Data Acquisition, Storage, and Linkage

Rayid Ghani and Kit Rodolfa

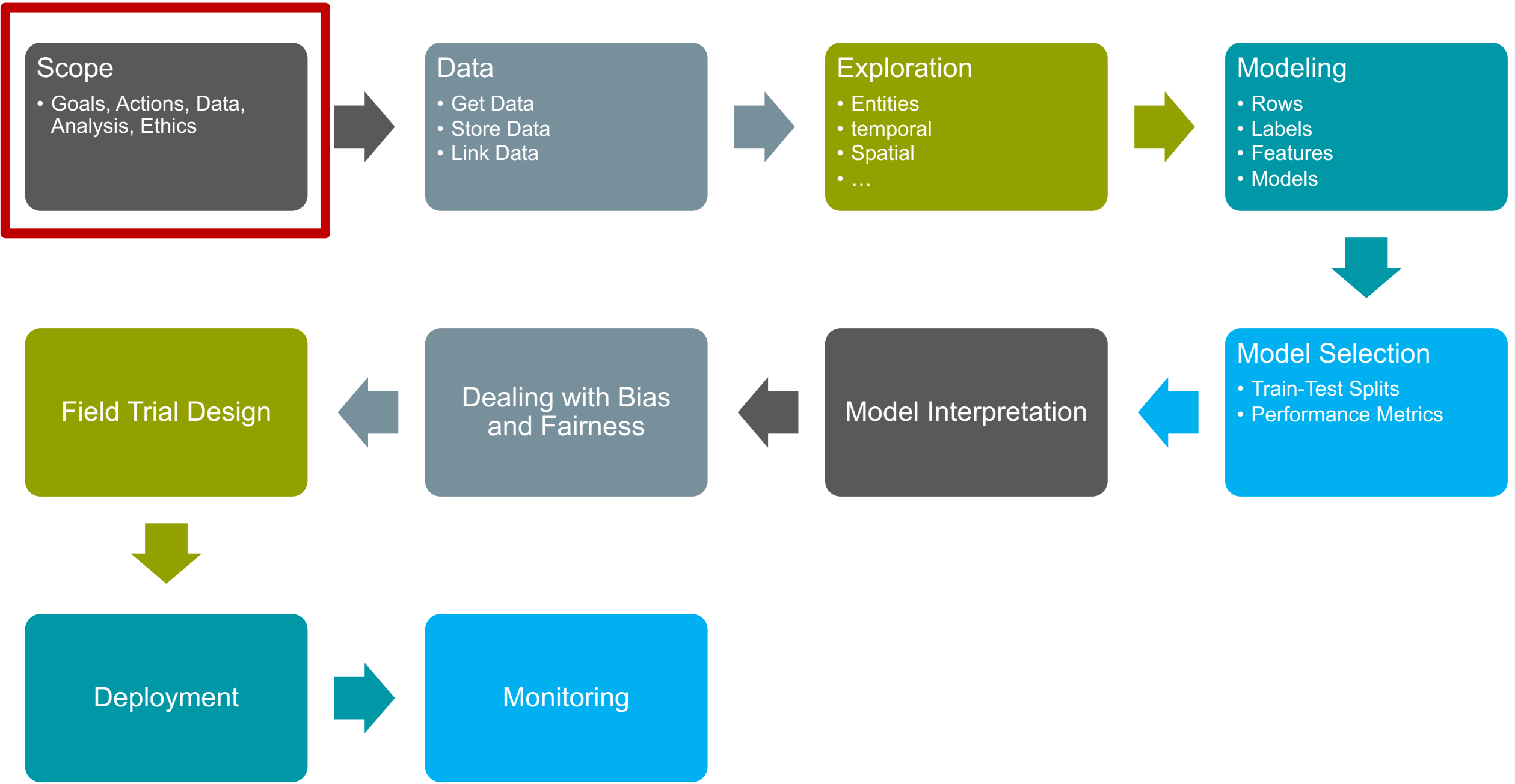
Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Things to remember

- **Make sure and check that you can access your project data today**
- Thursday:
  - Readings (Obermeyer et al; Passi & Barocas)
- What's coming next week:
  - Assignment due Monday
    - Project Formulation and Baselines
  - Readings


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

Creating a more equitable society

- How do we **define** equity?
- How do we **detect** inequity?
- How do we **increase** equity?

