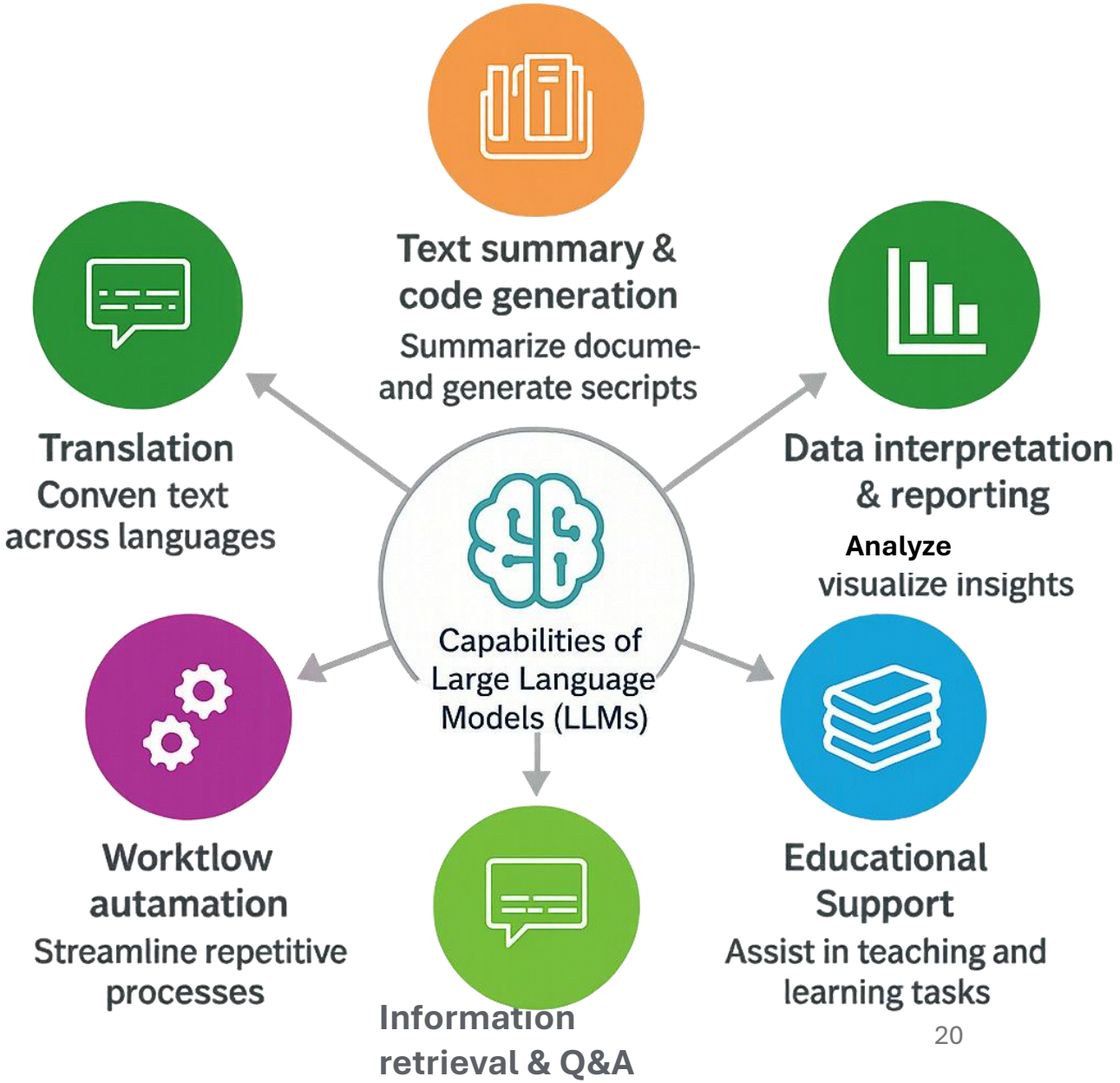


@

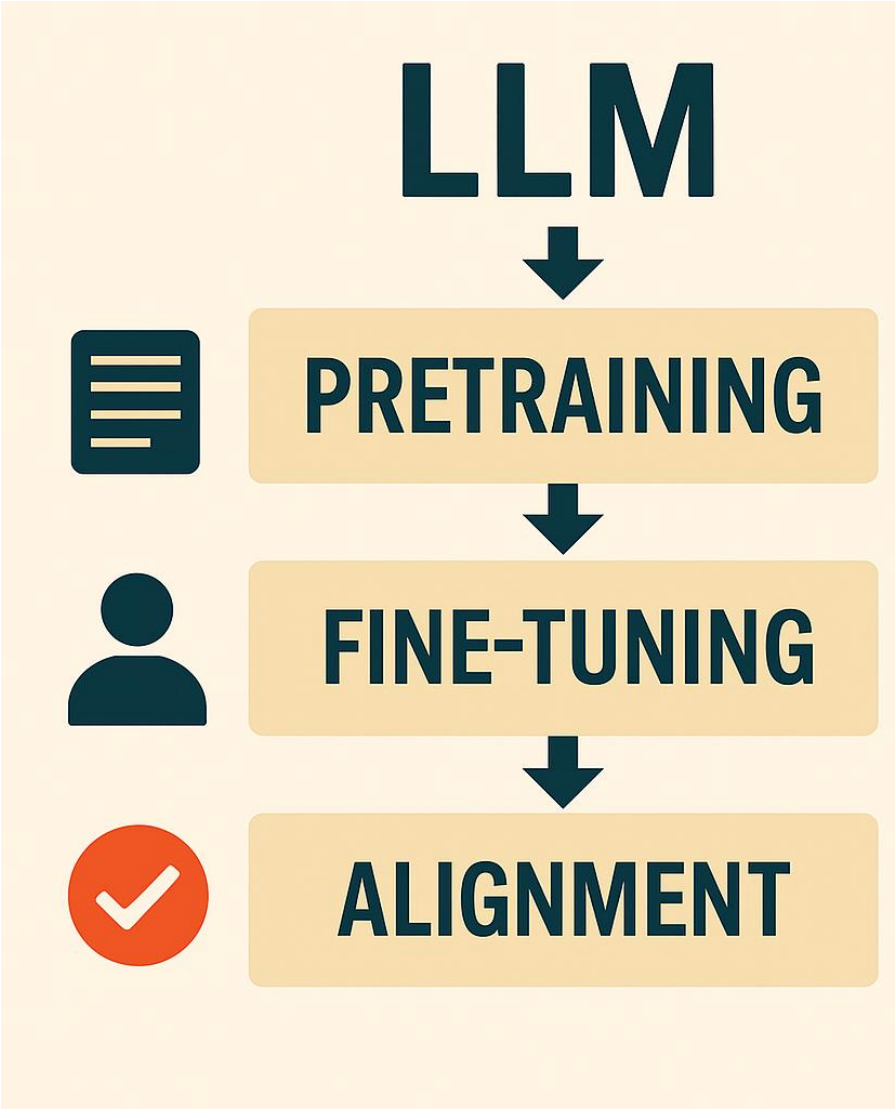
Book 3- P123
Book 4-P137

Large Language Models





Large Language Models



Pretraining: Unsupervised

Fine-tuning: Supervised (human-labeled examples)

