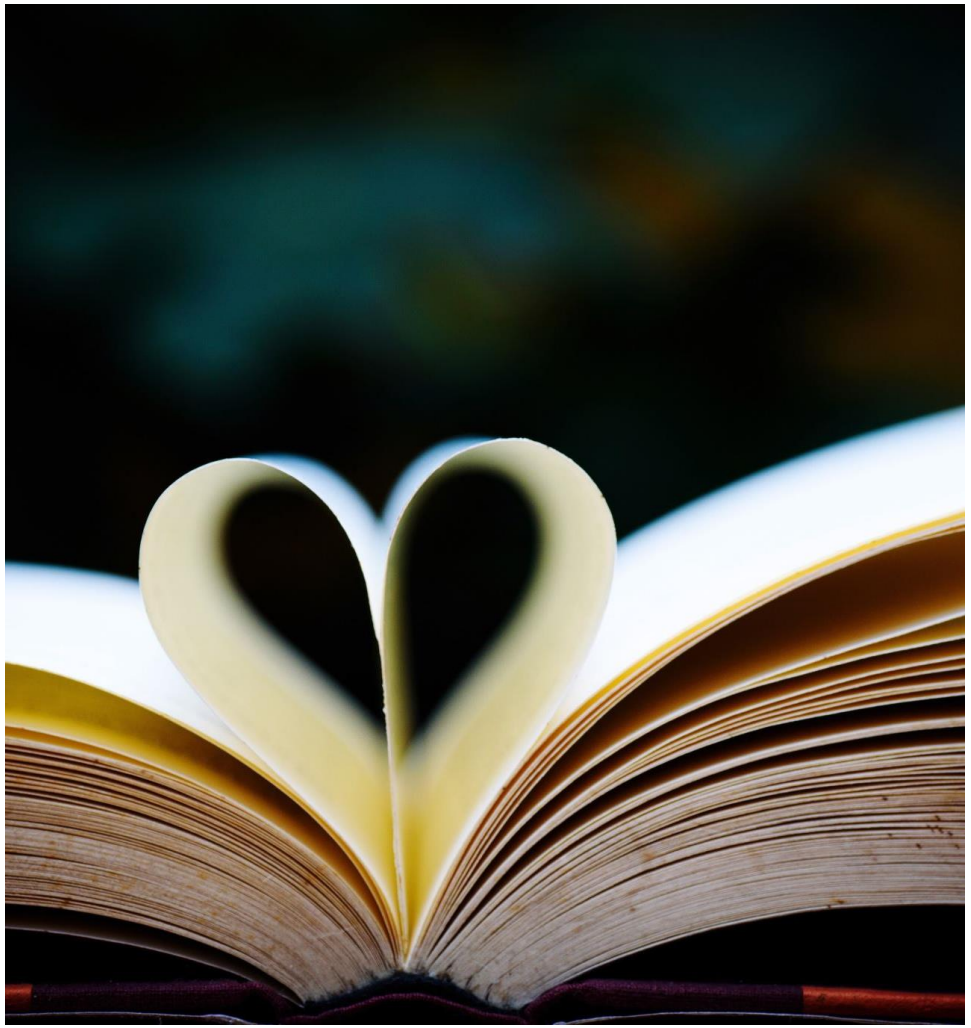
Designed big data

Lecture 7

Literature:

- **Chapter 3**: "Asking questions" in *Bit by bit: Social science in the digital age (2017),* Matthew Salganik **3.6**
- **Chapter 3** "Record linkage" in *Big data and social science (2021),* Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane **3.1 – 3.4**
- Eckman, S., Plank, B and Kreuter, F. (2024) "Position: Insights from Survey Methodology can Improve Training Data"



Lecture goals

- Understand the role of record linkage in designed big data
- Know what steps to take to link multiple datasets and their challenges.
- Understand how enriched and amplified asking work and their differences.
- Know in what ways labelling data knows issues like survey data.

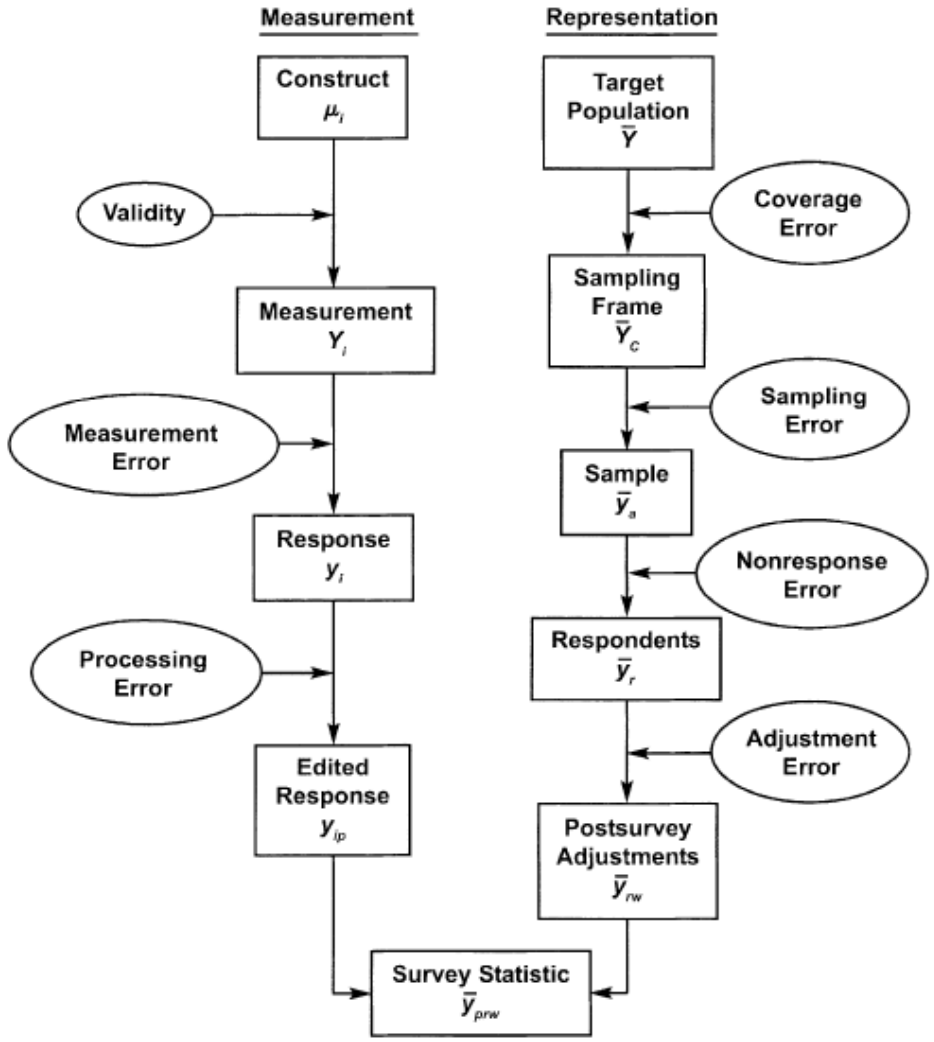
Recap week 4: Errors in digital trace data