Data and AI Ethics Issues

Privacy

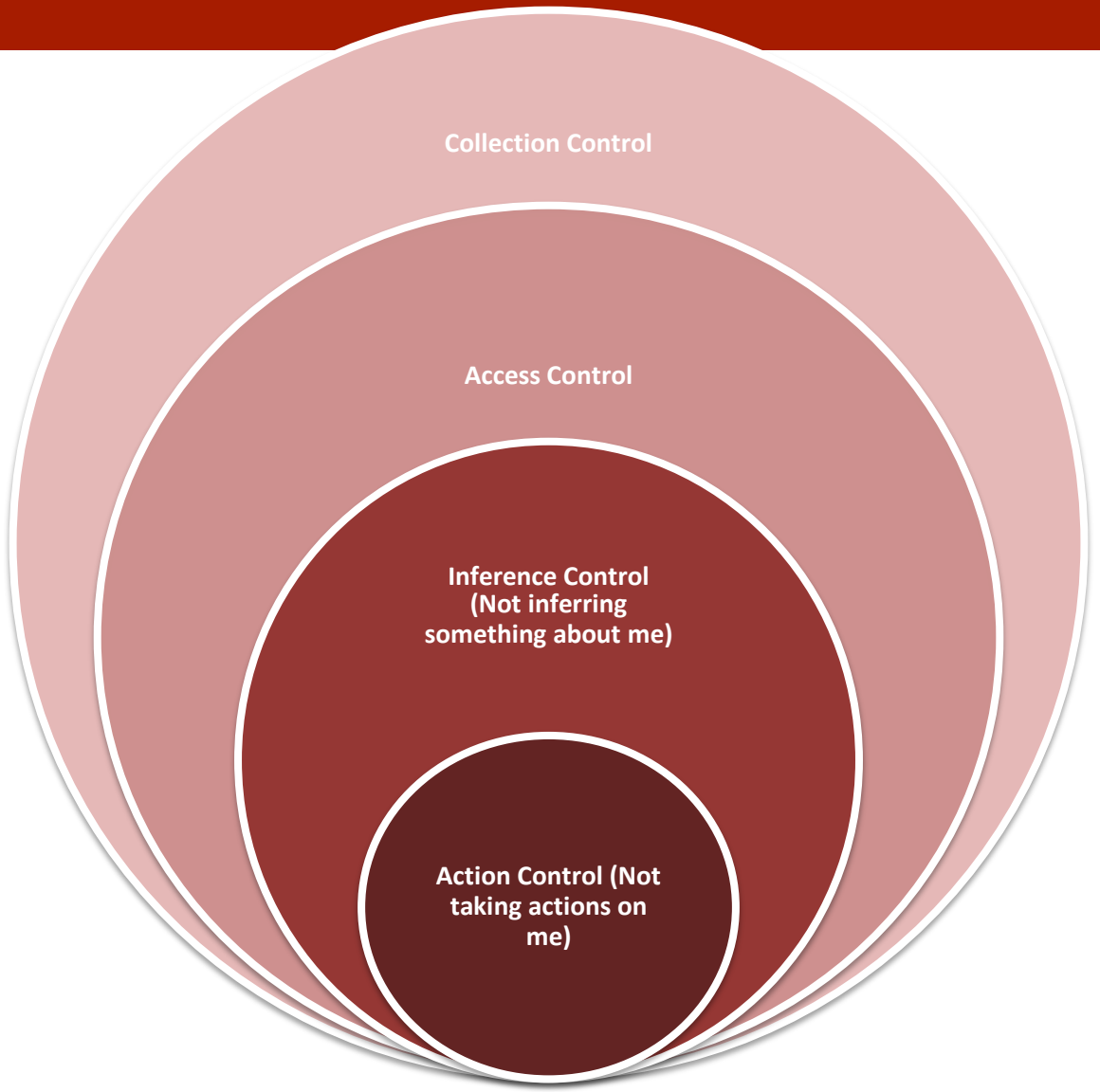
Data Ownership

Bias, Equity, & Fairness

Transparency

Trustworthiness and Accountability

Levels of control



Data Ethics Questions

- Are you using data for purposes it's intended for?
- How are you protecting the data?
- Do the people who "own" the data know you're using it?
- Do you have their permission? How was it obtained?
- What actions are you taking on individuals based on this data?
- Do the people you're targeting know why and if they're being targeted?
- What recourse do they have?
- Would it make the front page of the national newspaper if they found out what you're doing?