Alignment: Reinforcement learning

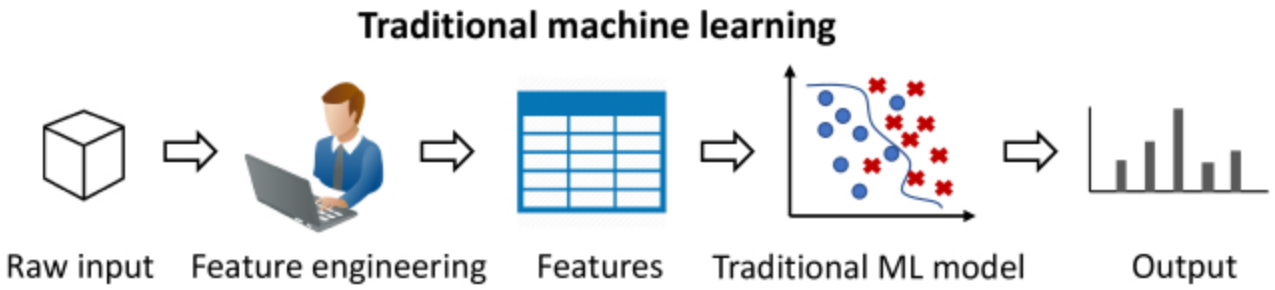
Machine Learning vs. LLMs

Goal: Predict outcomes from structured data

Examples: Random Forest, XGBoost, LSTM

Strengths: Tabular, geospatial, time-series, and clustering tasks

Limits: Needs labeled data, feature engineering

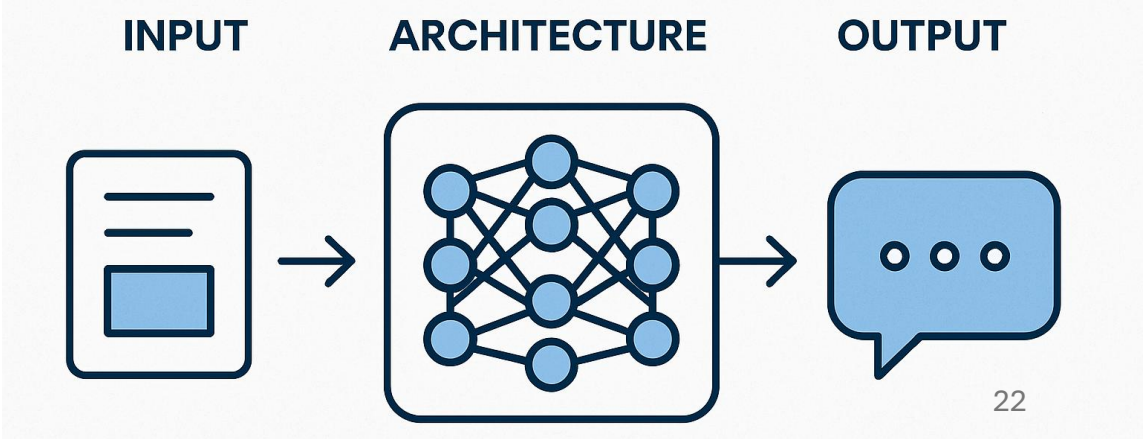


Goal: Understand & generate text

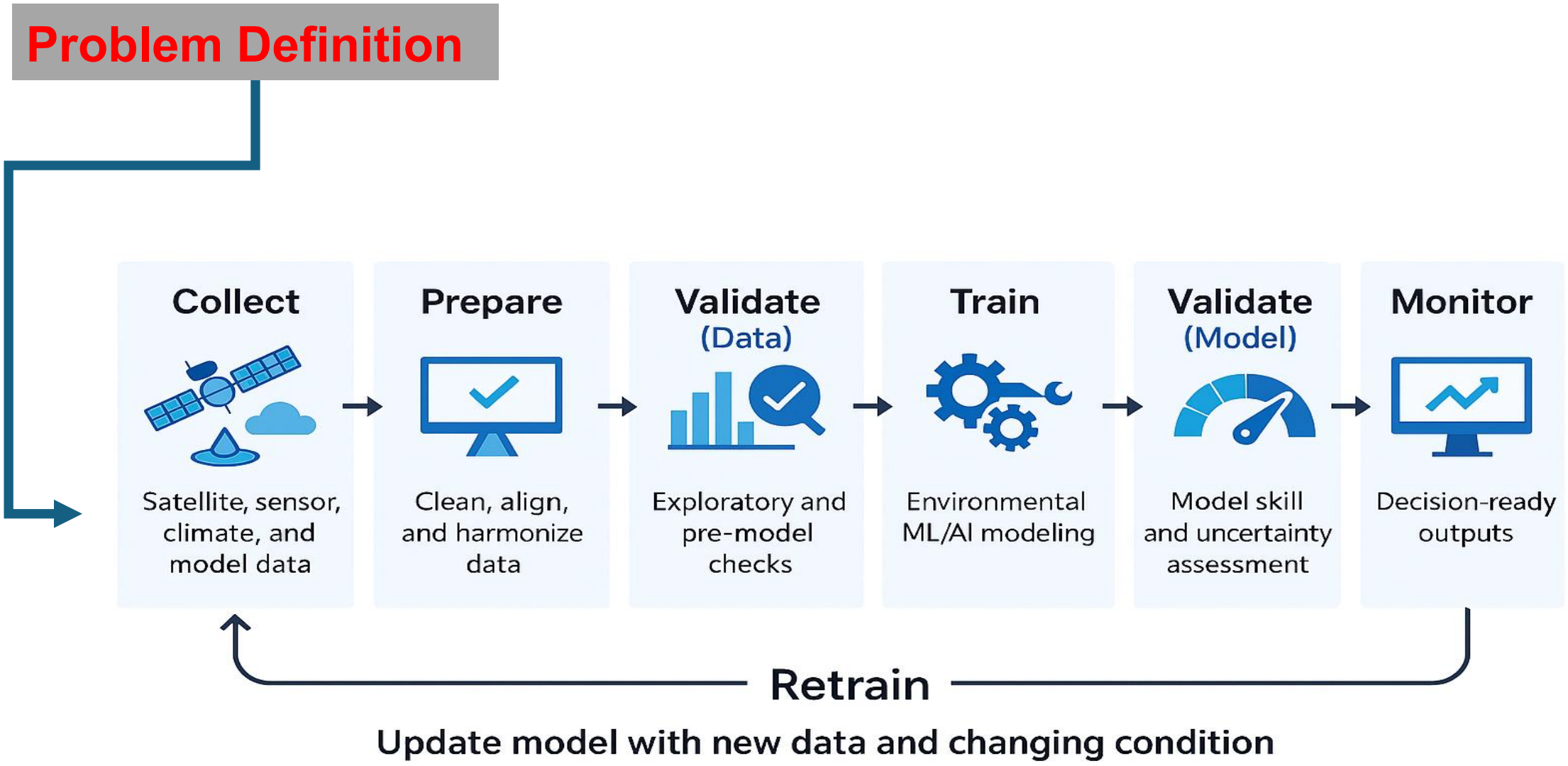
Examples: ChatGPT, Gemini, Claude

Strengths: Code generation, text summarization, and text generation

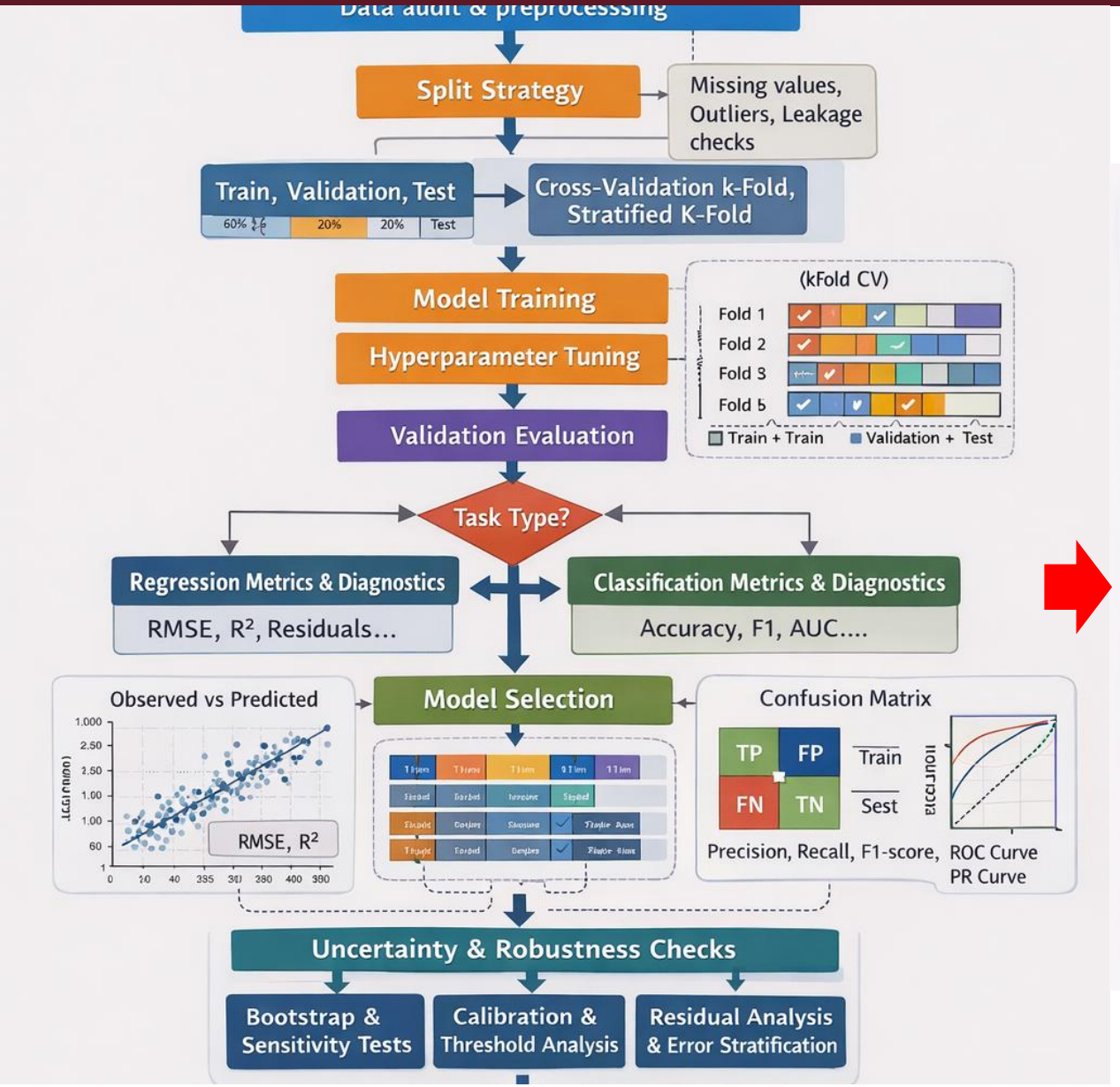
Limits: Not for numeric/spatial prediction; risk of hallucination and Bias



Machine Learning Workflow

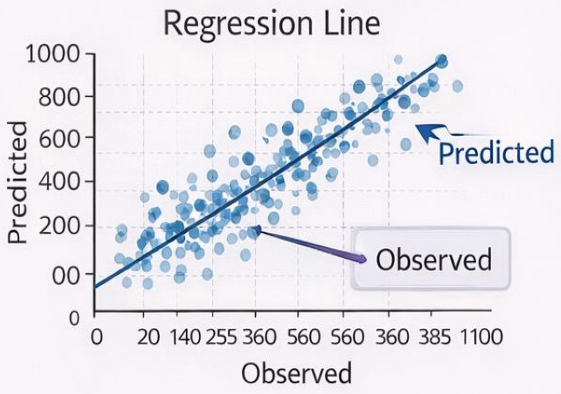


Model Evaluation and Validation



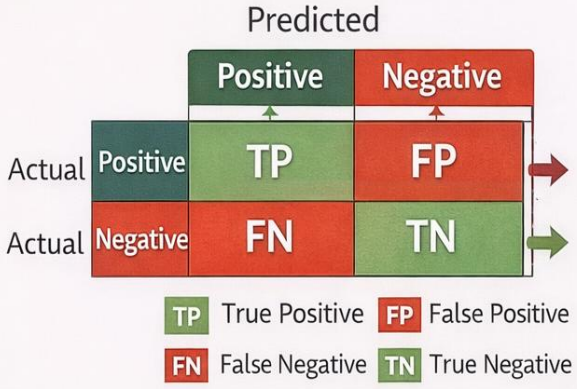
Regression Metrics

- RMSE
- R^2
- MAE
- MAPE
- MSE, MSE
- Residual Analysis



Classification Metrics

- Accuracy
- Precision
- Recall
- F1 Score
- AUC
- Confusion Matrix

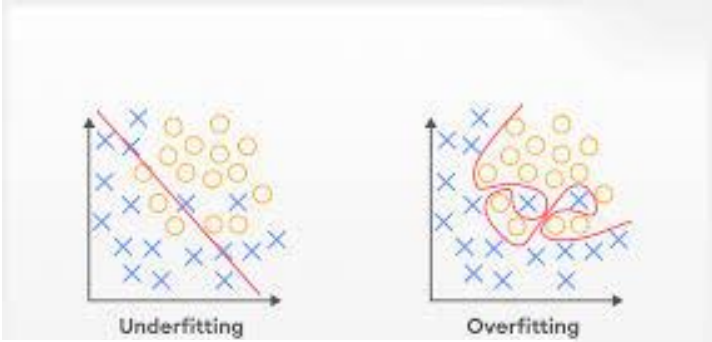


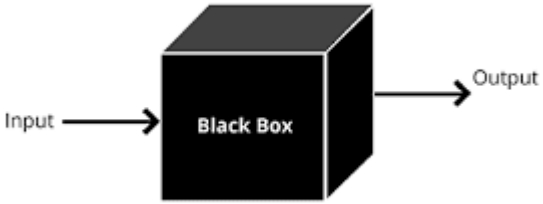
Model Interpretation and Explainability

- **Interpretable without explicit equations and code**
- **Feature importance and key drivers**
- **Explainable AI improve transparency**

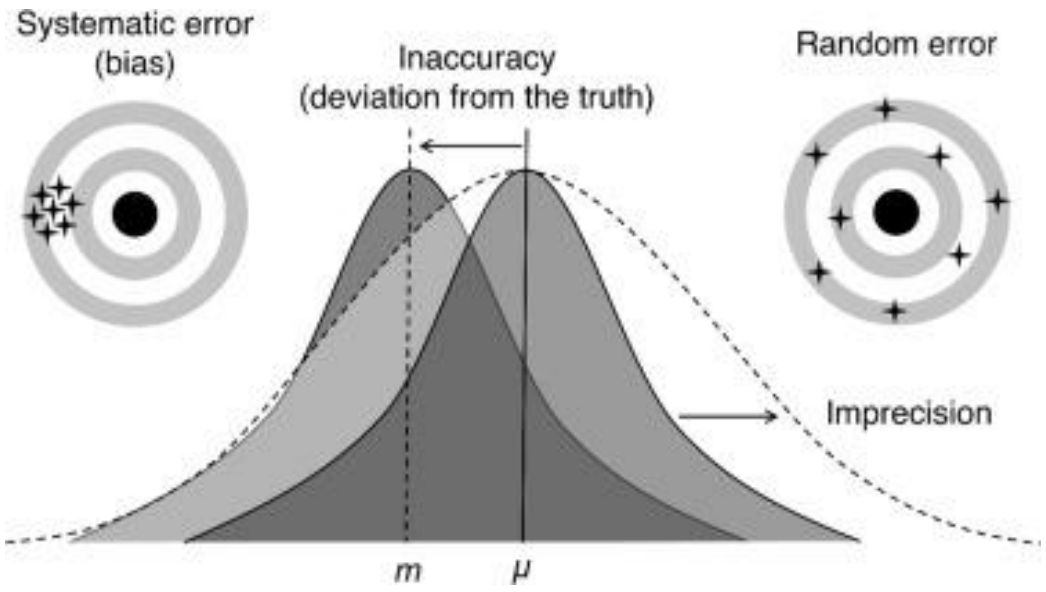
Limits of Machine Learning

- Data bias
- Extreme events
- Risk of overfitting
- Concept drift
- Black-box behavior