Survey methodology is an old and established field of research. It knows about:

- Errors in data collection
- Dealing with expensive data collection
- Collecting data of high quality

Suggestion: **Combine survey data and digital trace data**

- Use the knowledge from survey methodology
- "Control" the unknown issues in digital trace data



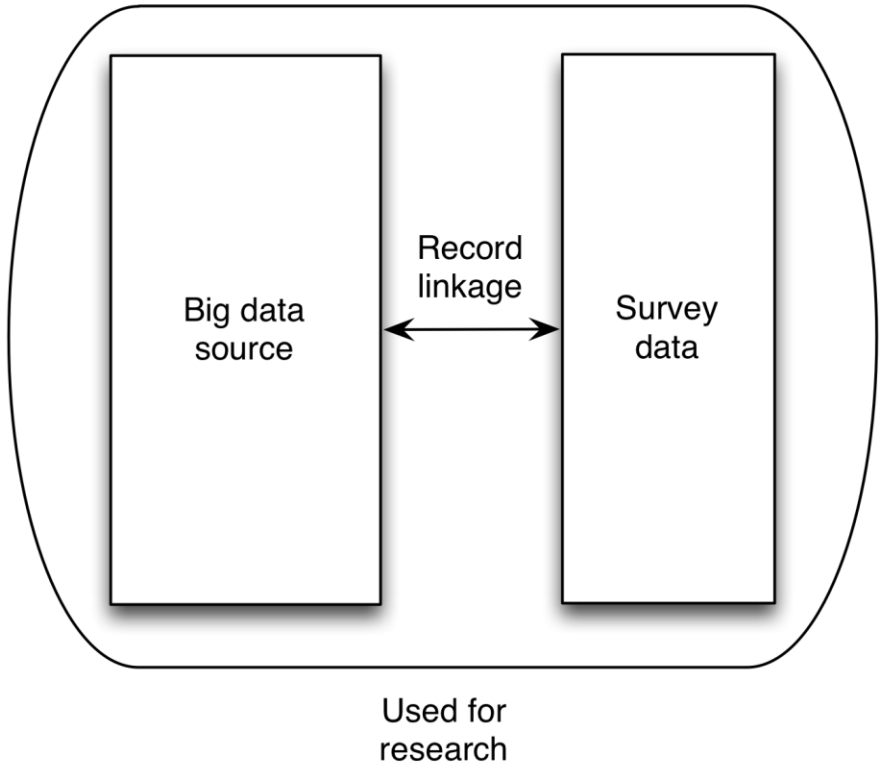
How to combine survey data and digital trace data?

Enriched asking

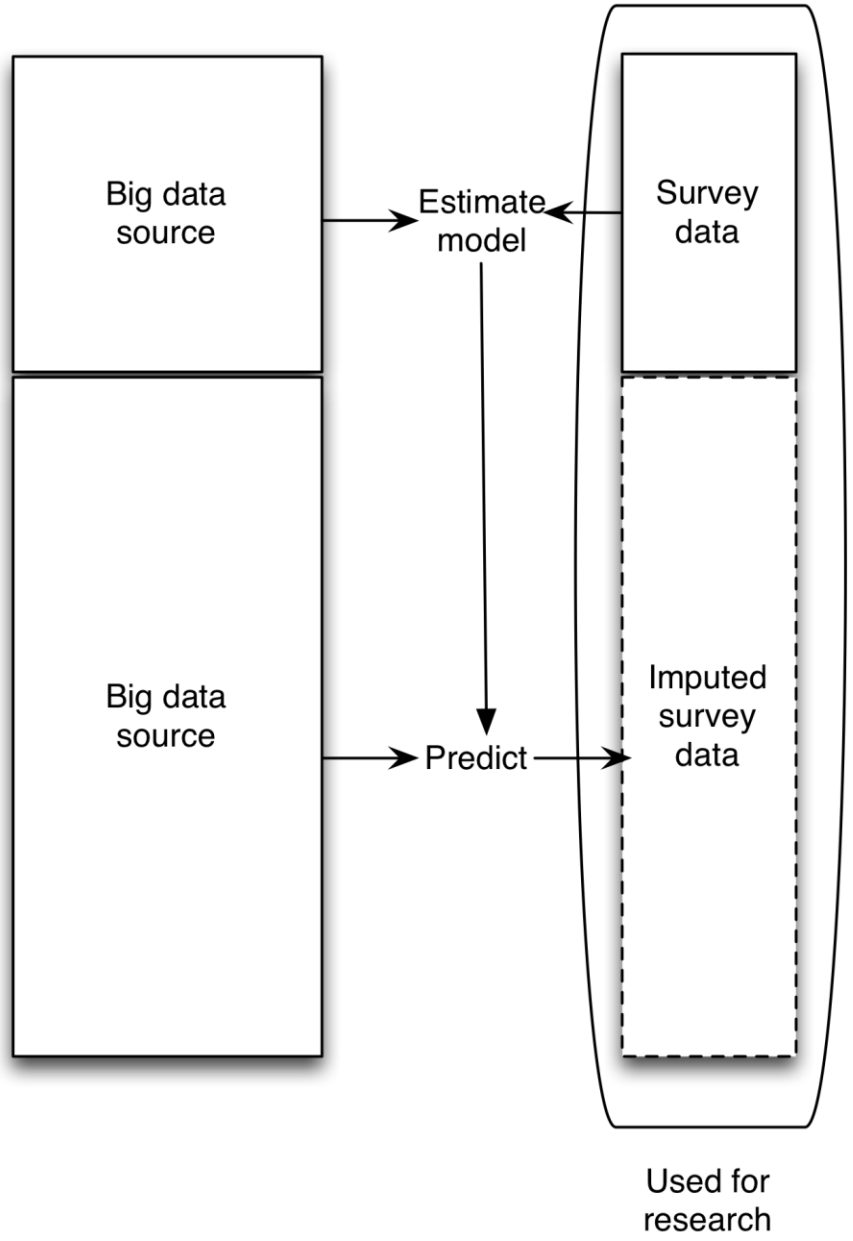
Amplified asking

Surveys and big data are complements and not substitutes!