A Few Things to Remember

- Don’t be afraid to ask naïve questions
- Spend time discussing goals and metrics – don’t forget equity as a goal
- Understand what the current process/solution is
- Communication is critical – before, during, and after
- We need to make sure that we tackle these problems responsibly and ethically
- Data and ML does not solve problems, people do. Is what you’re doing helping solve the problem?

Project

Support a classroom.
Build a future.

Teachers and students need your support more than ever. Get crayons, books, cleaning supplies, technology, and more to help students get back to learning.

See classroom projects

Our efficiency and transparency have earned us the highest rating on Charity Navigator.

DONORS CHOOSE

Working with Your Project Team

- Spend some time up front figuring out how to work as a team and work styles
- Everyone should participate in **all** aspects of project
  - No individual should do the majority of the coding, report writing, etc.
- However, you should divide up different pieces of the work
  - For instance, working on different parts of the pipeline code in parallel or splitting up sections of the report

Working with Your Project Team

- We're providing class time to make it easier to coordinate with your group and get feedback from peers and instructors, but you won't be able to complete the project just in this allotted time

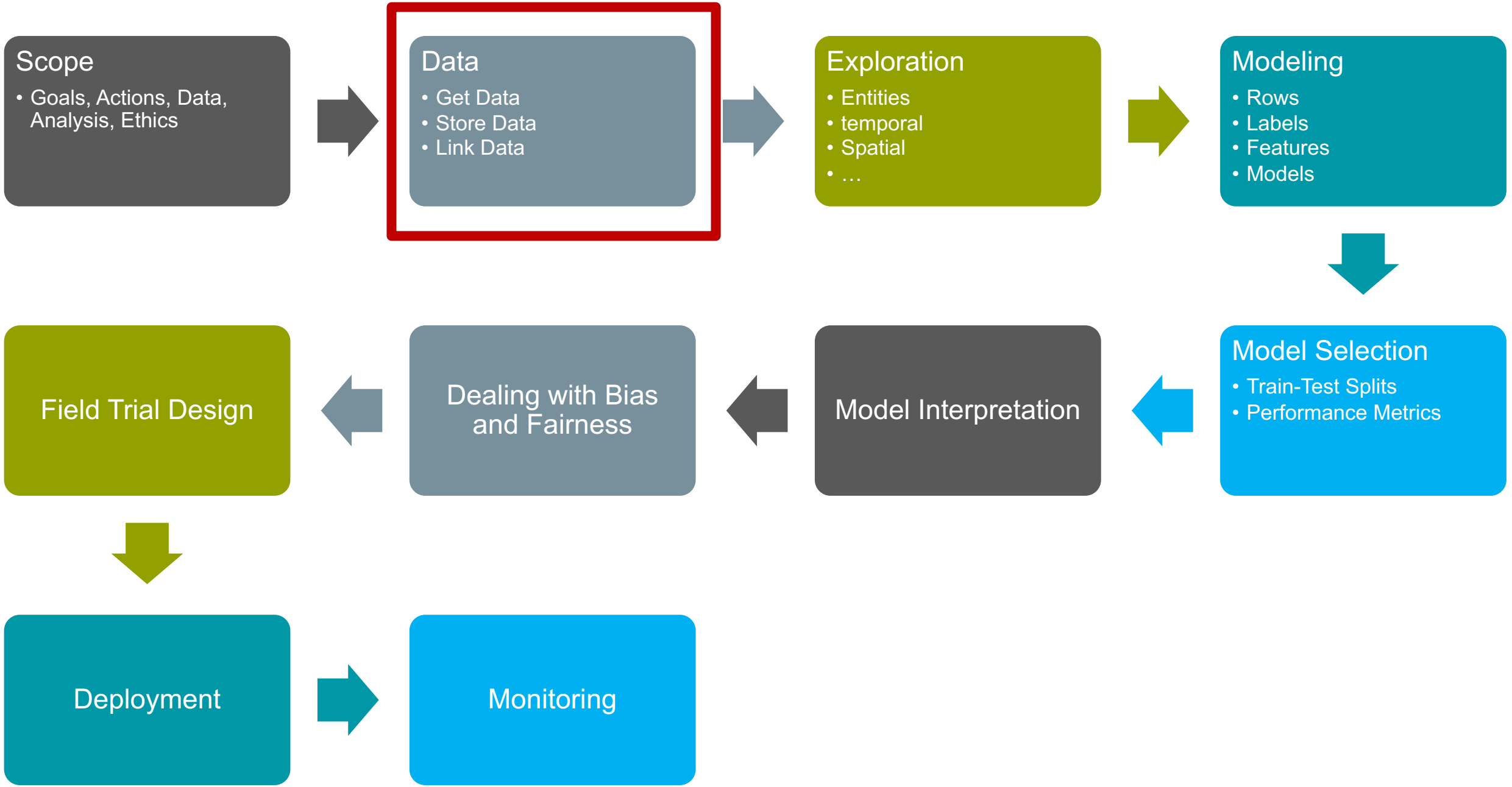
- Tools for coordination:
  - Slack: we'll create group-level and project-level channels. Additionally, feel free to use group DMs and video calls with your group to coordinate as well.
  - Many good free options for task tracking/management: github issues or project boards, trello, etc.

