# Interpretability

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

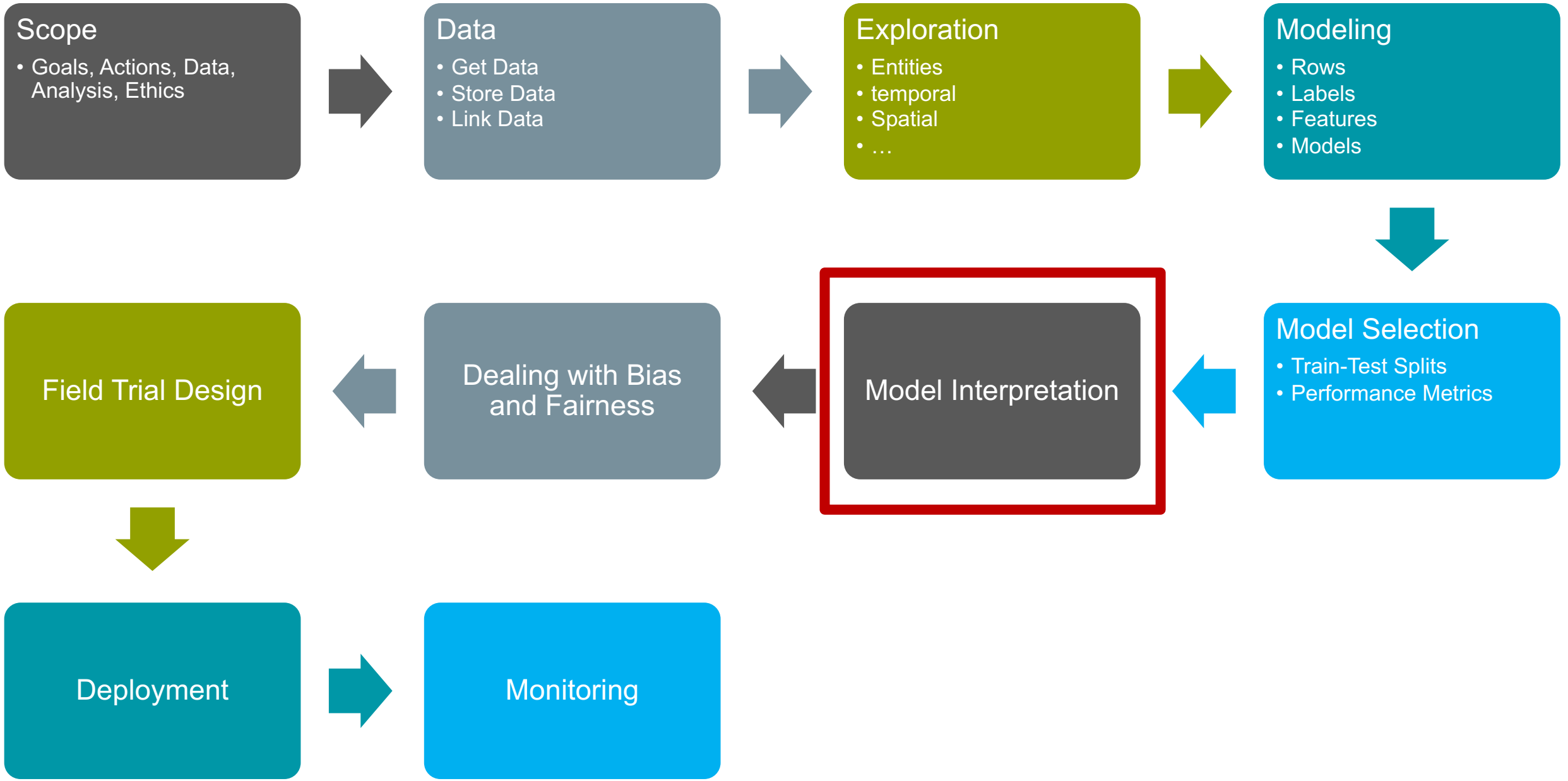
# Going Forward: Modules 2 and 3

- Module 2 & 3 classes focus on 2-3 methods/approaches each day
- Each group responsible for applying one approach from each class (may implement from scratch or use existing packages)
- “Extended Abstract” (3-4 pages) at end of the module comparing these results
- Presentations and “discussants” (15 minutes)
  - We will assign one of each, other methods up to you
  - Method overview and preliminary results

# Reminders

**This week:**

**Coming up next week:**


Scope
• Goals, Actions, Data, Analysis, Ethics
Data
• Get Data
• Store Data
• Link Data
Exploration
• Entities
• temporal
• Spatial
• …
Modeling
• Rows
• Labels
• Features
• Models
Model Selection
• Train-Test Splits
• Performance Metrics
Model Interpretation
Dealing with Bias and Fairness
Field Trial Design
Deployment
Monitoring

# Why do we want Interpretability?

# What we'll cover today

- Why we want ML models to be interpretable?
- Real-world use cases
- Types of Interpretability

# Why do we want Interpretability?

- Debugging the Model
- Improving the performance of the system
- Creating and increasing trust and adoption
- Selecting appropriate actions/interventions
- Legal Recourse

# Deeper Dive into Use Cases

| Use Case | User | Task | What type of method? |
|---|---|---|---|
| Debugging | | | |
| Improving performance | | | |
| Trust | | | |
| Interventions | | | |
| Recourse | | | |

# Deeper Dive into Use Cases

| Use Case | User | Task | Global or Local |
|---|---|---|---|
| Debugging | ML Developer | Sanity check and improve “correctness” | Both |
| Improving performance | Action-taker | Decide to agree or override | Local |
| Trust | Policymaker/Action-Taker | Model use => better outcomes | Both |
| Interventions | Action-Taker | Improve outcomes | Local |
| Recourse | Individual(s) affected | Provide information that the individual can understand and ideally change | Local |

# What do we want from a global explanation?

# How to interpret specific models

- Decision Trees
- KNN
- SVMs
- RFs
- NNs
- Are there other methods that are easier to interpret/understand out of the box?

# What we do want from a local explanation?

# How would we validate if an interpretability method is effective?

| Use Case | Evaluation Methodology |
| --- | --- |
| Debugging | |
| Improving performance | |
| Trust | |
| Interventions | |
| Recourse | |