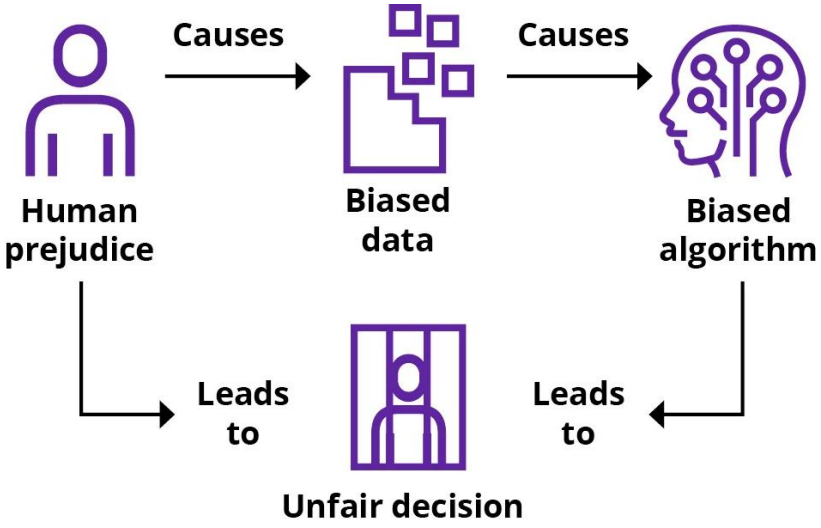




Bias and Uncertainty



Konno et al. 2025

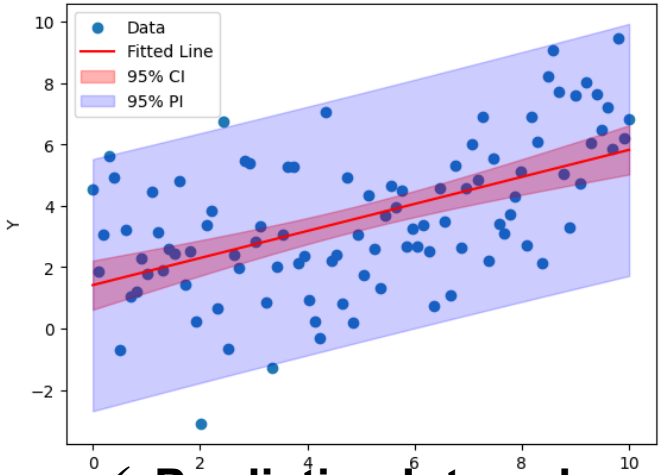


Book

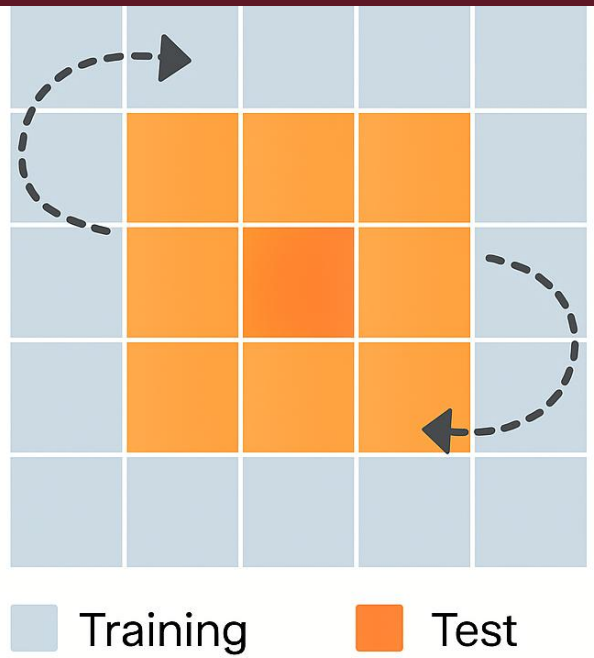
More

Detection & Quantification

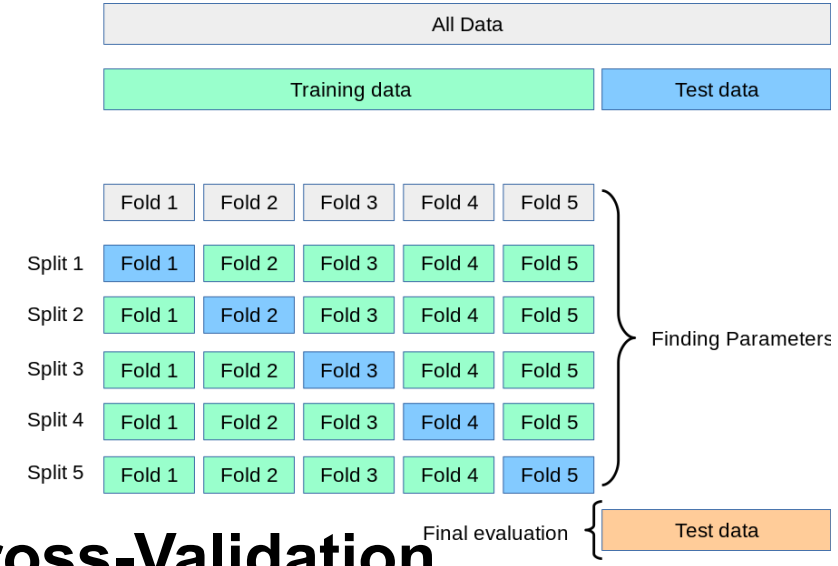
Ask and Check



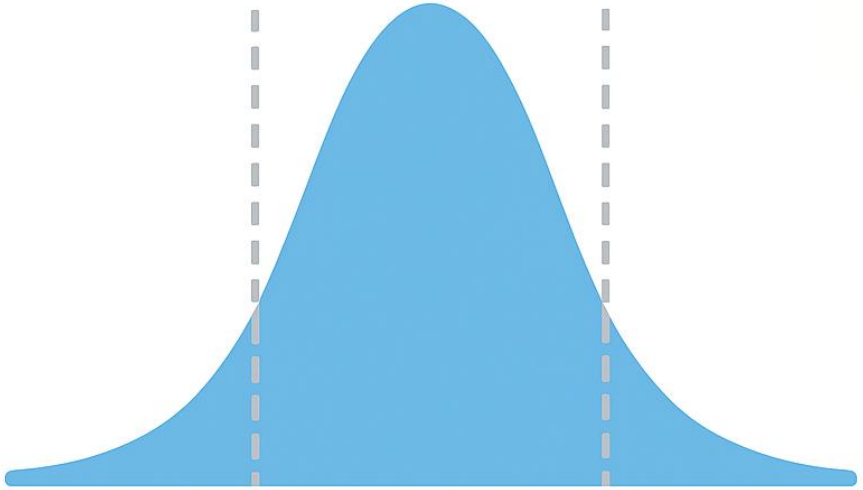
✓ **Prediction Intervals**



✓ **Spatial block, temporal block**

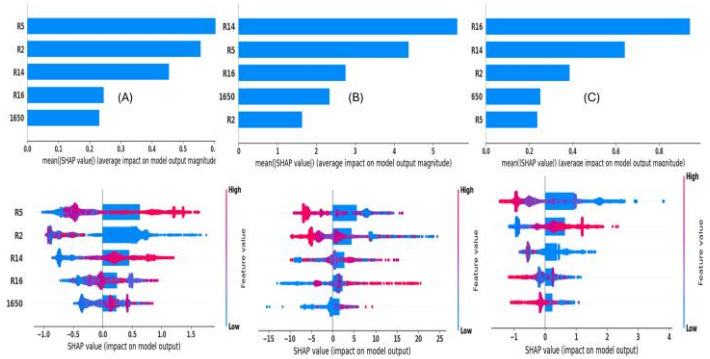


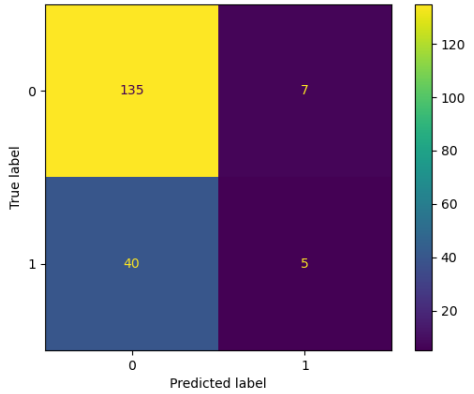
✓ **Cross-Validation**

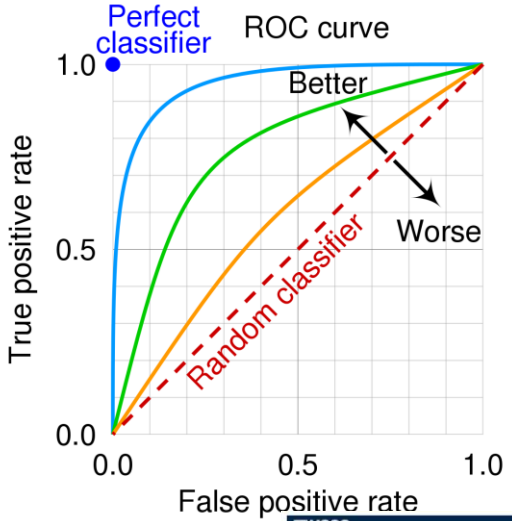


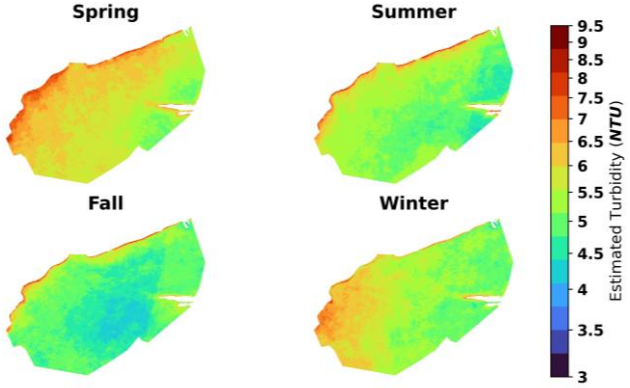
✓ **Probabilistic Models**

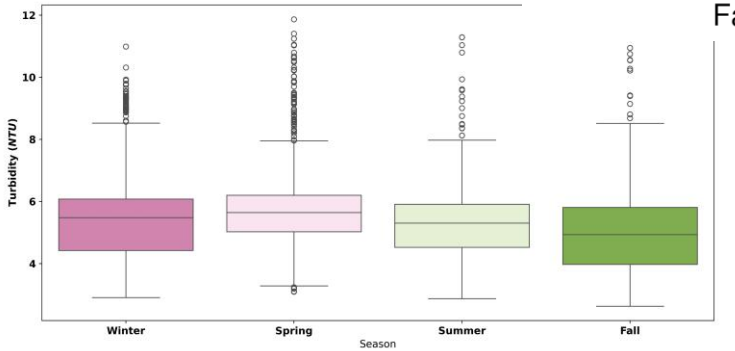
Visualizing Model Outputs

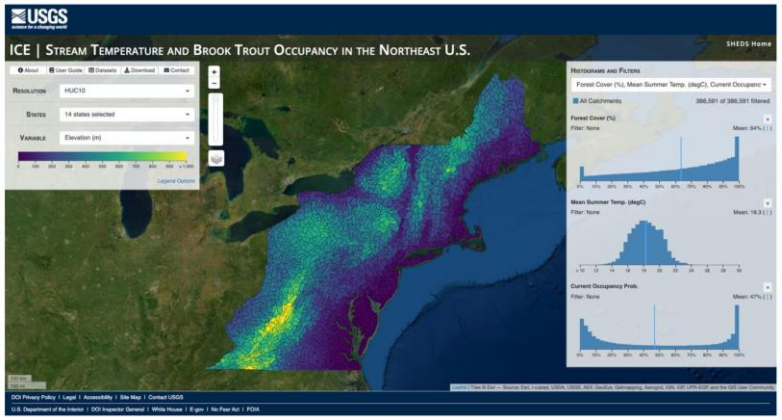


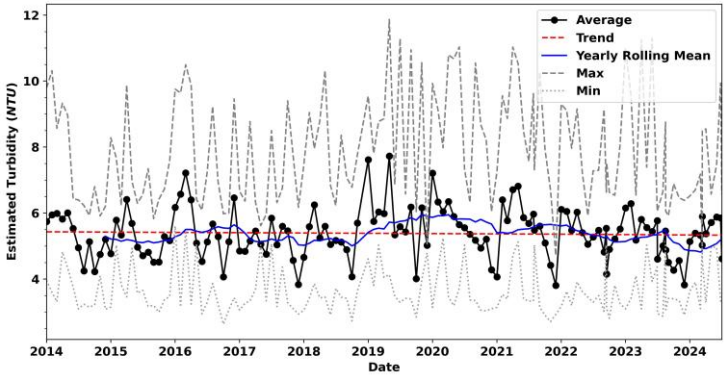