Enriched asking



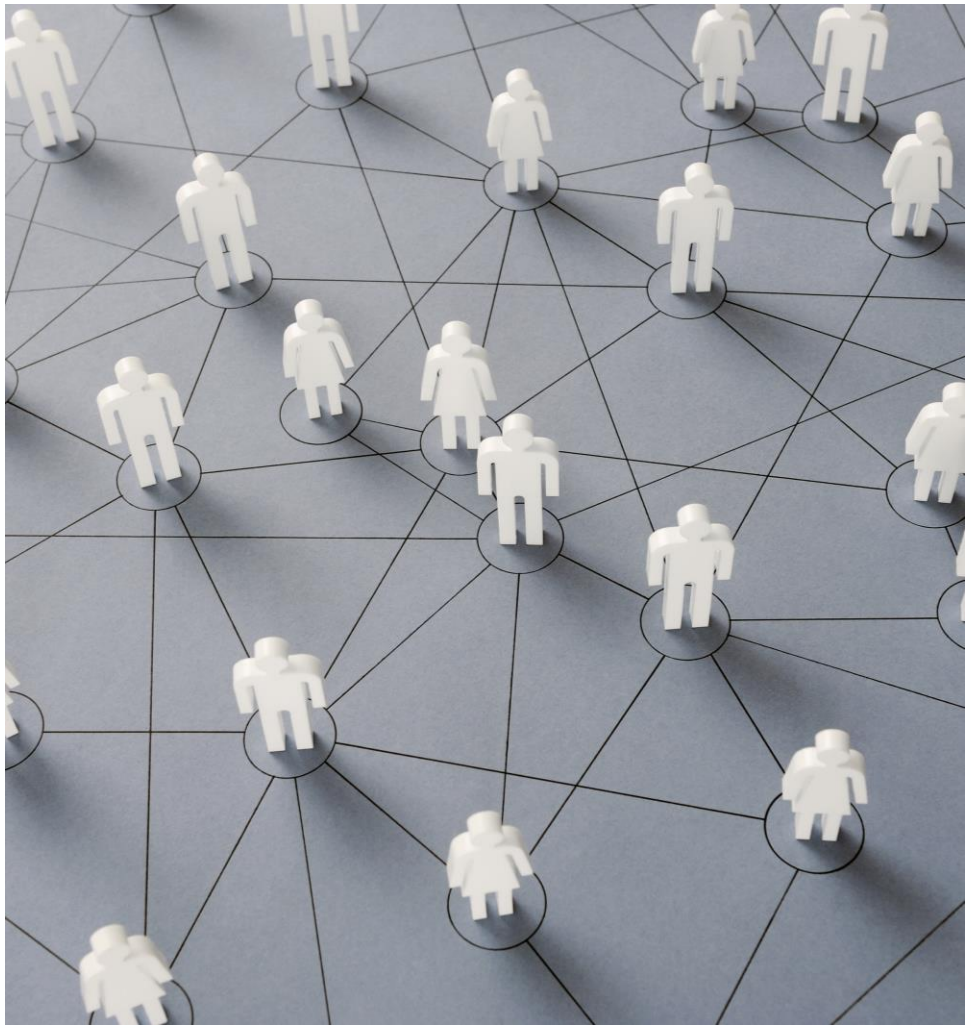
Amplified asking



Enriched asking

Example 1:

Communication via Facebook and feelings of closeness



The practical and fundamental limitations of big data sources, and how they can be overcome with surveys, are illustrated by Moira Burke and Robert Kraut's (2014) research on how the strength of friendships was impacted by interaction on Facebook. At the time, Burke was working at Facebook so she had complete access to one of the most massive and detailed records of human behavior ever created. But, even so, Burke and Kraut had to use surveys in order to answer their research question. Their outcome of interest—the subjective feeling of closeness between the respondent and her friend—is an internal state that only exists inside the respondent's head. Further, in addition to using a survey to collect their outcome of interest, Burke and Kraut also had to use a survey to learn about potentially confounding factors. In particular, they wanted to separate the impact of communicating on Facebook from communication through other channels (e.g., email, phone, and face to face). Even though interactions through email and phone are automatically recorded, these traces were not available to Burke and Kraut so they had to collect them with a survey. Combining their survey data about friendship strength and non-Facebook interaction with the Facebook log data, Burke and Kraut concluded that communication via Facebook did in fact lead to increased feelings of closeness.

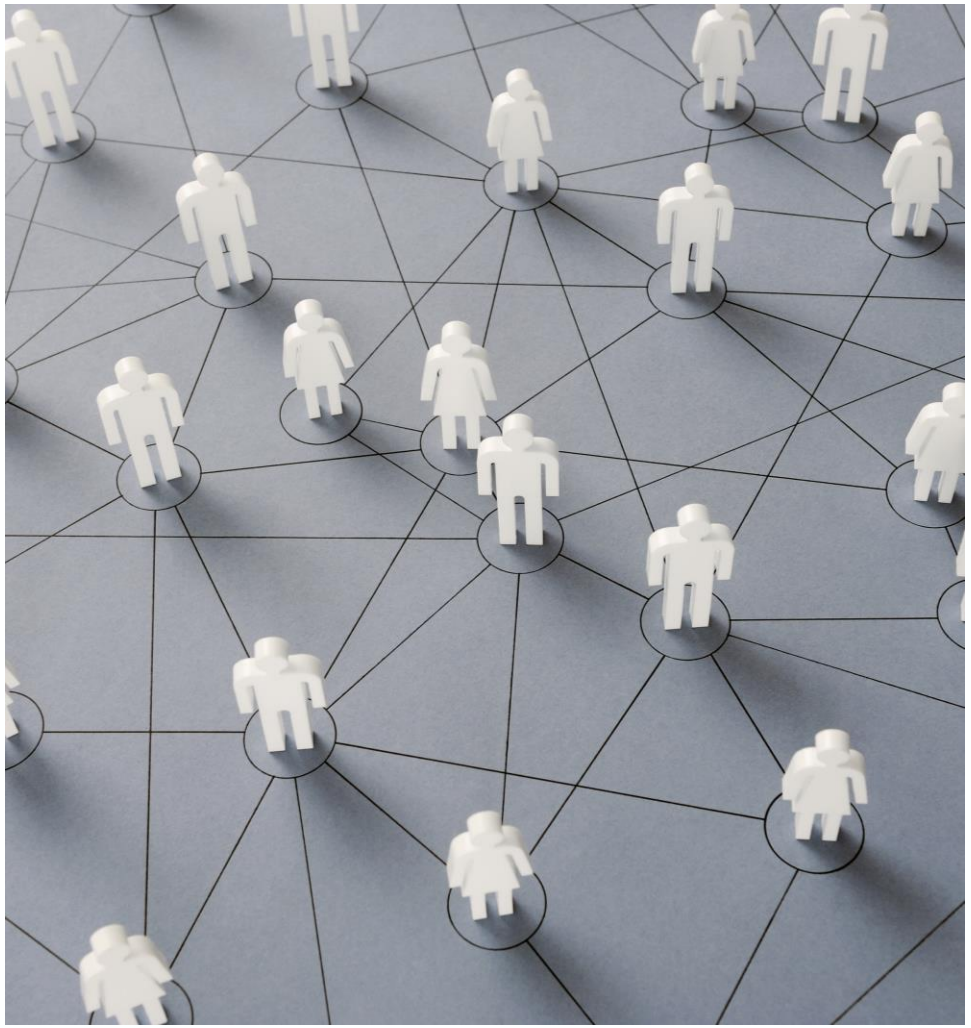
Enriched asking

Example 1:

Communication via Facebook and feelings of closeness

They did not have to deal with the following challenges:

- Data from the survey and Facebook need to be linked (record linkage, unique identifier needed).
- Quality of the big data source is often difficult or impossible to assess.



Project results

- Directed, composed communication is linked with increases in tie strength.
- So does passively reading a partner’s posts.
- Both broadcasting by yourself and by your partner is linked with declines in tie strength when those stories are not read.
- Family ties are less affected by Facebook activity than non-family.

Enriched asking

Example 2:

How does what people say about voting differ from their voting behavior?

