Hybrid Data — Designed + Gathered By James Wagner

Hybrid Data (1)

- Lecture 2: Designed vs Gathered Data
- Often, these are distinct types
  - Designed exemplar: Surveys
  - Gathered examples: Administrative records, social media data, web scraping
- *In some situations, we combine elements of both types*
- *We label these situations "**hybrid**"*

Hybrid Data (2)

- Combining types of data unites strengths of each type
  - ***Designed***
    - ***Strengths****: Designer controls quality, data aligns with concepts of interest (validity)*
    - ***Weaknesses****: Expensive*
  - ***Gathered***
    - ***Strengths****: Large amounts of data, often less expensive to collect*
    - ***Weaknesses****: Not necessarily aligned with concepts, may require labelling or other procedures to create training data*

Hybrid Data (3)

- Two main ways in which “hybrid” data are created

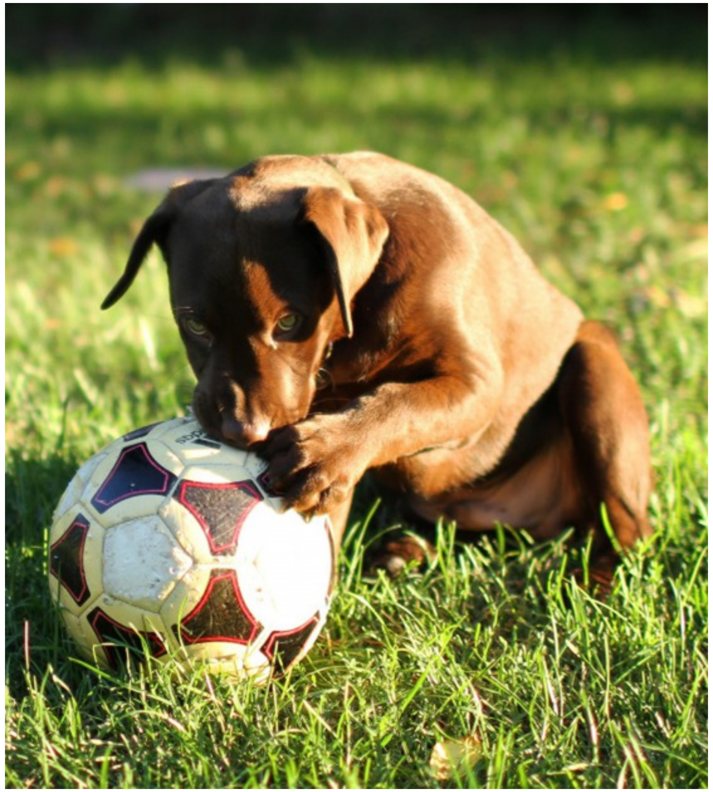
1. Adding designed data to gathered data to create training data
2. Adding gathered data to designed data to enrich/”widen” the data

Hybrid Data (4)

- Example: Adding gathered data to designed data
  - Survey data asks for permission to link administrative records
  - Link tax records to survey data
    - Allows for *(partial)* validation of survey reports on income
    - May provide additional detail
  - However, some may not consent, creating risk of representation bias
  - Tax records do not measure all income, measurement error

Hybrid Data (5)

- Example: Adding designed data to gathered data
- Adding **labels** to a set of images
  - This process creates **training** data that can be used to train an algorithm which will then be used to label new data
  - This process is subject to **measurement error**, just like other designed elements
    - Is this a “soccer” ball or a “football”? Is this a dog or a puppy? Could someone describe the dog as a retriever?
    - “Noisy human labeling” (Misra, et al. 2016)



What’s next?

- Next, we will introduce the Total Data Quality framework
- This framework allows us to identify and discuss potential sources of errors

UNIVERSITY OF MICHIGAN