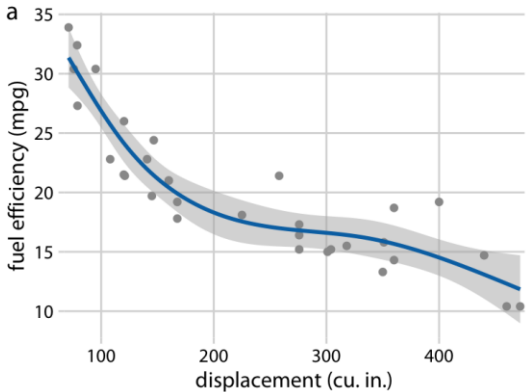


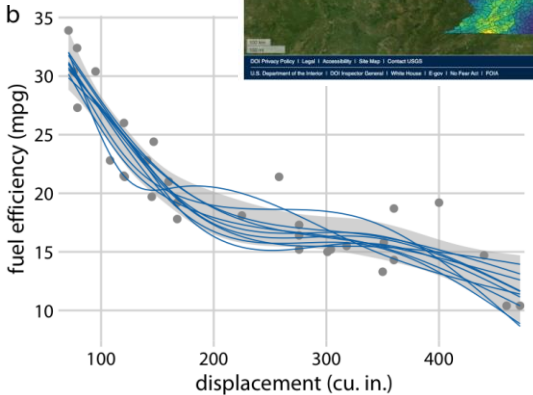












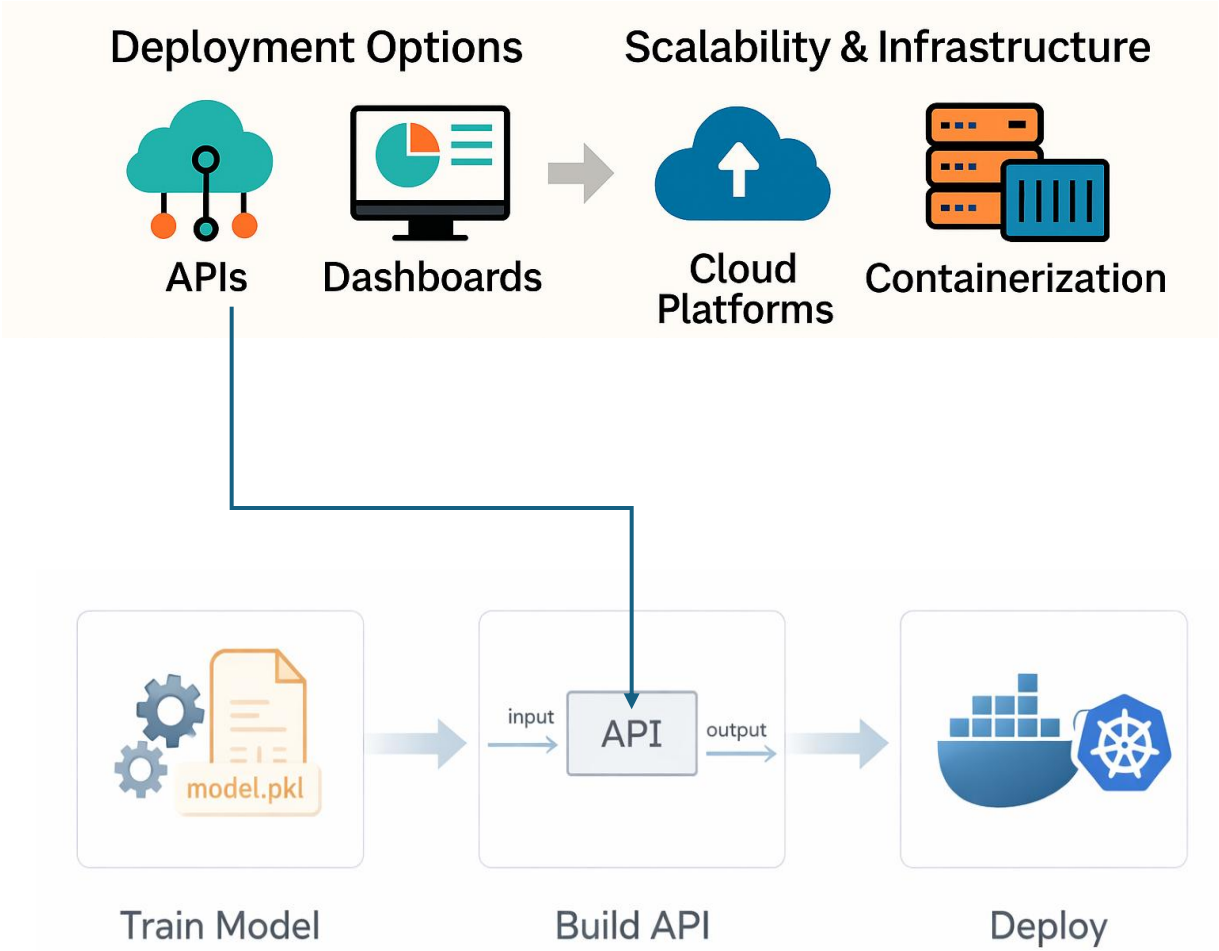
From Deployment to Scalability

Deployment Options:

- ✓ APIs: Flask or FastAPI to serve models as endpoints.
- ✓ Dashboards: Build interactive tools with Streamlit, Dash, or R Shiny.

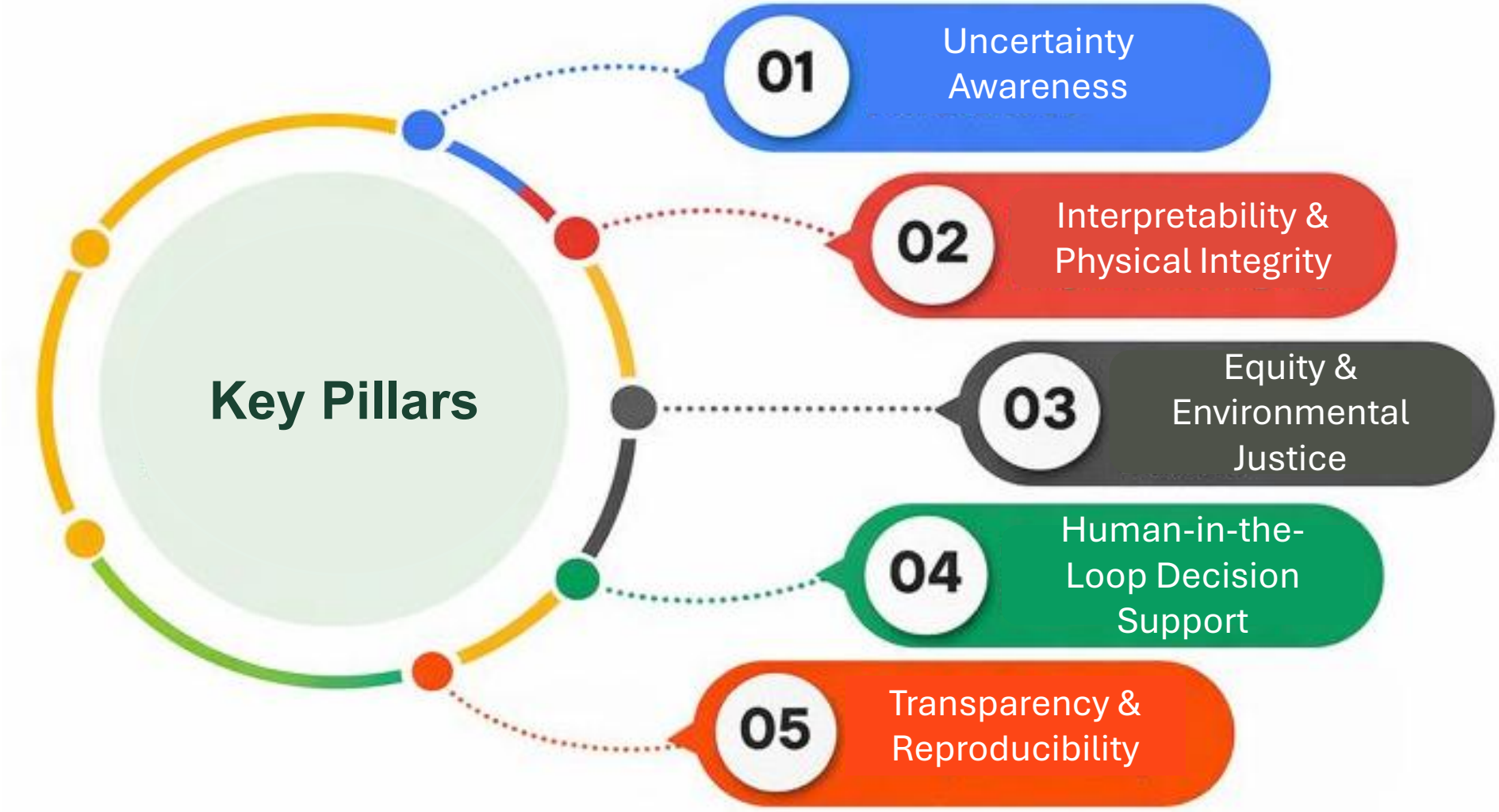
Scalability & Infrastructure:

- ➢ Cloud Platforms: AWS, Azure, or Google Cloud
- ➢ High-Performance Computing: Leverage clusters
- ➢ Containerization: Docker or Kubernetes

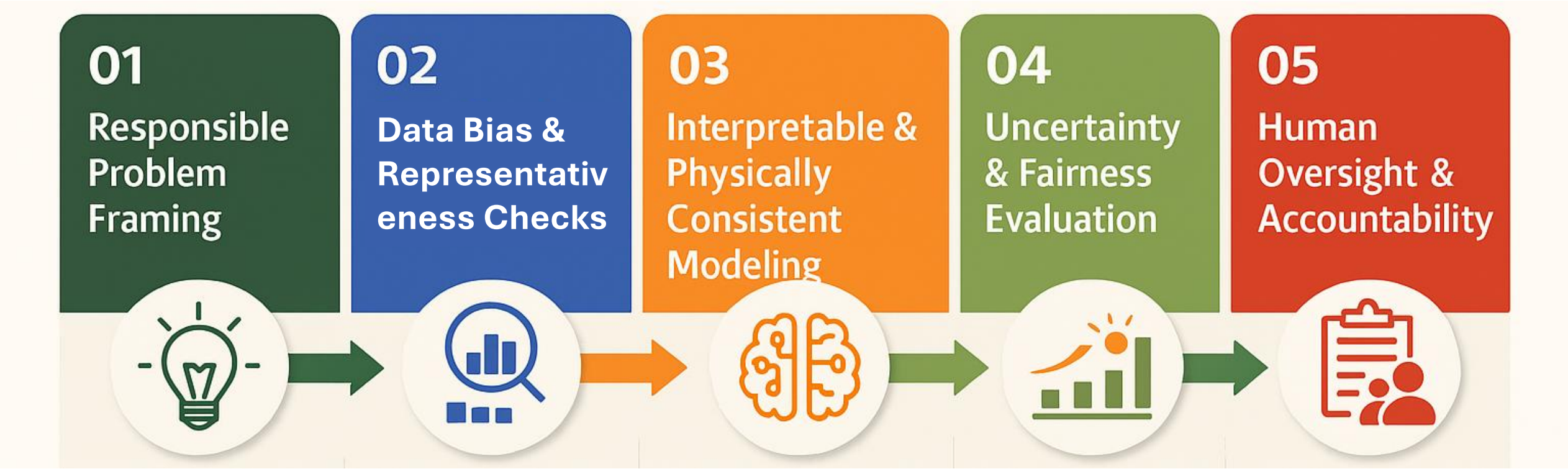


Ethical Environmental AI

Design AI systems for environmental science that are trustworthy, transparent, equitable, and decision-ready.



