ML Formulation and Baselines

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

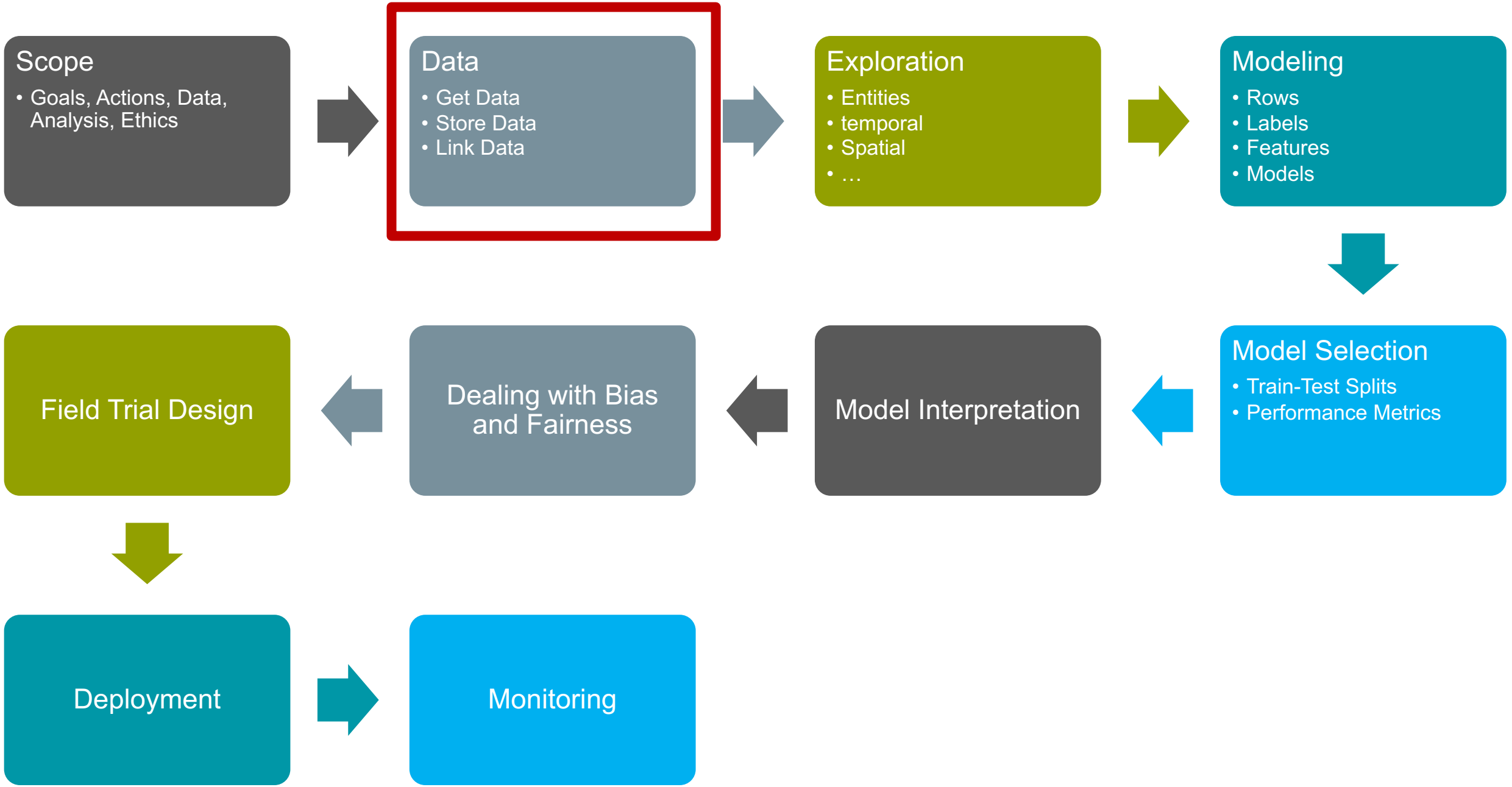
Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Things to remember

- **Should be able to access: github, server, database**

- Due Next Week:
  - Monday – project update assignment (Formulation and Baselines)
  - Tuesday – weekly feedback form

- Readings for Tuesday


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

How do we not compare every pair?

- How do we avoid looking at |A| * |B| pairs?
- *Blocking:* choose a smaller set of pairs that will contain all or most matches.
  - Simple blocking: compare all pairs that “hash” to the same value (e.g., same Soundex code for last name, same birth year)
  - Extensions (to increase *recall* of set of pairs):
    - Block on *multiple* attributes (soundex, zip code) and take union of all pairs found.
    - *Windowing:* Pick (numerically or lexically) *ordered* attributes and sort (e.g., sort on last name). The pick all pairs that appear “near” each other in the sorted order.

Common reasons for mismatches

- Case (capital, lower case, etc.)
- Nicknames
- Prefixes
- Suffixes
- Initials
- Punctuation
- Spaces
- Digits
- Transpositions
- Abbreviations

Discussion Topic

What are downstream ethical issues when dealing with errors in record linkage?

Moving from Scope to Analytical Formulation

Turning the project goals/scope into an ML problem

- Scoping defines the goals and approach at a high level, the **analytical formulation** maps this scope to an ML problem and analytical approach
- Should be as detailed and specific as possible, making it possible to code it without ambiguity
- The analytical formulation should be guided by — and map back to — how the system you're building will be deployed and used

Decisions we need to make

- What type of analysis are you doing?

Decisions we need to make: analytical approach

- Description
- Classification
- Detection
- Prediction
- Optimization
- Causal Inference

Decisions we need to make

- What type of analysis are you doing?
- What are the relevant entities? How do you identify the cohort?

Decisions we need to make: cohort definition

- Every entity that exists?
- “Active” entities?
- Event-based?
  - Making predictions when the events occur?
  - All entities that have had an event in a certain time window?

Decisions we need to make

- What type of analysis are you doing?
- What are the relevant entities? How do you identify the cohort?
- How do you define the outcome/label that you care about?
- How far into the future are you trying to predict?

Analytical Formulation Examples

Baselines

What is the appropriate comparison to evaluate effectiveness of your ML model?

Baseline Options

- Common Sense
- What they do today
- What they could do today easily (without any or very simple ML involved)
- Prior/Base Rate
  - What expected value would you get if you just choose at random (based on the data distribution)?

Baseline Considerations

- How much better than baselines does our system need to be in order to deploy?
- Important to compare performance against the base rate/prior, but this prior rarely represents a “common sense” baseline
- Good baselines should provide an ordering to sort the entities
  - Heuristic rules (or shallow decision trees) might reflect current practice, but can yield a small number of unique scores with lots of ties
- In many real-world problems, a good baseline can be difficult to beat

Baseline Examples

CASE STUDY

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Things to remember

- **Should be able to access: github, server, database**

- Due Next Week:
  - Monday – project update assignment (Formulation and Baselines)
  - Tuesday – weekly feedback form

- Readings for Tuesday