Image Fredex8/iStock.com

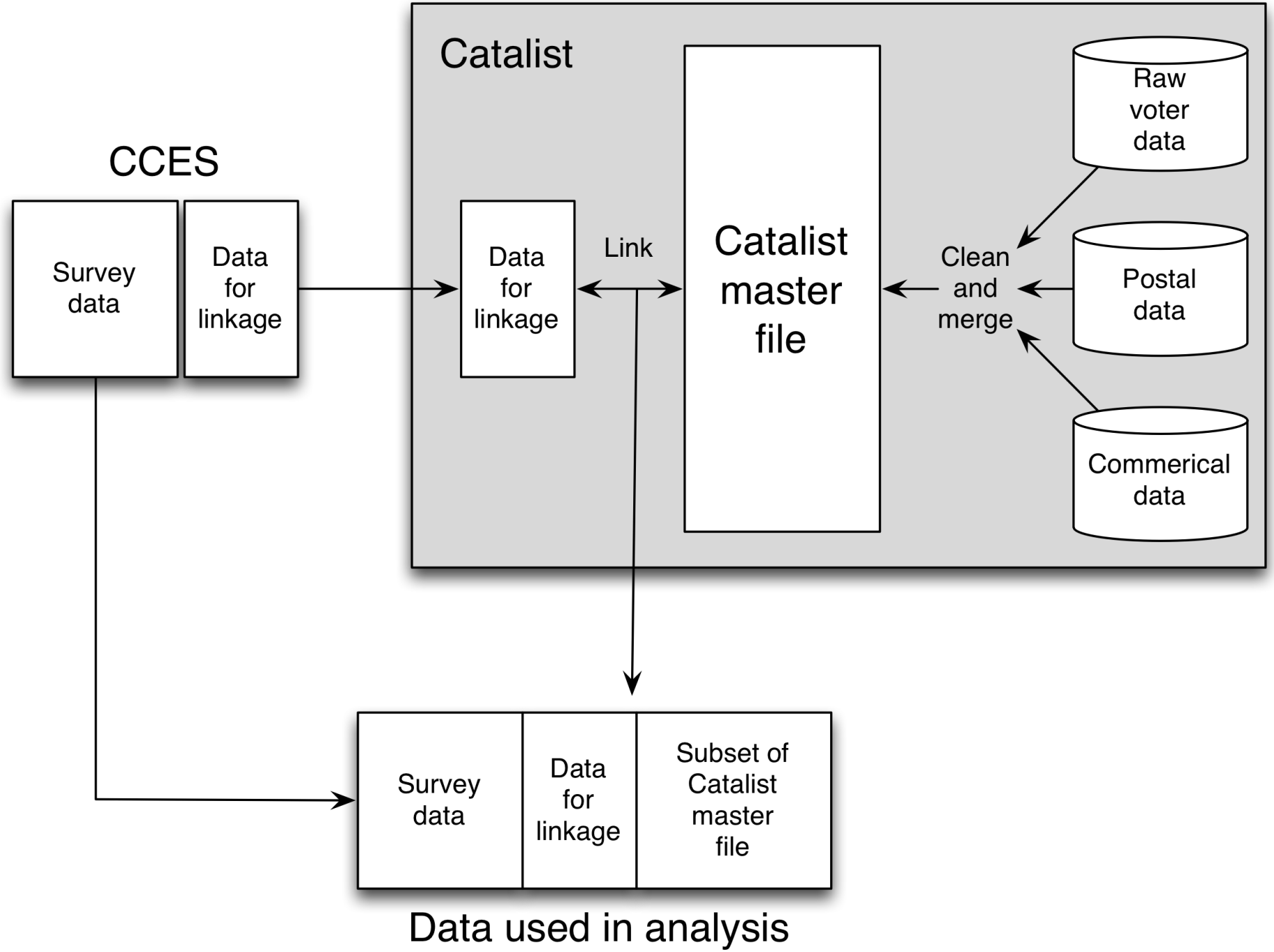
Enriched asking

Example 2:

How does what people say about voting differ from their voting behavior?

- US Government records whether each citizen has voted.
- This can be supplemented with attitudes of respondents from a large social survey

Image Fredex8/iStock.com


Catalist
CCES
Survey data
Data for linkage
Data for linkage
Link
Catalist master file
Clean and merge
Raw voter data
Postal data
Commerical data
Survey data
Data for linkage
Subset of Catalist master file
Data used in analysis

Enriched asking

Example 2:

How does what people say about voting differ from their voting behavior?

- US Government records whether each citizen has voted.
- This can be supplemented with attitudes of respondents from a large social survey
- **Allows to compare survey and admin voting behavior**

Sources of error:

1. The merging done to create the admin Masterfile
2. No unique identifiers, errors in record linkage

Image Fredex8/iStock.com

Project results

- Public opinion surveys overestimate voter turnout. If someone reported voting, there is only an 80% change that they actually did.
- Over-reporting is not random: A particular group consistently misreport: well-educated, high-income, partisan, politically active, church-attending.
- As a result, the differences between voters and non-voters are smaller than appears from literature.

How can enriched asking help?

- There is big value in enriching big data sources and in enriching surveys.
- We can do things that are not possible to do with just one of them.
- Researchers can benefit from the efforts done by private companies.

- Administrative or commercial datasets cannot be considered a “ground truth”
- But surveys also not!

Crucial: record linkage

- Enriched asking assumes perfect linkage between data sources.
- If linkage is not perfect, this can be handled statistically.
- But …. This comes with assumptions.
- Record linkage is essential to make use of enriched asking.

More about record linkage

Why is it important?

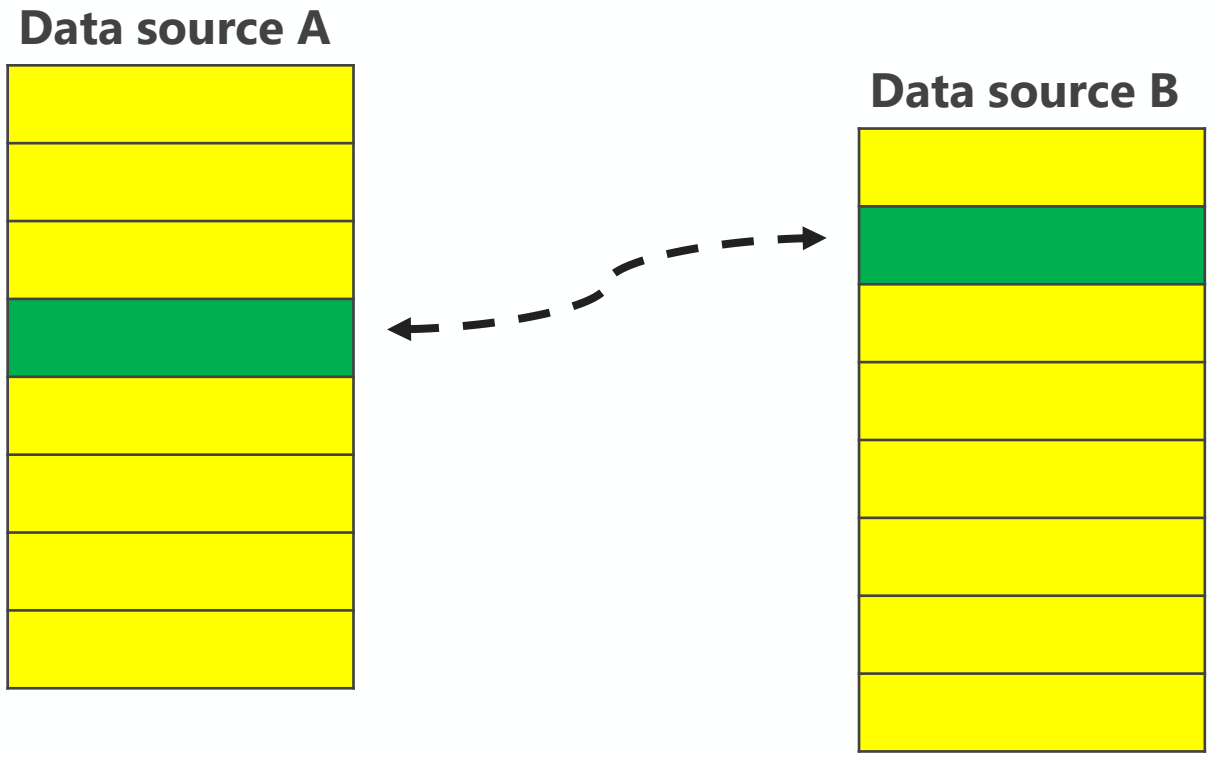
- Linking separate data sources allows to create a **combined** dataset that is richer in **coverage** and in **measurement** than any of the **individual** data sources