How to Implement Ethical AI



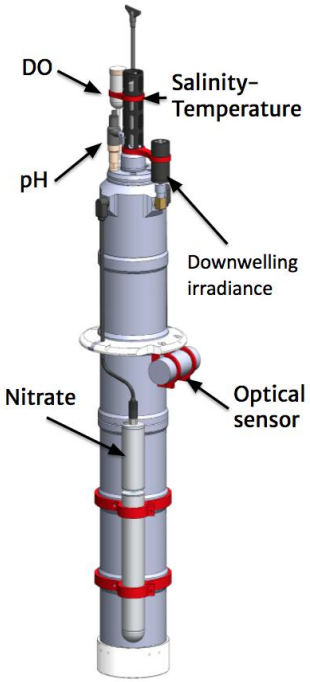
MSU AI Innovation Hub

Lab Demonstration

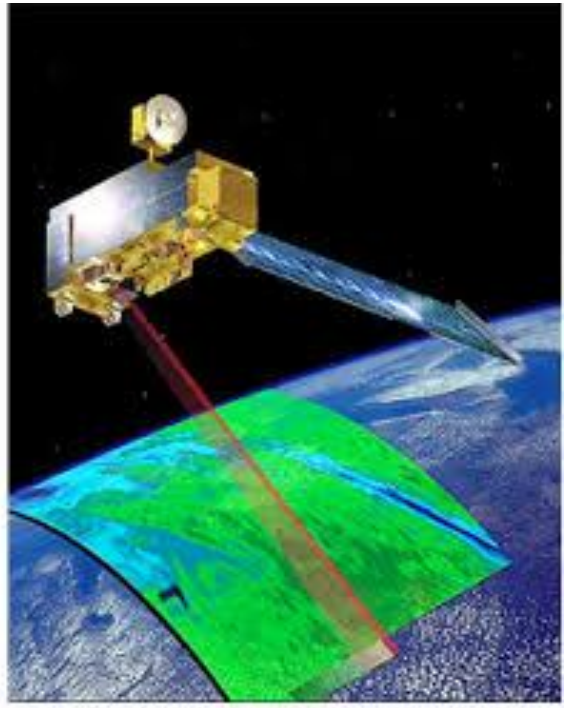
Early Detection of Hypoxia from Field and Remote Sensing Data

Program

Coding Editor



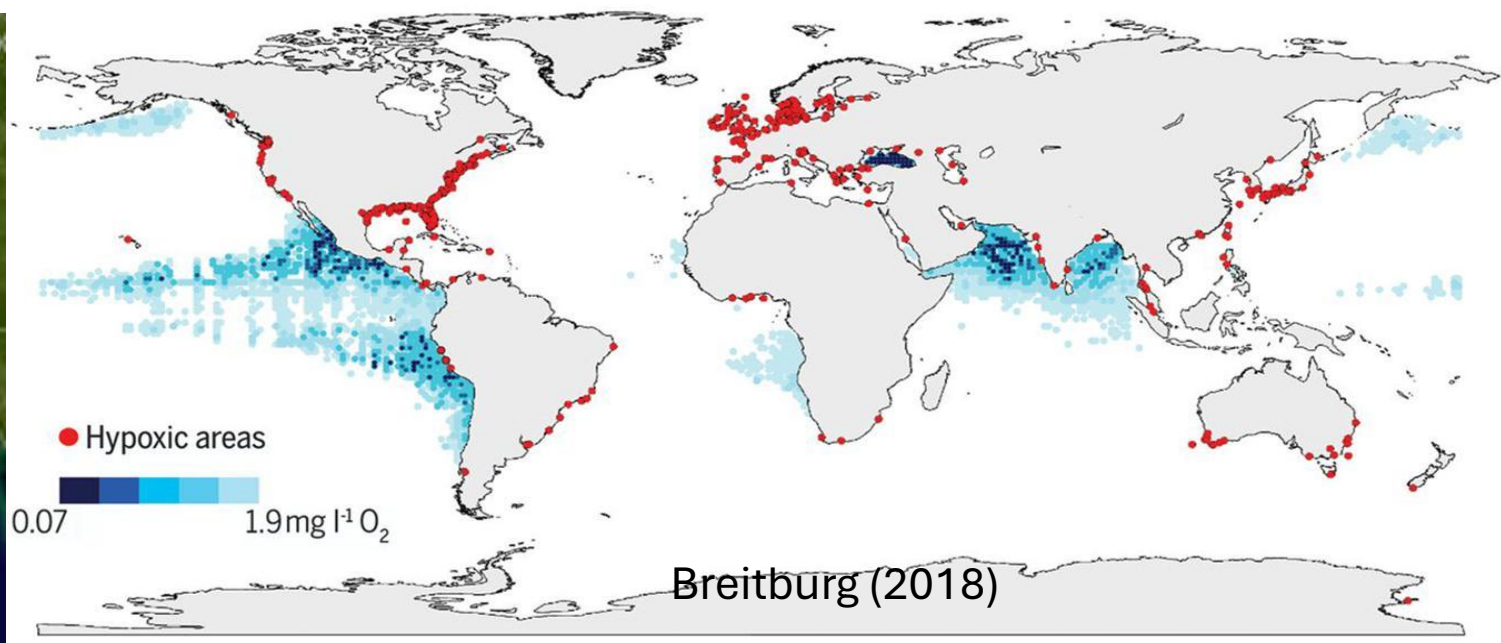
Field sensor



Satellite

Problem

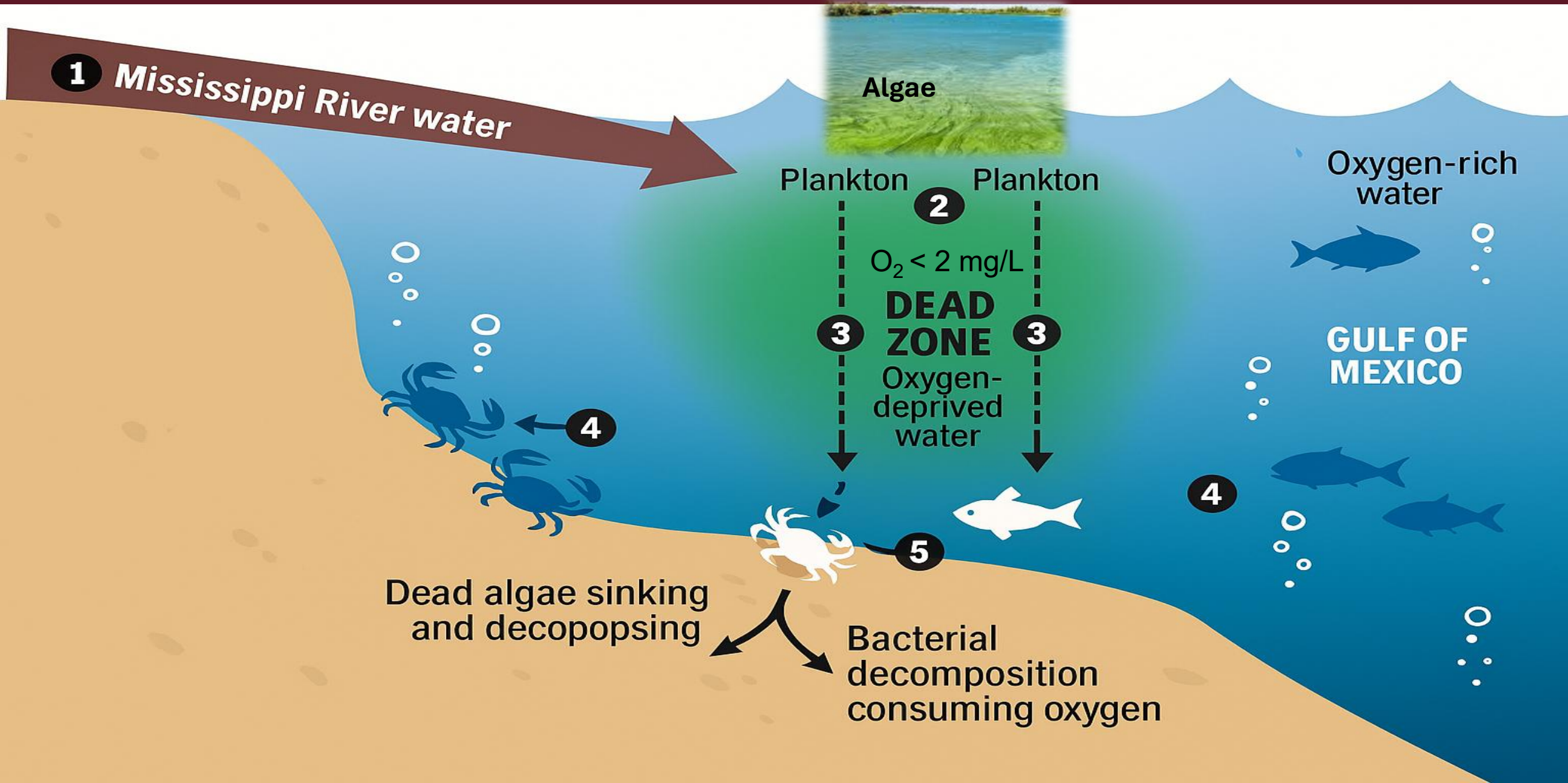
ARTISTIC RENDERIN
Mississippi


Hypoxic areas
0.07
1.9 mg l-1 O2
Breitburg (2018)

@LUMCON

a
b
c

How Hypoxia Develops



Economic and Health Impacts

Tourism: Fewer visitors; losses for charter boats, hotels, restaurants, coastal businesses

