outrageously
AMBITIOUS

Module 3: Ethics in AI

Duke
PRATT SCHOOL *of* ENGINEERING

Why consider ethics?

- Ethical issues in AI systems can be particularly dangerous
  - Can have significant impacts on people’s lives
  - Difficult to detect
  - May not violate any laws

Module 3 Objectives:

At the conclusion of this module, you should be able to:

1) Explain the goals of Fair, Accountable and Transparent AI
2) Identify sources of bias in AI projects
3) Implement strategies to mitigate potential ethical risks

outrageously AMBITIOUS

Fair, Accountable & Transparent AI

Duke
PRATT SCHOOL *of* ENGINEERING

Ethical risks of AI

Allocative harm

- Opportunities or resources are withheld from certain people/groups
- Examples:
  - Automated resume review system selects primarily male candidates for interviews for a technical role
  - Men and women with identical backgrounds receive different credit limits in applying for a credit card

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing, Information and Society (2017).

Ethical risks of AI

Representational harm

- Certain people/groups are stigmatized or stereotyped
- Examples:
  - Computer vision model which identifies all female doctors as nurses

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing, Information and Society (2017).

Ethical AI

- Three criteria of ethical AI systems:
  - **Fair**
  - **Accountable**
  - **Transparent**

Fairness

- Roots in anti-discrimination laws
- No single universal definition of fairness

Individual fairness

People who are similar should receive similar outcomes

Group fairness

Different groups should experience similar levels of positive outcome or rates of errors

- Individual and group fairness can come into tension

Accountability

- Clear responsibility for outcomes
- Users have recourse if they identify issues
- Key considerations:
  1. **Who is responsible** for system performance?
  2. On **what set of values and laws** is the system based?
  3. **What recourse do users have** if the system is not behaving in accordance with values and laws?

Transparency

- Users have visibility into data usage and model functioning
- Methods of providing transparency:

