Module 3: Why Do We Collect a Lot of Data?

Background

This module introduces the concept of why data is collected and how it can be used. Specifically, the module emphasizes how to identify patterns in data and how those patterns might be used as the basis for a model.

Additional Reading for Facilitators

- Elish, M. C., and danah boyd. "Situating Methods in the Magic of Big Data and AI." *Communication Monographs*, vol. 85, no. 1, Jan. 2018, pp. 57–80, https://doi.org/10.1080/03637751.2017.1375130.
- Fischer, Christian, et al. "Mining Big Data in Education: Affordances and Challenges." *Review of Research in Education*, vol. 44, no. 1, 2020, pp. 130–60, https://doi.org/10.3102/0091732X20903304.
- Hasselbalch, Gry. *Data Ethics of Power: A Human Approach in the Big Data and AI Era*. Edward Elgar Publishing, 2021, https://doi.org/10.4337/9781802203110.
- Maffie, Michael David. "The Mythology of 'Big Data' as a Source of Corporate Power." *British Journal of Industrial Relations*, vol. n/a, no. n/a, https://doi.org/10.1111/bjir.12728.
- Milner, Yeshimabeit. "Abolish Big Data." *University of California, Irvine*, vol. 8, 2019.
- Pietsch, Wolfgang. *Big Data – The New Science of Complexity*. 2013, https://philsci-archive.pitt.edu/9944/?aggregate.

Module Motivation

This module launches learners into a conversation around AI and ML and their relationship with data. By the end of this module, learners should feel comfortable with basic ideas about how such models are designed to conduct automated pattern recognition.

Learning Goals

Students will be able to:

1. Understand and discuss the relationship between a dataset and a predictive system generated (or trained) off of it
2. Theorize about potential dataset use in the design and deployment of predictive systems, such as AI and ML products

Activity 1: Identifying Patterns from Data (Slides 3-14)

Purpose:

Step learners through the process of identifying a pattern in data and designing a system that makes use of that pattern.

Alternative Examples:

Depending on learners' educational background and mathematical experience, consider substituting with familiar examples such as:

- Stocking grocery store shelves
- Finding a book at a library using the Dewey Decimal System

Dataset: Colored Shapes

Initial dataset includes: Green square, blue triangle, red circle, green rectangle, orange trapezoid, and blue pentagon

Activity Progression:

Step 1: Number of Edges

- Begin with shapes and their respective number of edges
- **Key insight**: We cannot automatically assume the number of edges is specific to any one shape
- **Challenge**: We cannot define a shape based solely on its number of edges
- **Partial success**: If a shape has four edges, it can only be three of our six options

- **Problem**: This might not be accurate enough for our purposes

Step 2: Adding Edge Length Equality

- Introduce second data dimension: whether all edge lengths for a shape are equal
- **Question**: Can we now define a shape based on number of edges AND equal edge lengths?
- **Answer**: Not quite…

Important Note: Handling Missing Data

Take time to discuss "NA" as it pertains to the circle in this data dimension. Address:

- How is "NA" (something that doesn't exist, cannot be defined, or is unknown) marked in your software?
- Does it appear the same way in all data-processing tools you use?

Step 3: Adding Parallel Edges

- Integrate third data dimension: number of parallel edges
- **Success**: Now we can identify a shape based on:
    - Number of edges AND
    - Whether edges have equal lengths AND
    - How many sets of parallel edges the shape has

Introducing Decision Trees

At this point, introduce the **decision tree** concept:

- The most basic kind of automated template for identifying an unknown entity
- Based on characteristics we know
- **Facilitator note**: Demonstrate that learners already use decision trees in everyday life—they're so simple we often engage with them without realizing it

Step 4: Adding Color (and Demonstrating "Bad" Data)

- Add final data dimension: shape color
- Replace final decision tree step with color instead of parallel edges

- **Question**: Can we identify a shape based only on its color?
- **Answer**: Not quite...

Key Concepts to Address:

Correlation vs. Causation

Revisit this concept (hopefully now familiar from previous modules).

Data Quality

- While data cannot be truly "good" or "bad," its ability to depict events, concepts, or objects can be of better or lesser quality
- This measure is often subjective
- **Example**: Using color instead of parallel edges doesn't provide accurate predictions
- We deem it "bad" (more accurately, of "lesser" quality for predictive purposes)

Model Application and Limitations

Making Predictions

Walk learners through how models can be used to make predictions, demonstrating how new data items can be categorized or identified using the developed model.

Understanding Model Limits

Use another new data item example to demonstrate model limitations—we can only recognize things we've built our model to understand.

Time Management Notes

From DataWorks experience: Pattern generation and application took the bulk of activity time. Informal conversations outside workshop time substituted for deeper discussion of predictive system development and how pattern recognition can be both useful and harmful.

Recommendation: Depending on your setup, consider setting aside additional time to ensure these important conversations happen within the module's allotted time.

Optional Discussion Topics

If time permits, discuss with learners:

- What accuracy looks like within their organization
- Correct predictive potential (avoiding more complex characteristics of confusion matrices for now)

Reflection & Journal Entry

Assignment:

Practice creating your own model based on a logic chain (introduced in Module 1) that takes some input and returns some classification.

Sharing Options:

- **Option 1**: Share back in group discussion (time permitting)
- **Option 2**: Submit for later review by facilitator

Key Skills to Practice:

- Logic chain development
- Input-output relationship design
- Classification system creation