Fishing: Lower stocks, higher fuel costs, reduced profits

Seafood Markets: Price shifts

Job losses: Dependent on Gulf

@NOAA

Annual Losses:~$1 Billion

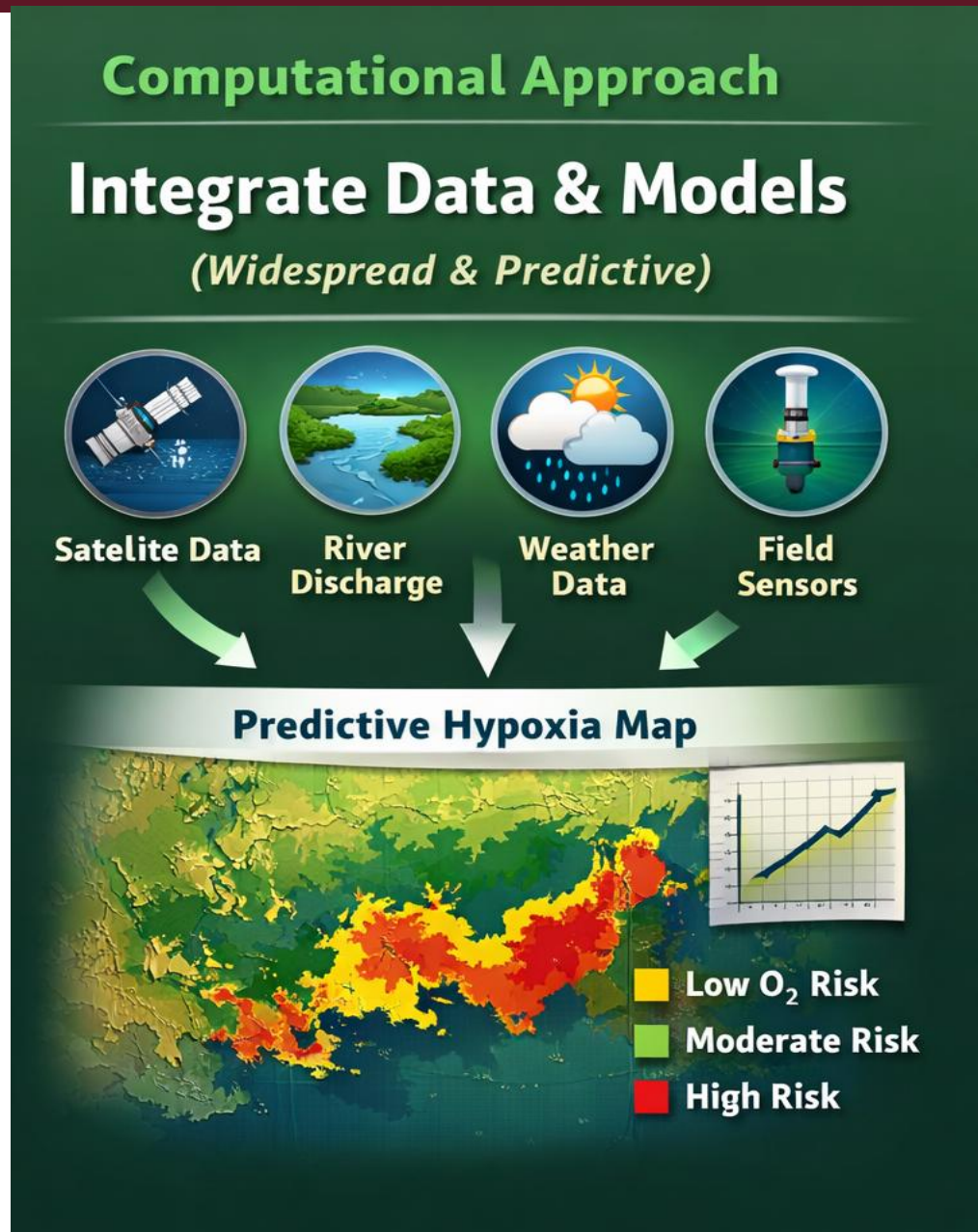
Report

Why This is a Computational Problem

1. **Scale** : Thousands of km²
2. **Frequency**: Daily Changes
3. **Multiple data sources**: Satellite data, buoys, ship surveys, weather stations
4. **Complex relationships**: Interacting variables, nonlinear
5. **Prediction needs**: Forecasts days to weeks in advance
6. **Uncertainty quantification**: Lives and economies depend on these predictions.

The decision is yours

?



The Computational Opportunity

We can

1. Use **satellite data** to monitor surface conditions across the entire region daily

2. **Integrate** field measurements with remote sensing

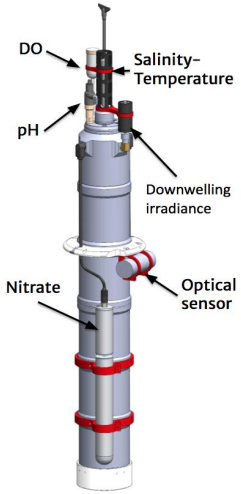
3. Build **predictive ML models**

4. Generate **forecasts** that guide monitoring and management decisions

Data Sources

1. Field Measurements

- Dissolved oxygen



2. Environmental Data

- Chlorophyll-a concentration
- Particulate Organic Carbon
- Sea Surface Temperature
- Normalized Fluorescence Line
- Remote Sensing Reflectance

Data Integration Strategy

- API
- Cloud Computing
- Sampling

- Spatial resolution differences
- Temporal frequency differences
- Units and scales
- Missing data
- Data quality

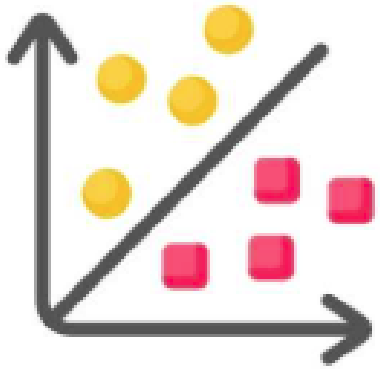
- Tabular time series
- Multispectral image
- Spatial points

- Manual
- **Write a Script**

Modeling Approaches & Decisions

Option 1: Classification

- ✓ Predict: Hypoxic (DO < 2 mg/L) or Not Hypoxic (DO ≥ 2 mg/L)
- ✓ Output: Binary prediction
- ✓ Use case: Rapid screening, alert systems

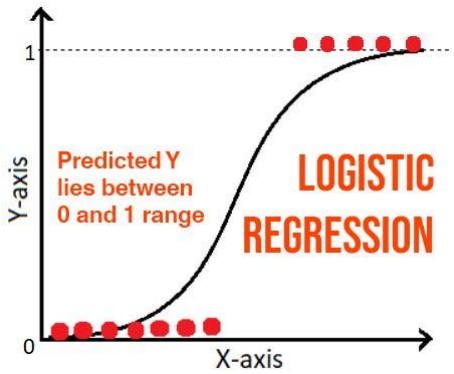


Option 2: Regression

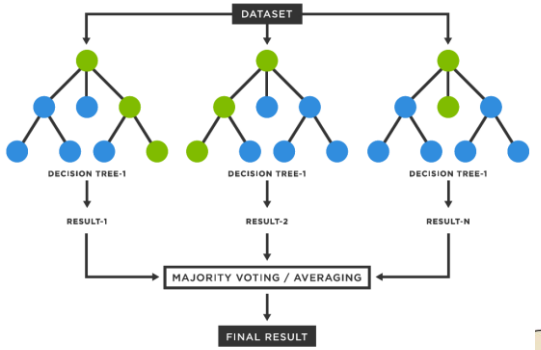
- ✓ Predict: Actual DO concentration (continuous value)
- ✓ Output: Numerical prediction (e.g., 1.5 mg/L, 3.2 mg/L)
- ✓ Use use case: Detailed monitoring, understanding severity

Model Selection

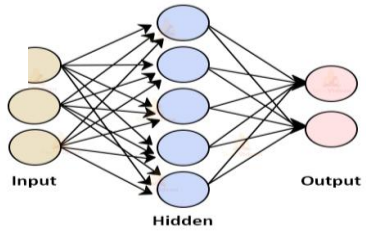
Approach 1: Logistic Regression



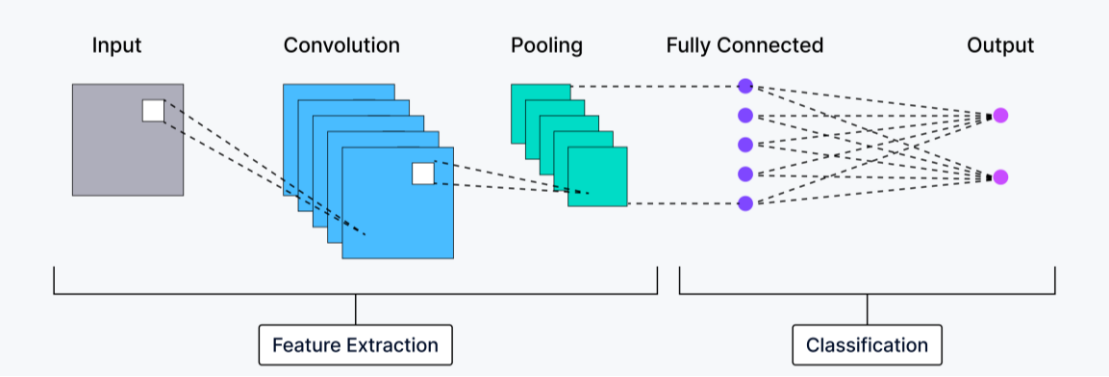
Approach 2: Ensemble Classification (RF, XGB!)



Approach 3: Artificial Neural Network



Approach 4: Convolutional Neural Network



Model Evaluation Considerations

- ✓ **Precision:** Of locations we predicted hypoxic, how many actually were?

  Important for: Avoiding false alarms

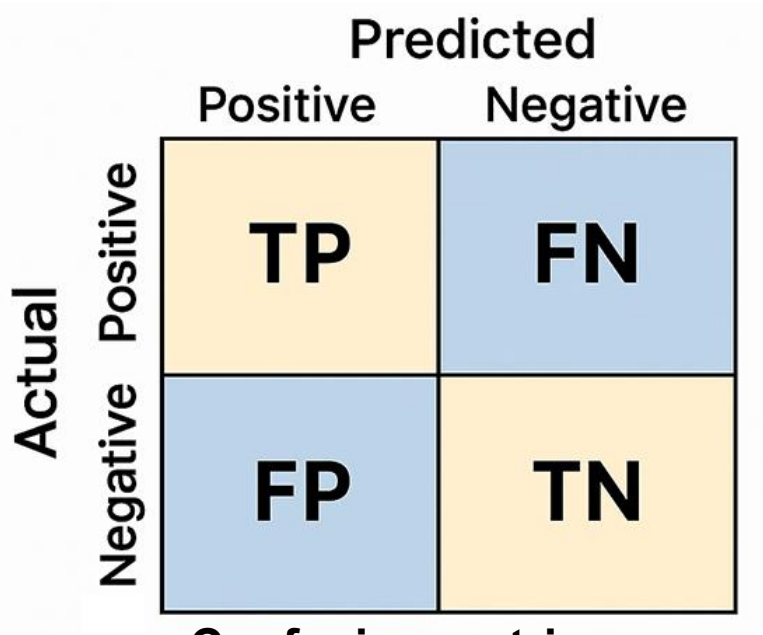
- ✓ **Recall (Sensitivity):** Of actual hypoxic locations, how many did we detect?