Interpretable models

Feature importance

Simplified approximations

Counterfactual explanations

outrageously AMBITIOUS

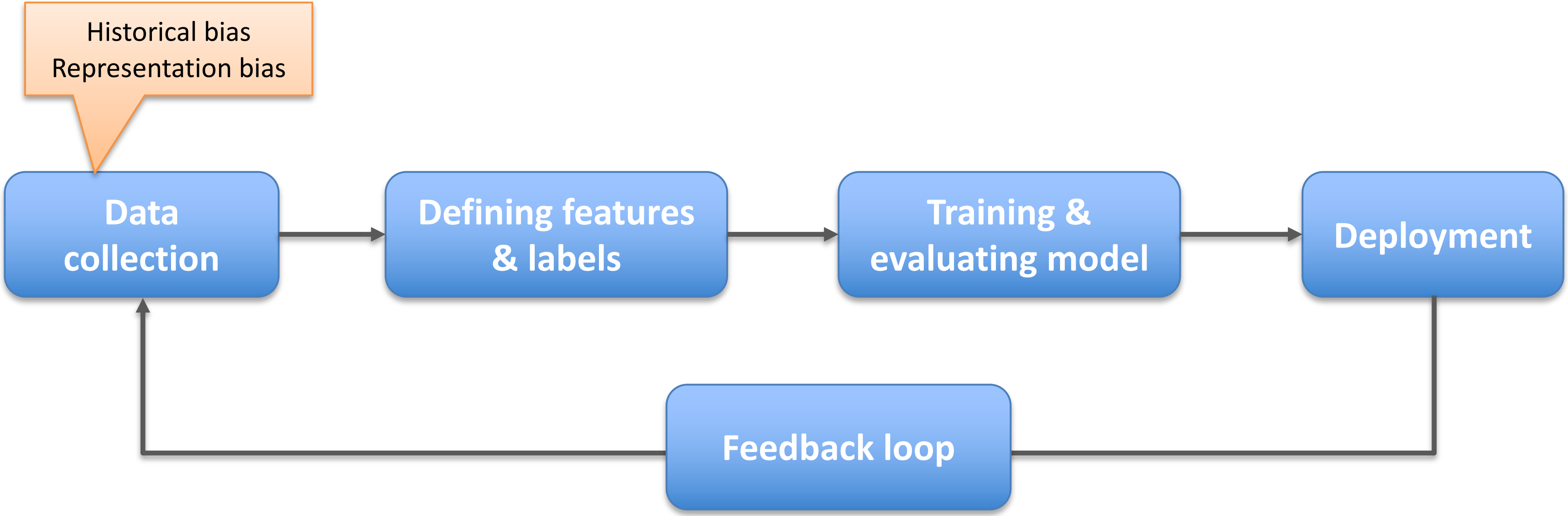
Types & Sources of Bias

Duke PRATT SCHOOL of ENGINEERING

Algorithmic bias

- AI systems often considered to be neutral, but can have many biases
- Systemic errors that create unfair outcomes for individuals or groups
- Can enter into AI systems in many ways:
  - Pre-existing perceptions of system creators
  - Design of data collection or model
  - Unanticipated use of system

Sources of bias



Suresh, Harini, and John V. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." arXiv preprint arXiv:1901.10002 (2019).

Historical bias

- Collected data reflect existing biases in the world around us at the time of data collection

Example

- Word embeddings trained on large-scale text associate occupational words such as “nurse” or “engineer” more strongly with women and men, respectively

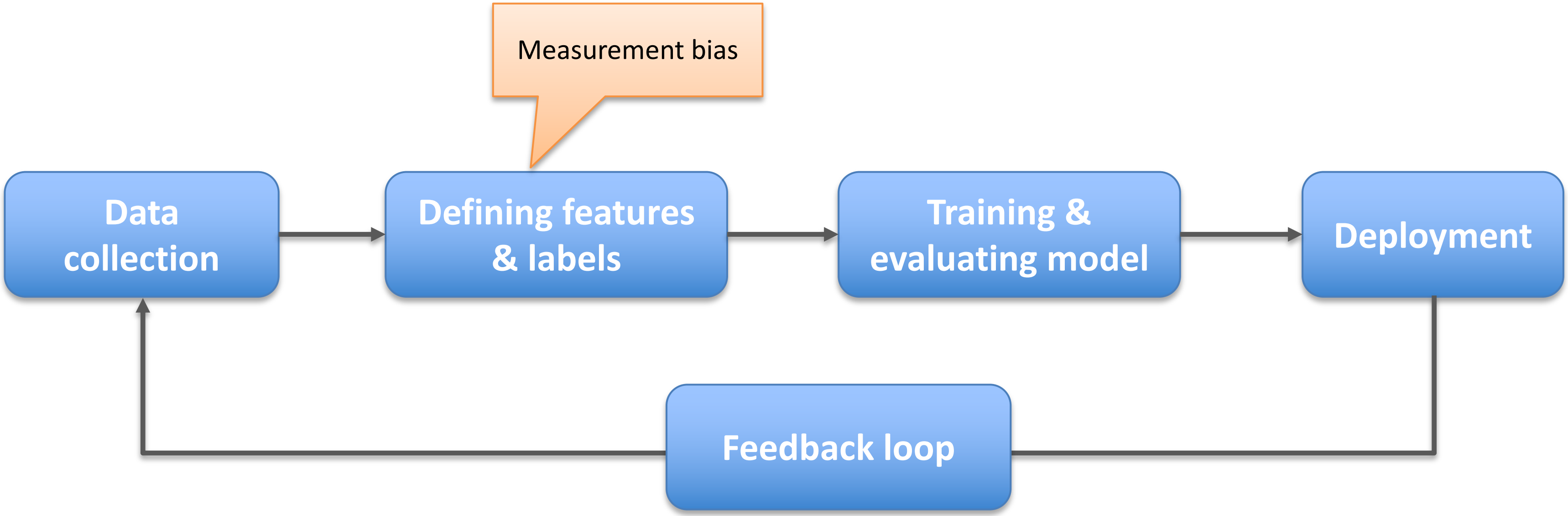
Representation bias

- Training dataset is not representative of the entire target population
  - Certain groups are naturally under-represented in the training data
  - Sampling method results in uneven data collected

Examples

- Certain medical dataset contains only a small % of pregnant women
- City of Boston’s pothole app flagged issues in younger, affluent neighborhoods

Sources of bias



Suresh, Harini, and John V. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." arXiv preprint arXiv:1901.10002 (2019).