When is record linkage easy?

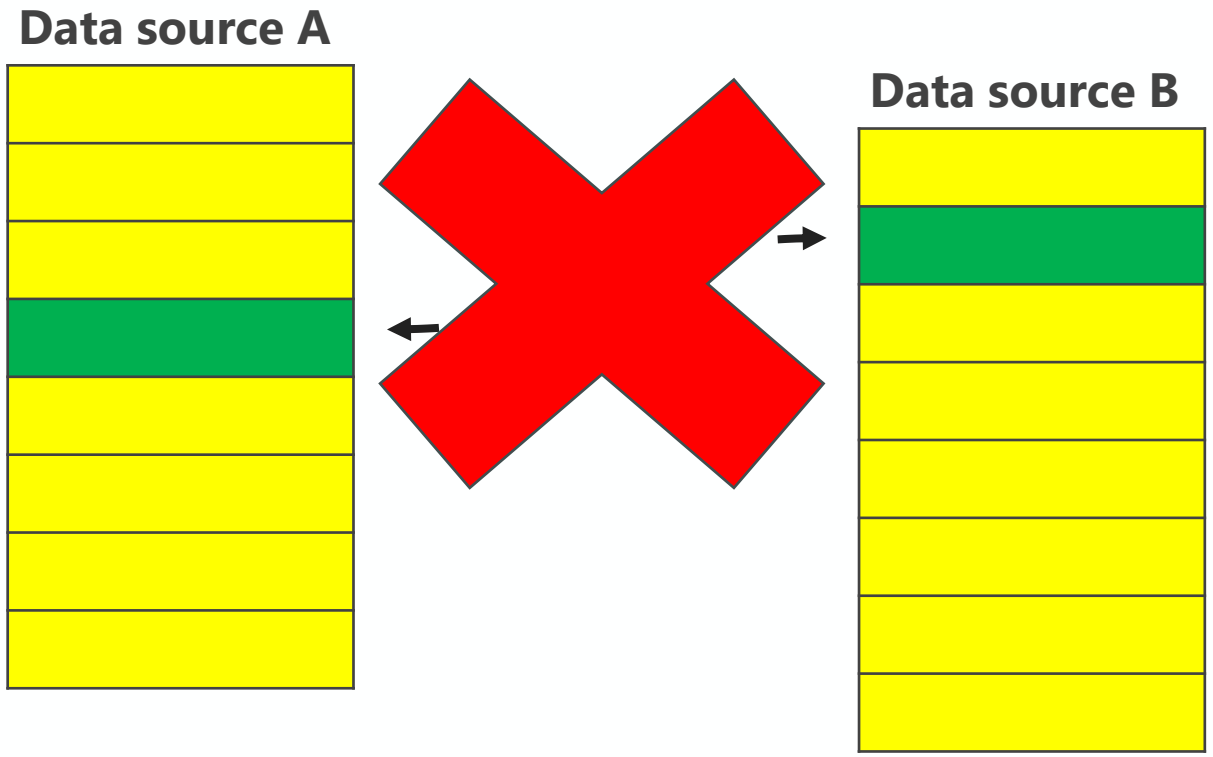
- Each entity has a corresponding **unique identifier** that appears in the data sets to be linked. No special techniques are needed to join the datasets.

If that is not the case, what can go wrong:

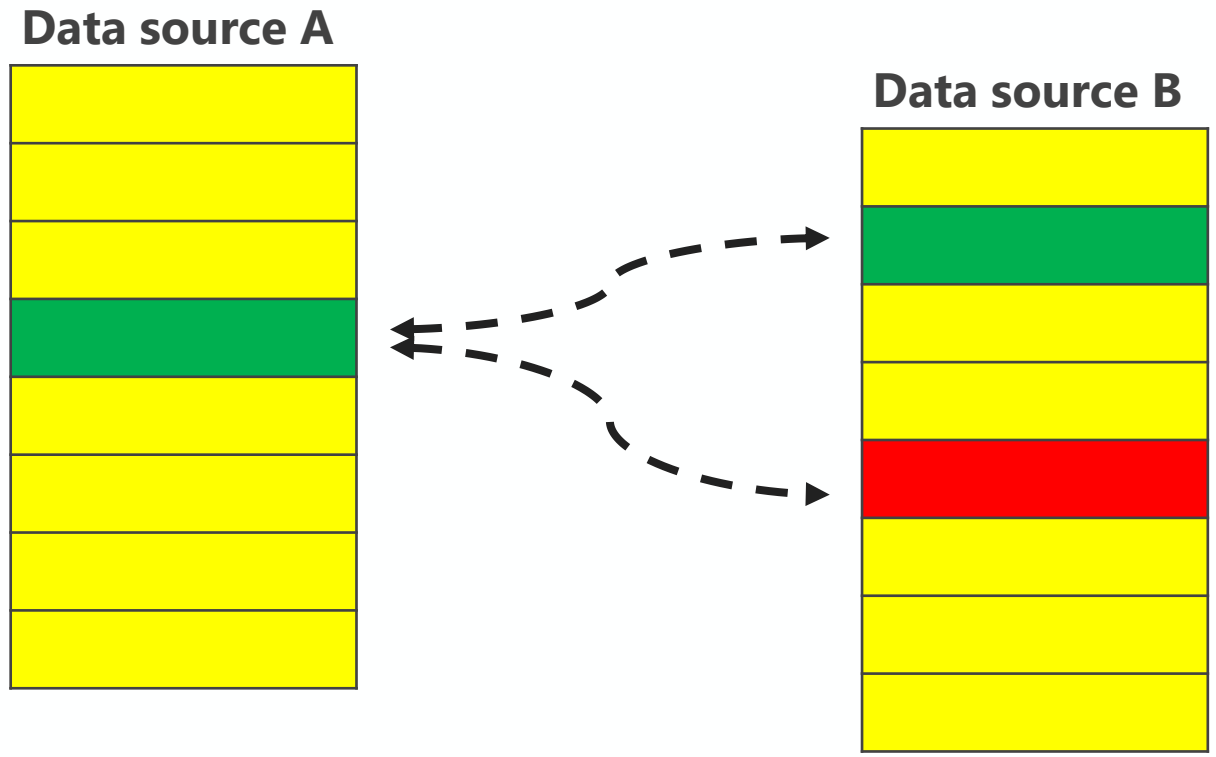
- **Missing links** → no link being made
- **Duplicative links** → linked to multiple entities
- **Erroneous links** → linked to wrong entity