  Important for: Not missing critical events

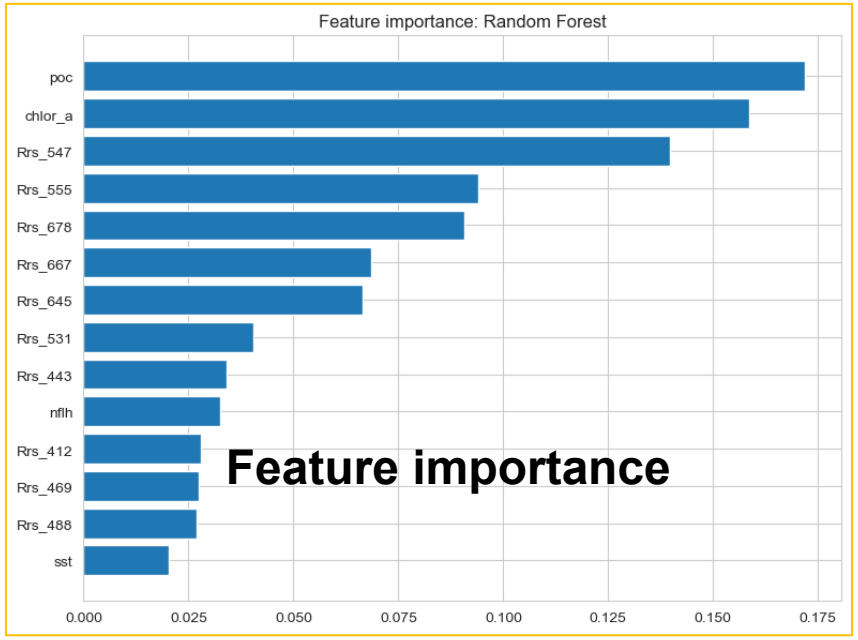
- ✓ **F1-Score:** Harmonic mean of precision and recall

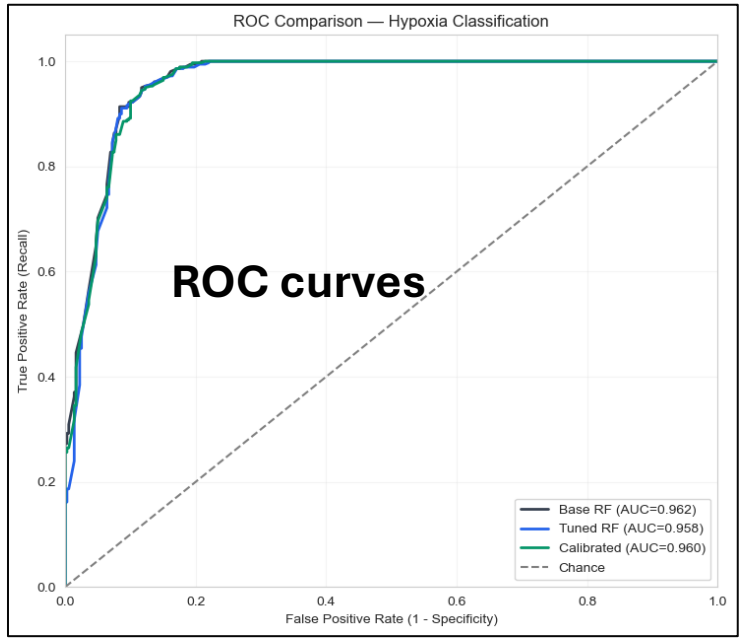
  Balanced metric for imbalanced data

Interpretation & Diagnostics

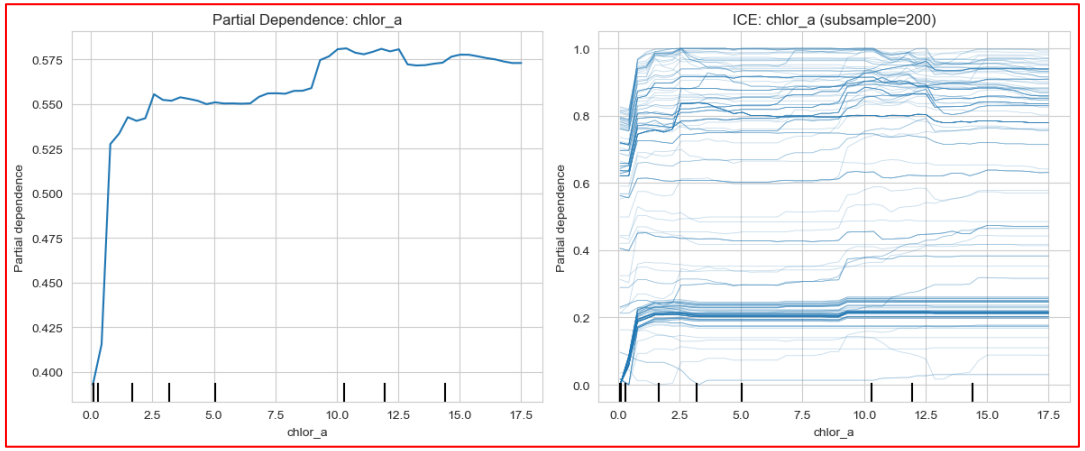


Confusion matrix





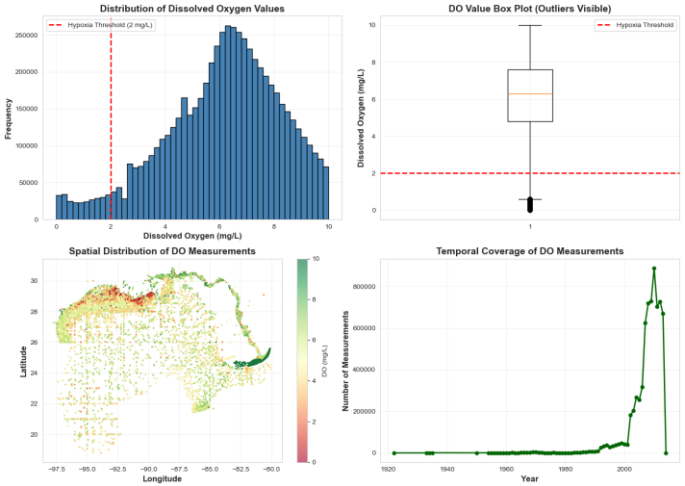
Error analysis

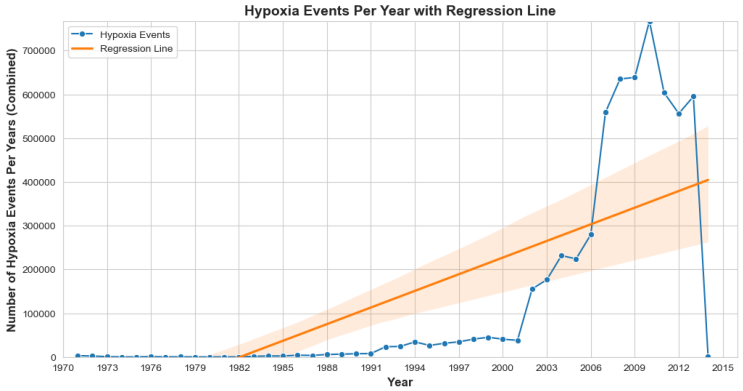


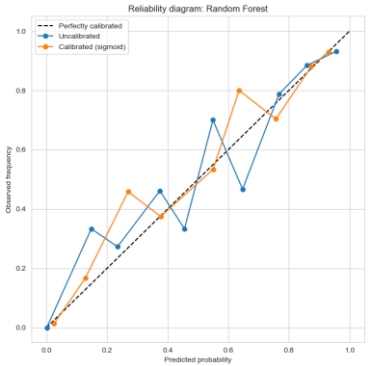
Partial Dependence Plot

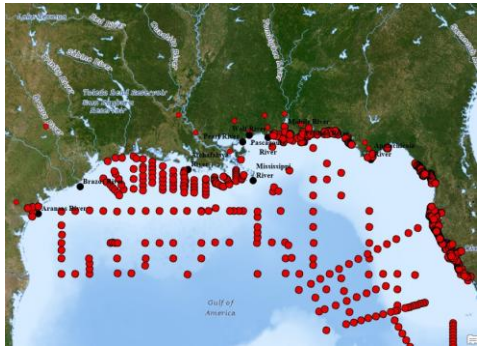
Visualization

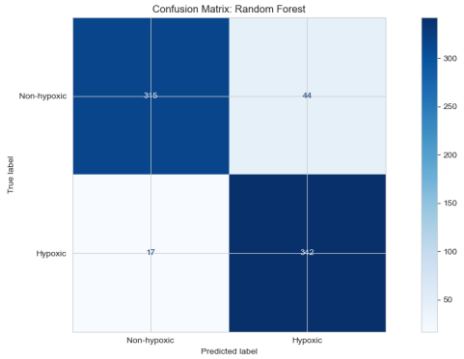
- Sample Distribution of DO
- Seasonality of Hypoxia
- All variables
- Performance Metrics
- Precision-Recall Tradeoff Curve
- What Drives Predictions?
- Partial Dependence Plot
- Prediction Map – Monthly/ Yearly
- Time Series of Prediction Maps
- Prediction Confidence

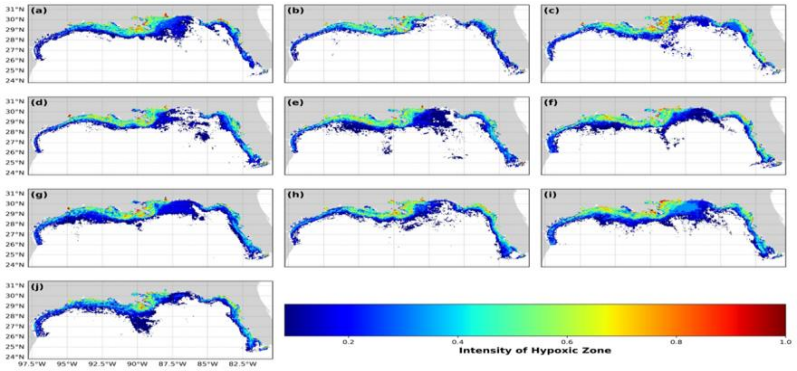


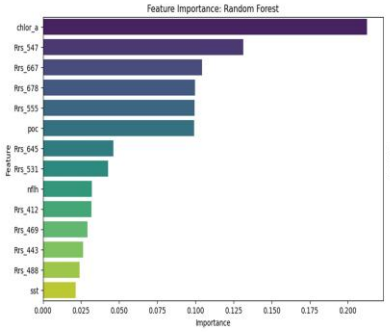


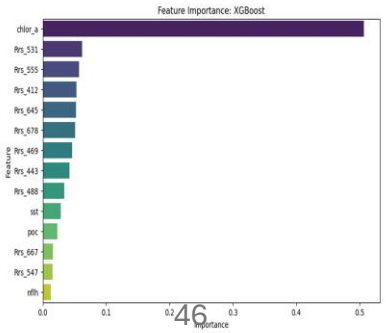












End-to-End Machine Learning Pipeline

Problem Definition & Prediction Target
- Target: DO
- Task: Classification
- Spatial & Temporal Unit

Data Sources & Resolution Tradeoffs
- In-situ
- Satellite imagery
- Ancillary data

Geospatial & Temporal Alignment
- Consistent WGS84
- Matching satellite & field dates

Data Integration & Feature Engineering
- Balanced
- Scale

Model Development & Validation
- Baseline Model
- Choice of object
- Tune

Model Interpretation & Uncertainty
- Feature importance
- Partial dependence
- Prediction uncertainty

Spatial Prediction & Temporal Dynamics
- Prediction map
- Uncertainty map
- Seasonal & Annual

Model to Decision
- Early Warning
- Risk comparison
- Communication

Update Model
- New Data
- New Sensor
- Drift detection

Future Research
- Physics informed
- Deep Learning
- More Context
- Transfer Learning

Ethical Considerations

When models inform **real-world decisions**

Questions to consider:

1. What accuracy is "**good enough**" when livelihoods are affected?
2. How do we **communicate uncertainty to non-experts**?
3. Who has **access** to predictions (equity)?
4. What happens when **models are wrong**?
5. How do we maintain **trust** and **transparency**?

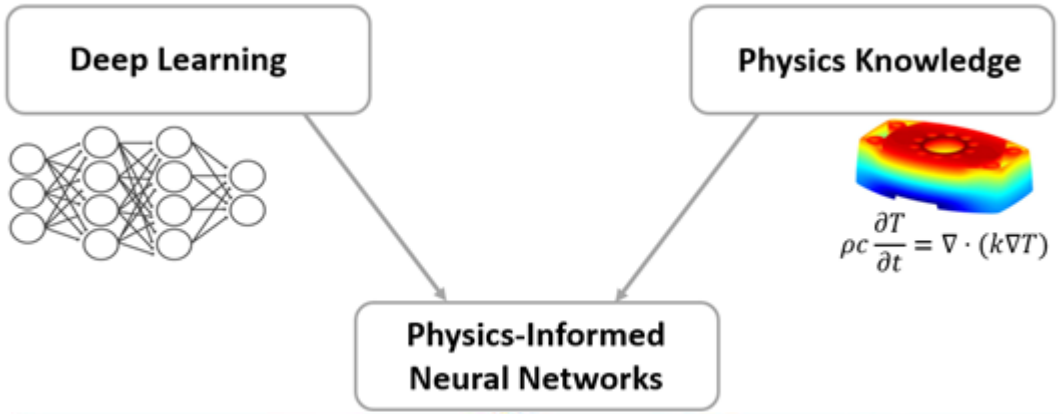
Lab Time!