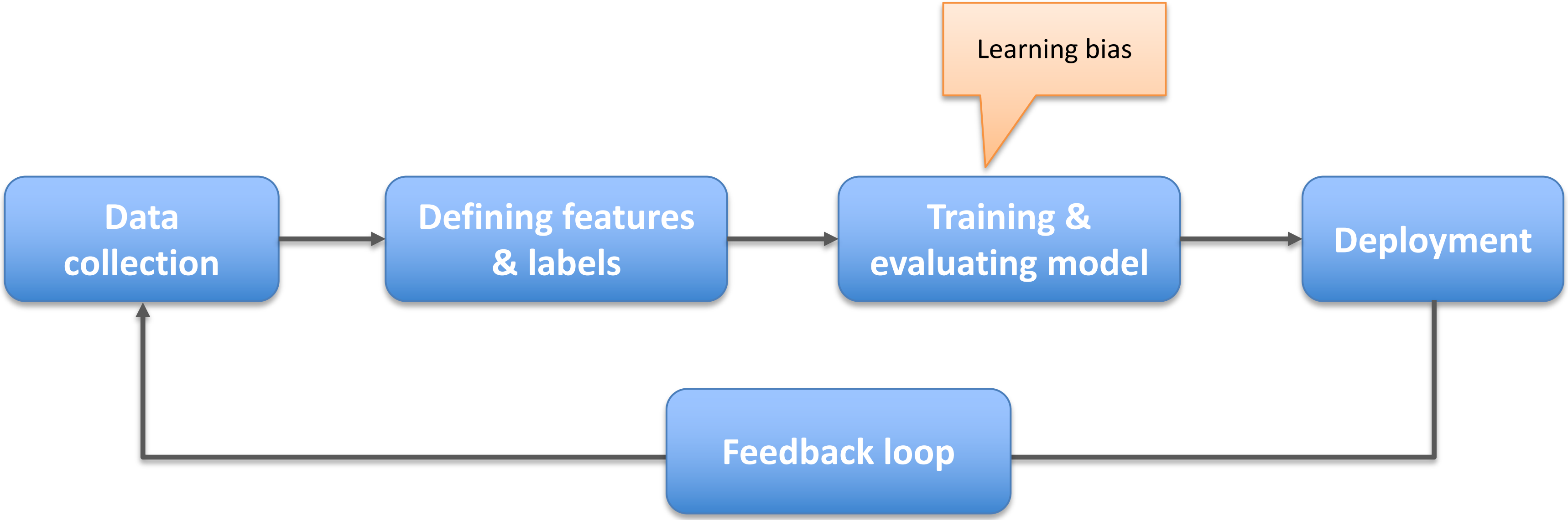
Measurement bias

- Features or labels chosen to represent some construct are poor reflections of it, or vary across groups
  - Proxy is an oversimplification
  - Method of measurement or accuracy varies across groups

Examples

- GPA as a proxy for student learning success
- Count of manufacturing anomalies across sites

Sources of bias



Suresh, Harini, and John V. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." arXiv preprint arXiv:1901.10002 (2019).