Data source A
Data source B

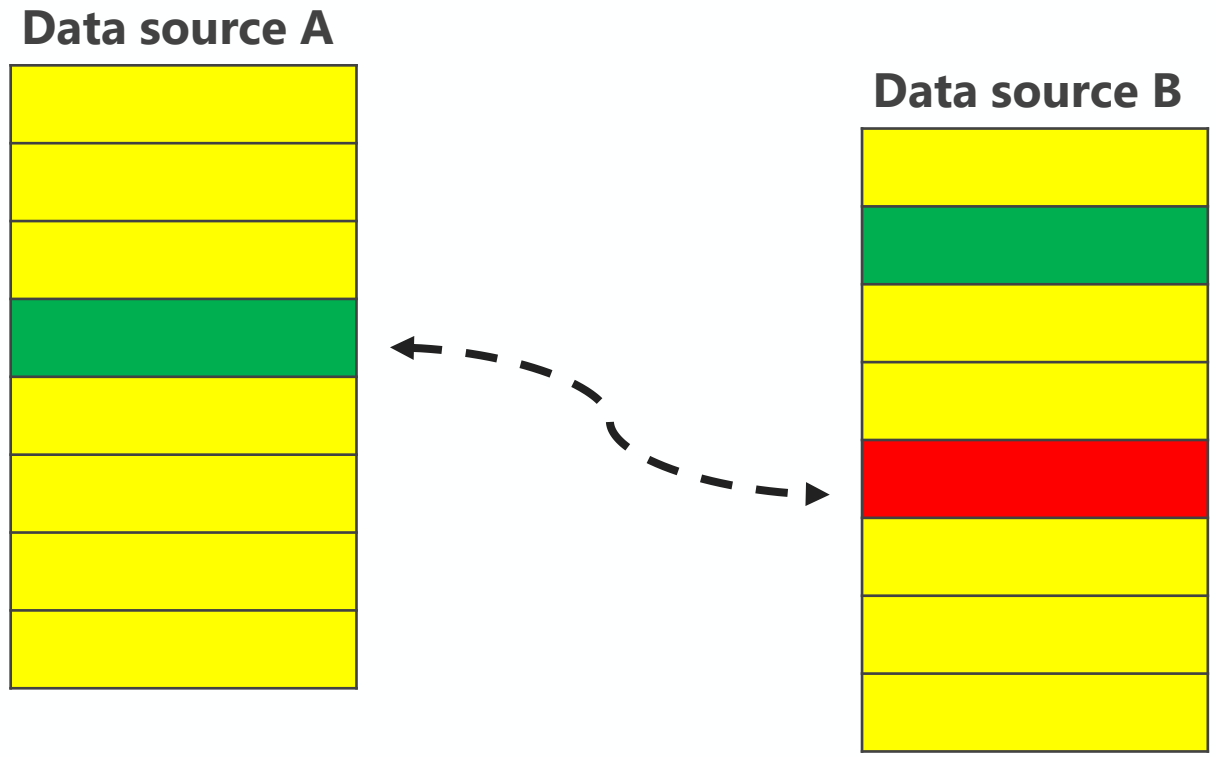
Missing links



Duplicate links



Erroneous links



What if there is no unique identifier?

We need to rely on fields that are only partially identifying the entities, such as names, addresses or dates of birth.

It becomes extra challenging if:

- Poor data quality
- Duplicate records

Steps in record linkage

Goal: Find all possible links between two data sets.

1. **Blocking:** What comparisons to make?
2. **Matching**: Compare each record of dataset 1 with each record of dataset 2.
3. **Outcome**: Set of links: Record pairs that correspond to the same entity.
4. **Evaluate:** Estimate the resulting error rates

Consider:

- Why are you linking these datasets?
- What are the costs of a false negative versus the costs of a false positive in linkage?