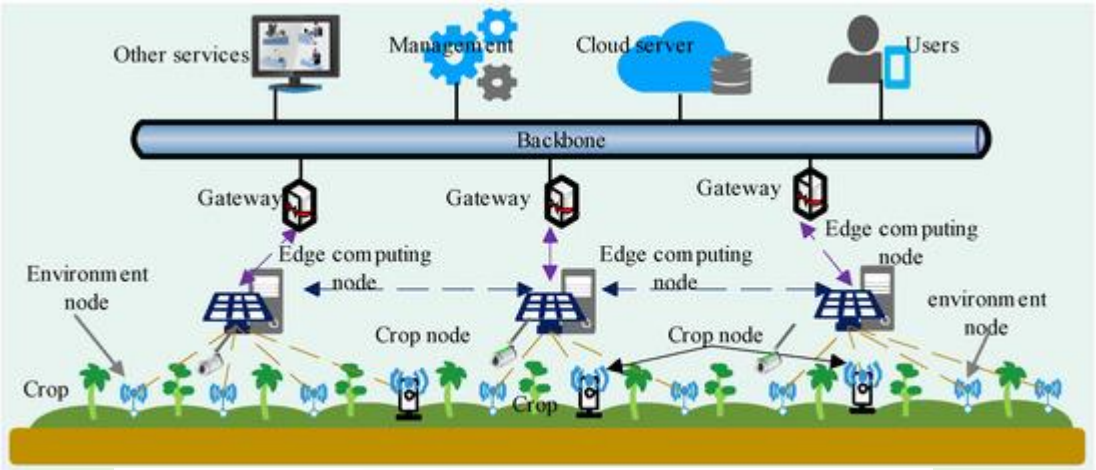
Q&A

Code+ Data Link

What's Next for Environment?

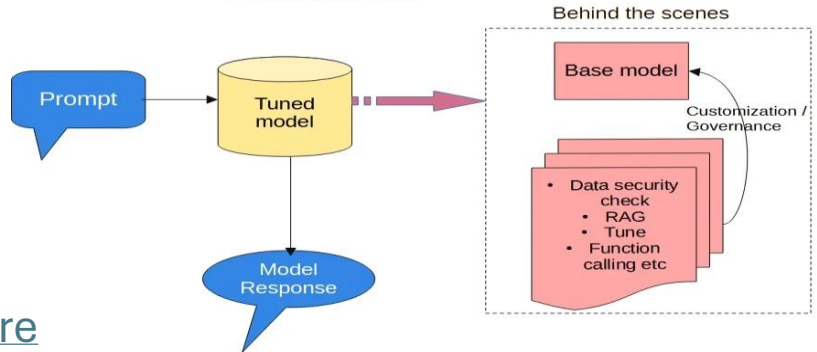
- Integrate ML + Scenario + Projection + Simulation + Optimization + Decision Support
- Physics-informed machine learning
- Edge computing + IoT sensor networks
- Real-time Earth system analytics
- Generative AI for scenario modeling





Generative AI

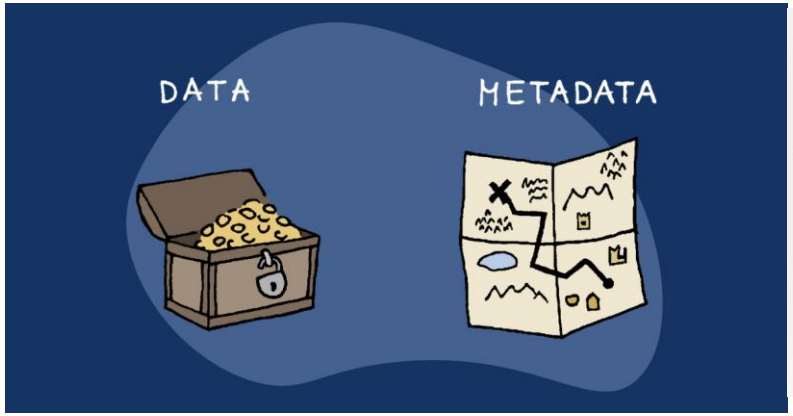
@

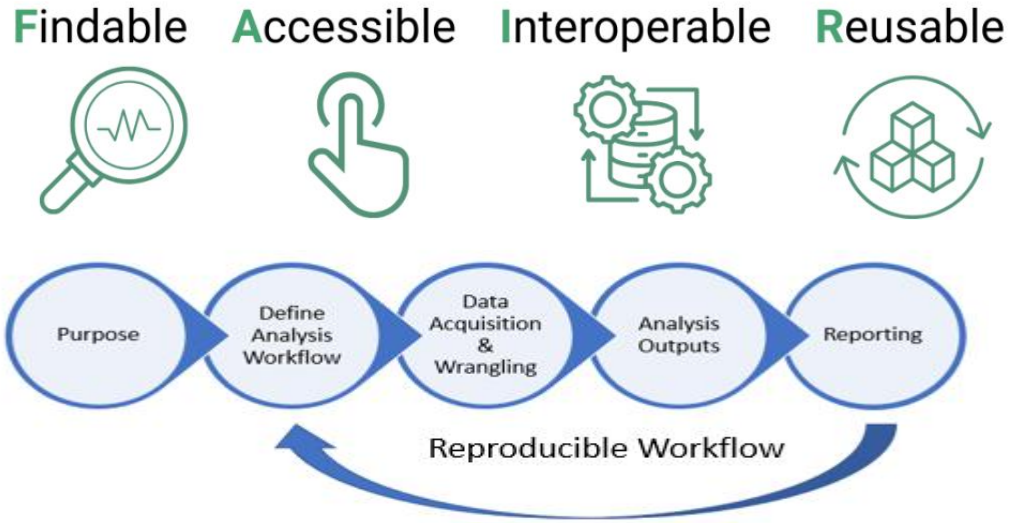


Link

More

Best Practices for Environmental Data Science





awesome-gee-community-catalog

HYDROSHARE

GitHub

LSHTM Data Compass

zenodo

FLOW Repository

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

git

figshare

Open Science Framework
A scholarly commons to connect the entire research cycle

UK Data Servic

DRYAD

GEO
Gene Expression Omnibus

NCBI GenBank

PlasmoDB
Plasmodium Genomics Resource

Key Takeaways

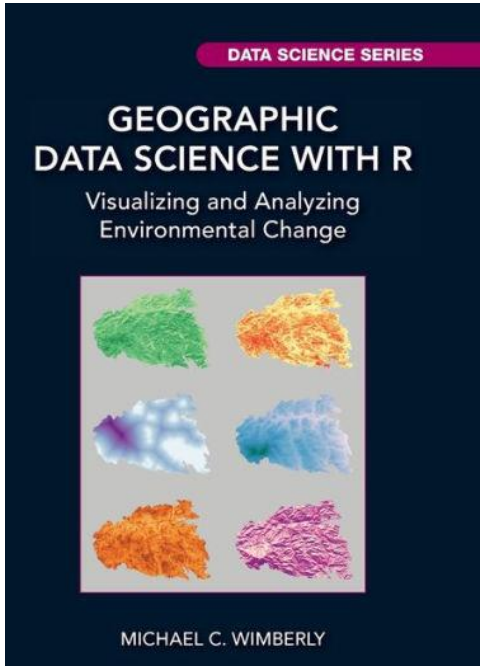
Robust Data

Honest Eval

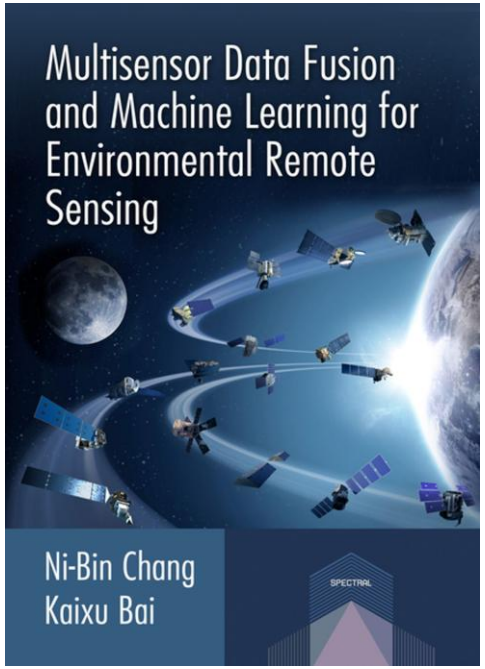
Domain Insight

Ethics First

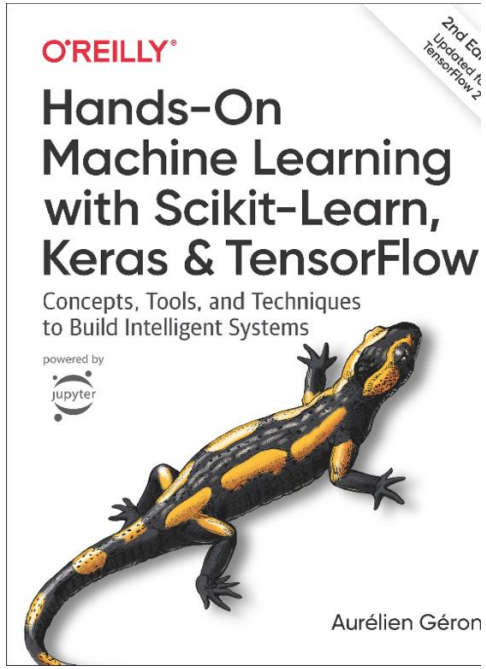
References



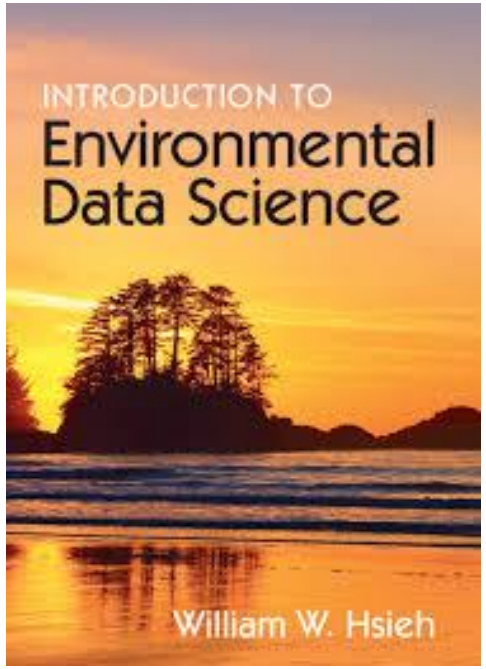
1*



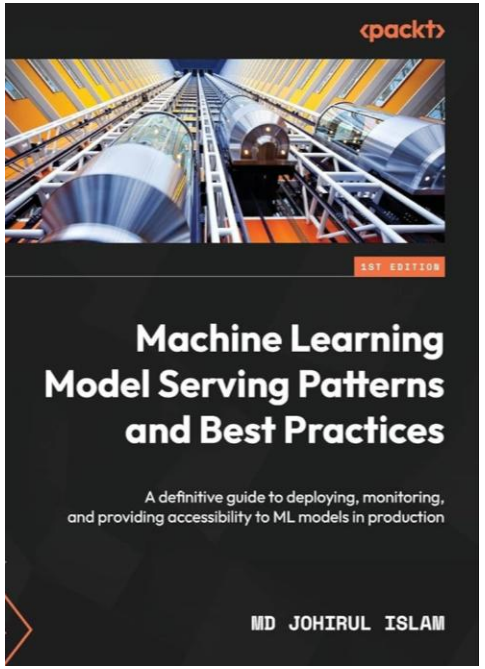
2*



3*



4*



5*

Questions + Discussion