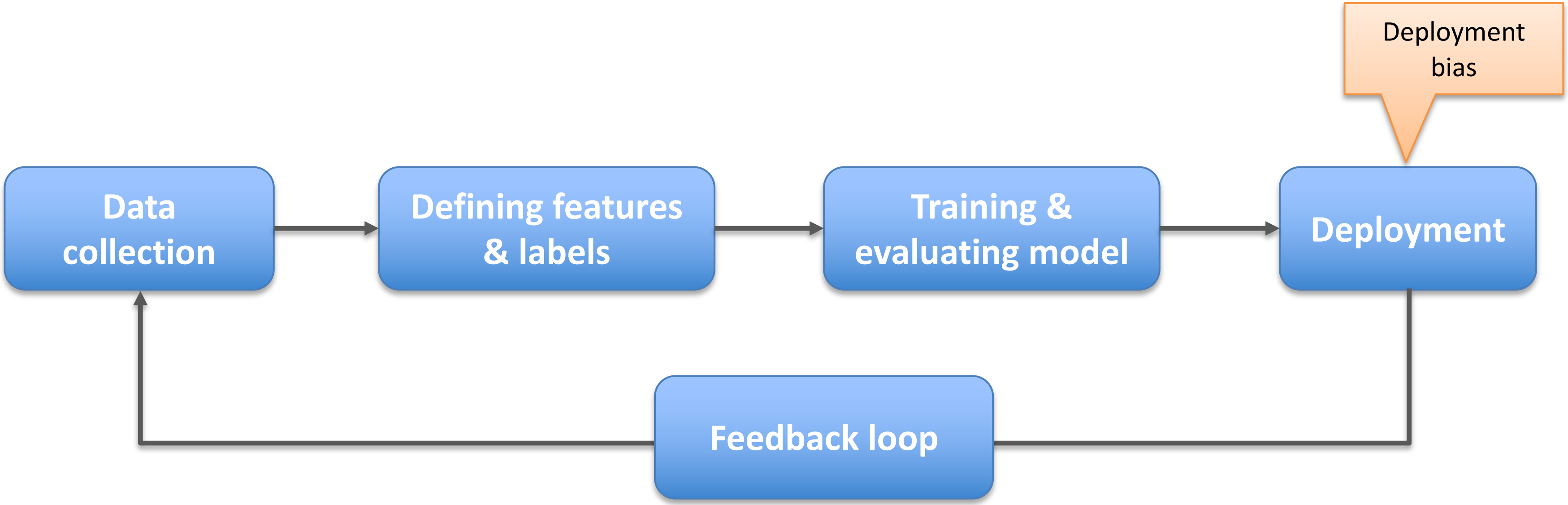
Learning bias

- Modeling choices amplify performance disparities across groups
- Cost function may optimize aggregate performance at the expense of consistency across groups (disparate impact)

Examples

- Use of demographic data to predict likelihood of criminals to re-offend
- Prioritizing smaller models at the expense of underrepresented attributes

Sources of bias



Suresh, Harini, and John V. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." arXiv preprint arXiv:1901.10002 (2019).

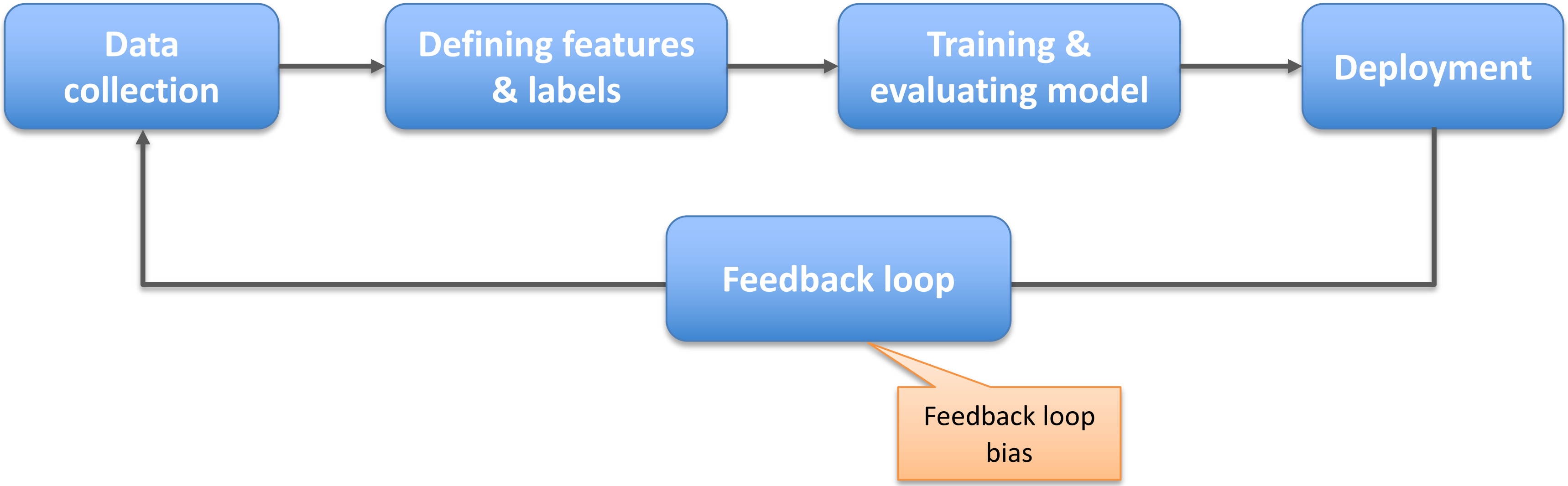
Deployment bias

- Mismatch between how a tool was intended to be used and how it is used in practice
- Occurs when system developers consider tool in isolation of usage environment

Example

- Automated teacher evaluation tool used to terminate low rated teachers

Sources of bias



Suresh, Harini, and John V. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." arXiv preprint arXiv:1901.10002 (2019).