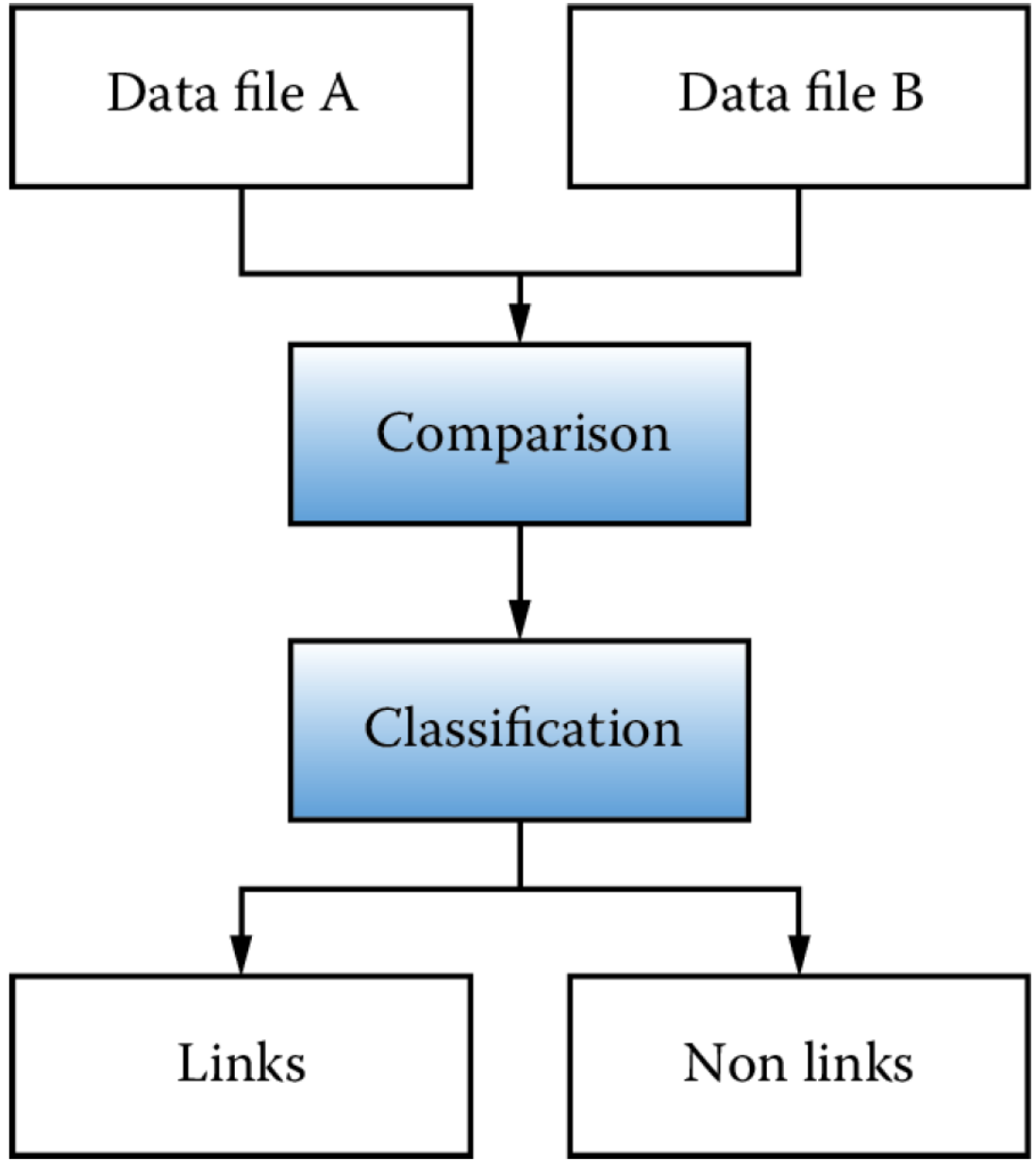


FIGURE 3.1 The preprocessing pipeline

Step 0: Preprocessing!

- Process the identifier variables before you start the matching
- Create a workflow to transform the messy and noisy data into well-defined, clearly structured, and quality tested data

This requires judgement and cannot be effectively automated

- Develop the link keys
- Complex link keys (like addresses) can be broken down into components so that the components can be compared independently. In this way, you can isolate the errors

Step 0: Preprocessing!

Example 1 Address:
- 1600 Pennsylvania
- 160 Pennsylvania Ave

Example 2: Different variants of a field
- Legal first name versus nickname in one dataset
- “First name” in the other dataset

Two purposes:
1. To correct for issues in ***data quality***
2. To account for the different ways that ***input files*** are ***generated***, resulting in the same underlying data being recorded on different scale or according to different conventions.

Step 0: Preprocessing!

4. Evaluate: Estimate the resulting error rates → We can do this for every step!

How can you create errors in the preprocessing step?

Step 1: Blocking

What comparisons to make?

- If we have *2* datasets of *n* records to link, we need to make *n^2* comparisons.
- The proportion of record comparisons that correspond to a link is only *1/n*

Conclusion:

- The linking algorithm spends the bulk of its time comparing records that are not matches.
- We can **speed up** by skipping comparisons between records of pairs that are not likely to be linked.

Step 1: Blocking

With **blocking**, you determine first what parts of the datasets will be compared:

1. Construct a **blocking key** for each record by concatenating fields or parts of fields.
2. Two records with identical blocking keys are said to be in the same block.
3. Only records in the same block are compared.

Step 1: Blocking

Example
two list of individuals:

1. Blocking key is the first letter of the last name + postal code.
2. Blocking on: the first character of the last name + postal code.
3. Total number of comparisons now consists of only comparing those individuals who live in the same postal code and last names begin with the same letter.

Step 1: Blocking

Example
two list of individuals:

1. Blocking key is the first letter of the last name + postal code.
2. Blocking on: the first character of the last name + postal code.
3. Total number of comparisons now consists of only comparing those individuals who live in the same postal code and last names begin with the same letter.

Last name	Postal code
Boeschoten	3584 CS
Bernardo	3584 CS
Carriere	3584 CS
….	….

Last name	Postal code
Boeschoten	3584 CS
Bernardo	3584 CS
Carriere	3584 CS
….	….

Step 1: Blocking

Example
two list of individuals:

1. Blocking key is the first letter of the last name + postal code.
2. Blocking on: the first character of the last name + postal code.
3. Total number of comparisons now consists of only comparing those individuals who live in the same postal code and last names begin with the same letter.

Last name	Postal code
Boeschoten	**3584 CS**
Bernardo	**3584 CS**
Carriere	3584 CS
….	….

Last name	Postal code
Boeschoten	**3584 CS**
Bernardo	**3584 CS**
Carriere	3584 CS
….	….

Step 1: Blocking

Example
two list of individuals:

1. Blocking key is the first letter of the last name + postal code.
2. Blocking on: the first character of the last name + postal code.
3. Total number of comparisons now consists of only comparing those individuals who live in the same postal code and last names begin with the same letter.

Last name	Postal code
Boeschoten	**3584 CS**
Bernardo	**3584 CS**
Carriere	3584 CS
….	….

Last name	Postal code
Boeschoten	**3584 CS**
Bernardo	**3584 CS**
Carriere	3584 CS
….	….

Step 1: Blocking

4. Evaluate: Estimate the resulting error rates → We can do this for every step!

How can you create errors in the blocking step?

1. Blocking creates a potential bias in the linked data because true matches that do not share the same blocking key will not be found.
2. Because blocking keys are compared exactly, there is an implicit assumption that the included fields will not have typos or other data entry errors.

Step 1: Blocking

4. Evaluate: What are resulting errors? → We can do this for every step!

How can you create errors in the blocking step?

Last name	Postal code
Boeschoten	3584 CS
Bernardo	3584 CS
Carriere	3584 CS
….	….

Last name	Postal code
Doeschoten	3584 CS
Bernardo	3584 CS
Carriere	3584 CS
….	….

Step 2: Matching

Compare each record of dataset 1 with each record of dataset 2.

Pairwise comparison:
Compare two records and output a score that quantifies the similarities between those records.

- Use all fields of the record.
- Code them binary (same or not)
- Or indicate a certain level of agreement
    - **Edit distance:** minimum number of edit operations needed to convert one string to another.
    - **Jaro-Winkler:** words with more characters in common will have a higher value (between 0 and 1) than those with fewer characters in common.

Step 2: Matching

Once computed, the field comparisons must be combined to produce a final prediction of match status.

Rule-based approach:

- A set of ad hoc rules to determine which pairs of records should be linked.
- You can also use information from fields that are only in one of the dataset.

  Example: education and occupation.

Step 3: Outcome

Set of links: Record pairs that correspond to the same entity.