Working with Your Project Team

- Active participation in the group work throughout the entire semester is required by all the team members, and a very large component of your grade

- Collaboration is encouraged, both within and across teams! Learning from each others' strengths is a big benefit of group work, and you should feel free to discuss strategies and approaches with other teams (i.e., it's not a competition)

- Pacing your work is important. You won't be able to do everything the week before the final report, and if everyone tries to, you'll break the server.


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

Challenges

- Political
- Internal Awareness
- Legal/Contractual
- Ethical
- Technical

Data Acquisition: Technical (challenges)

- How should you get data?
  - API access
  - Flat files
  - Database dumps
- How much should it be processed before you get it?
- How do you build a repeatable data acquisition pipeline?
- When do you collect new data?

Data Storage

- Use Databases whenever possible
  - Types of databases
- Deidentification when dealing with confidential/sensitive identifiable data
  - hashing

Linkage: Goals

- Determine if pairs of *records* describe the same entity

- Main applications:
  - *Joining* two different data sources
  - *Removing duplicates* from a single data source

Record Linkage: Synonyms

- (data) matching
- merge/purge
- duplicate detection
- de-duping
- reference matching
- co-reference/anaphora resolution

Factors to consider

- Deduping or Linkage
  - 1-1 or 1-many or many-1
- Rule-based or ML based
  - Do you have labeled training data?
- Domain specific or generic similarity metrics?
- Evaluation metric
  - Precision or recall
  - Task-specific - Implications on future analysis (bias for example)

Approaches

- Exact matching
- Rule-based
- Probabilistic linkage

Common reasons for mismatches

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

When are two records about the same entity?

- Examples of possible similarity metrics
  - Edit distance
  - Soundex

“Fuzzy” Matching System

- Apply set of cascading rules
- Assign confidence score based on which rules fire

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.

Machine Learning based Record Linkage

- Generate training data
  - Label pairs as match/no match
- Generate features over each pair
  - Distance metrics over different attributes (fname, lname, dob, etc.)
  - Tfidf scores
- Build and evaluate classifiers

One-off versus recurring matching

- Unique identifiers: persistence?
- What do we do with new or changed pairs?

Discussion Topic

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

Things to remember

- **Make sure and check that you can access your project data today**
- Thursday:
  - Reading (Obermeyer et al; Passi & Barocas)
- What's coming next week:
  - Assignment due Monday
    - Project Formulation and Baselines
  - Readings