Feedback loop bias

- The design of a system incorporates a feedback loop which influences the training data and thus the model outputs

Example

- Product recommendation engine which bases ordering of items on number of positive reviews

outrageously
AMBITIOUS

Mitigating Potential Ethical Risks

Duke
PRATT SCHOOL of
ENGINEERING

Tools to Mitigate Ethical Risk

- Datasheets for datasets
- Ethical checklist
- Ethical pre-mortems

Datasheets for datasets

- Creation and selection of data is the most common source of bias
- There is currently no standardized process for documenting datasets in the ML community
- In most other industries, all inputs are accompanied by a standard datasheet describing composition and use

See Gebru, et al. Datasheets for Datasets. https://arxiv.org/pdf/1803.09010.pdf

Objectives of a dataset datasheet

- For **dataset creators**
  - Encourage best practices in collecting data
  - Foster reflection on risks and implications of use
- For **dataset consumers**
  - Provide transparency to support decisions on whether/how to use dataset
- For **users of models**
  - Contribute to explainability of model outputs

Ethical Checklist (1/2)

Project Selection & Scoping

- Is the problem we are solving a symptom of a bigger issue?
- Is AI the right tool for the job?

Building the Team

- Does the team include or consider individuals who will ultimately be affected by the tool?
- Does our team reflect diversity of opinions and backgrounds?

Data Collection

- Does collecting data impede on anyone's privacy?
- Have we collected appropriate user consents to use the data?
- Were the systems and processes used to collect the data biased against any groups?
- Have we studied and understood possible sources of bias in our data?

Based on Fritzler, Alan. "An Ethical Checklist for Data Science." https://www.dssgfellowship.org/2015/09/18/an-ethical-checklist-for-data-science/

Ethical Checklist (2/2)

Analysis / Modeling

- Has the team introduced bias in the variable selection or modeling?
- Should the team include features that could be discriminatory?
- Is the analysis sufficiently transparent?
- Have we tested for fairness with respect to different user groups?
- Have we tested for disparate error rates among different user groups?

Implementation

- Are the people using our models aware of its shortcomings?
- Do we have a mechanism for redress if people are harmed by the results?
- Have we listed how this technology can be attacked or abused?
- Do we test and monitor for model drift to ensure our software remains fair over time?

Based on Fritzler, Alan. “An Ethical Checklist for Data Science.” https://www.dssgfellowship.org/2015/09/18/an-ethical-checklist-for-data-science/

Ethical pre-mortems

- Involve diverse group of stakeholders
- Anticipate ethical issues occurring
- What might cause them?
- Why might they turn into major issues?
- How can we prevent them?

Detecting & Resolving Fairness Issues

outrageously AMBITIOUS

Duke
PRATT SCHOOL of ENGINEERING

Defining fairness goals

Set specific goals for system to work fairly
across user groups:

1. **Define groups of significance**
   - Age? Race? Gender? Location? Etc.
   - Combinations?
2. **Determine what "fair" means**
   - Same error rates across groups?
   - Same level of positive outcome across groups?

Defining fairness goals

Example: automated loan approvals

- How should the groups be defined?
- How should we define fairness?
  - Give loans at same rate to different groups, even if they have different rates of historical payback?
  - Give loans proportional to each group's historical payback rate?

Fairness auditing

- Develop a fairness auditing plan
  - Training data collection
  - Test set formation
  - Test set performance
  - Production monitoring
- Fair AI tools simplify the process of evaluating fairness
- Requires access to demographic attributes of interest

Feedback Loops

- Risks may take time to materialize, and environmental factors change with time
- Feedback loop mechanisms
  - Invite user feedback
  - Triage systemic vs. individual issues
  - Regularly review identified ethical risks
- Accountability for executing feedback mechanisms and risk follow up

Resolving Fairness Issues

Three options for resolving fairness issues:

1. Change the data
2. Change the model
3. Change the system

outrageously
AMBITIOUS

Wrap-up

Duke
PRATT SCHOOL of ENGINEERING

Wrap-Up

- Many possible sources of bias in building models
- Ethical risks in AI systems can have significant consequences
- Objective is Fair, Accountable and Transparent AI
- Anticipation of fairness issues is key to mitigation