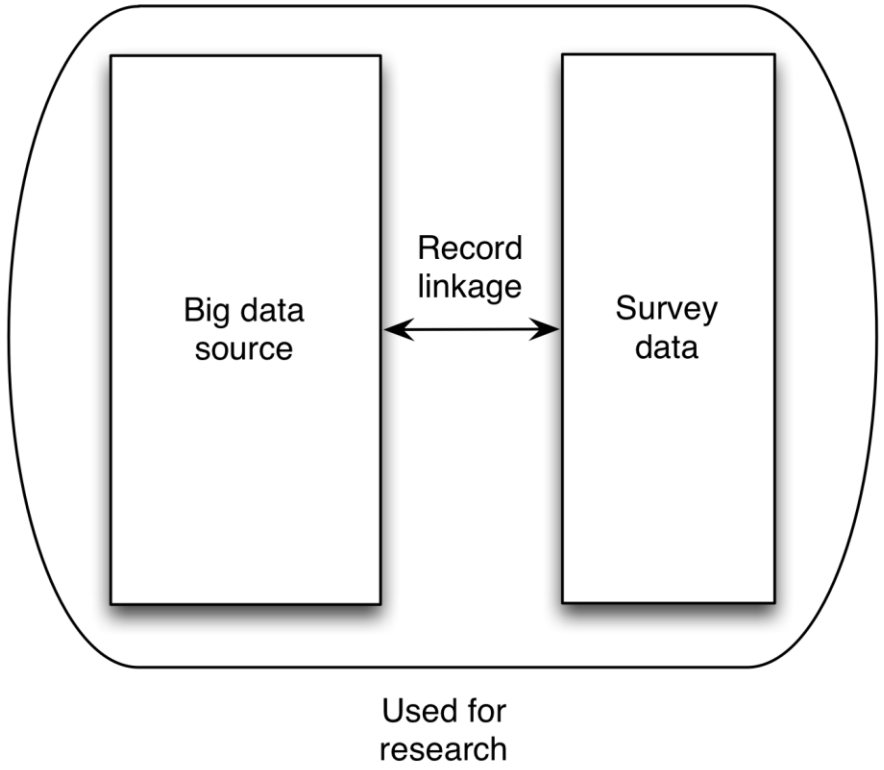
Step 2: Matching

4. Evaluate: What are resulting errors? → We can do this for every step!

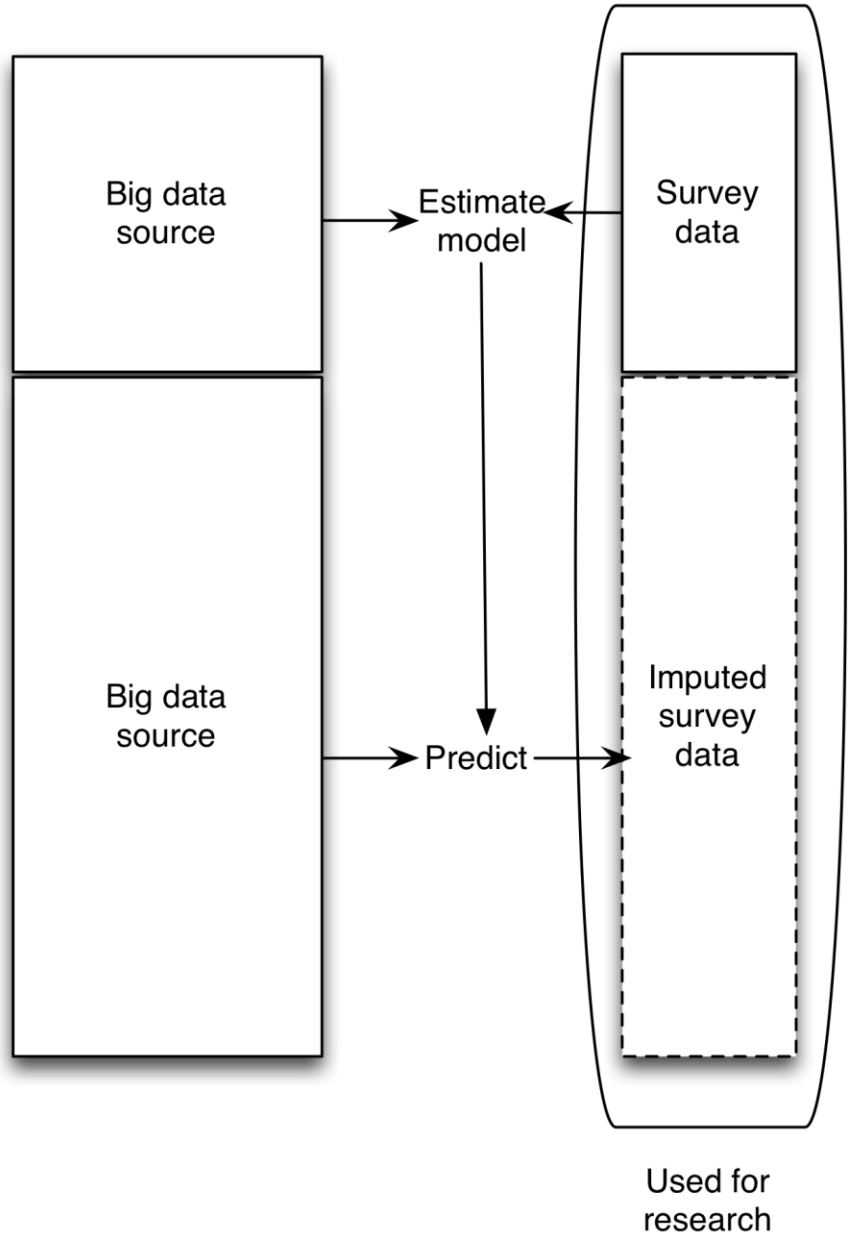
- A different approach to compute field comparisons …
- A different set of rules …
- …


Coffee break

Enriched asking



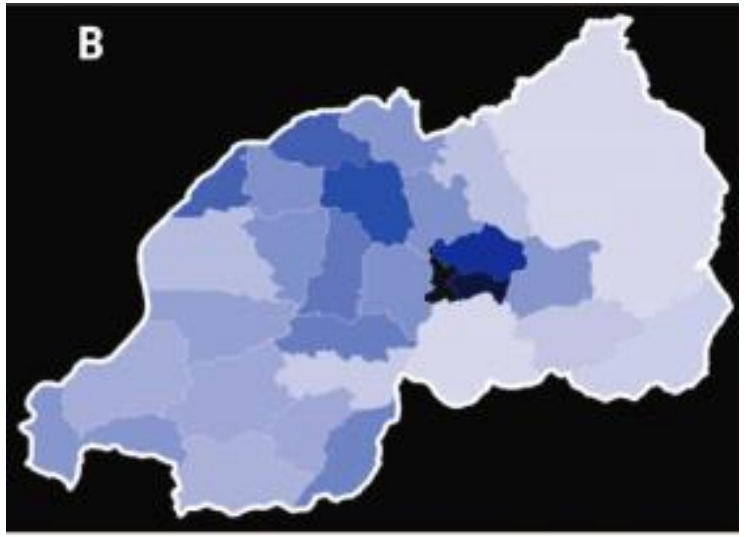
Amplified asking



Amplified asking

Use a predictive model to **combine** a small amount of **survey** data with a **big data** source to produce estimates at scale or granularity that would not be possible with either one of the two sources individually.

Example: Study by Blumenstock
Goal: help guide development in poor countries.
Concepts: measure wealth and wellbeing



(Blumenstock et al., 2015)

Traditional approaches

Survey:
- Hard to make estimates about specific geographical regions or demographic groups.

Census:
- Expensive
- Can only ask a few questions
- Does not happen often

Combine and get the best of both!

Collaboration with a phone company

Phone company:
- Has all phone record data

How to supplement with survey data:
- Select a random sample of phone numbers
- Call them
- Ask for consent (+ consent to link)
- Ask survey questions to measure wealth and well-being.

Two step procedure on phone data

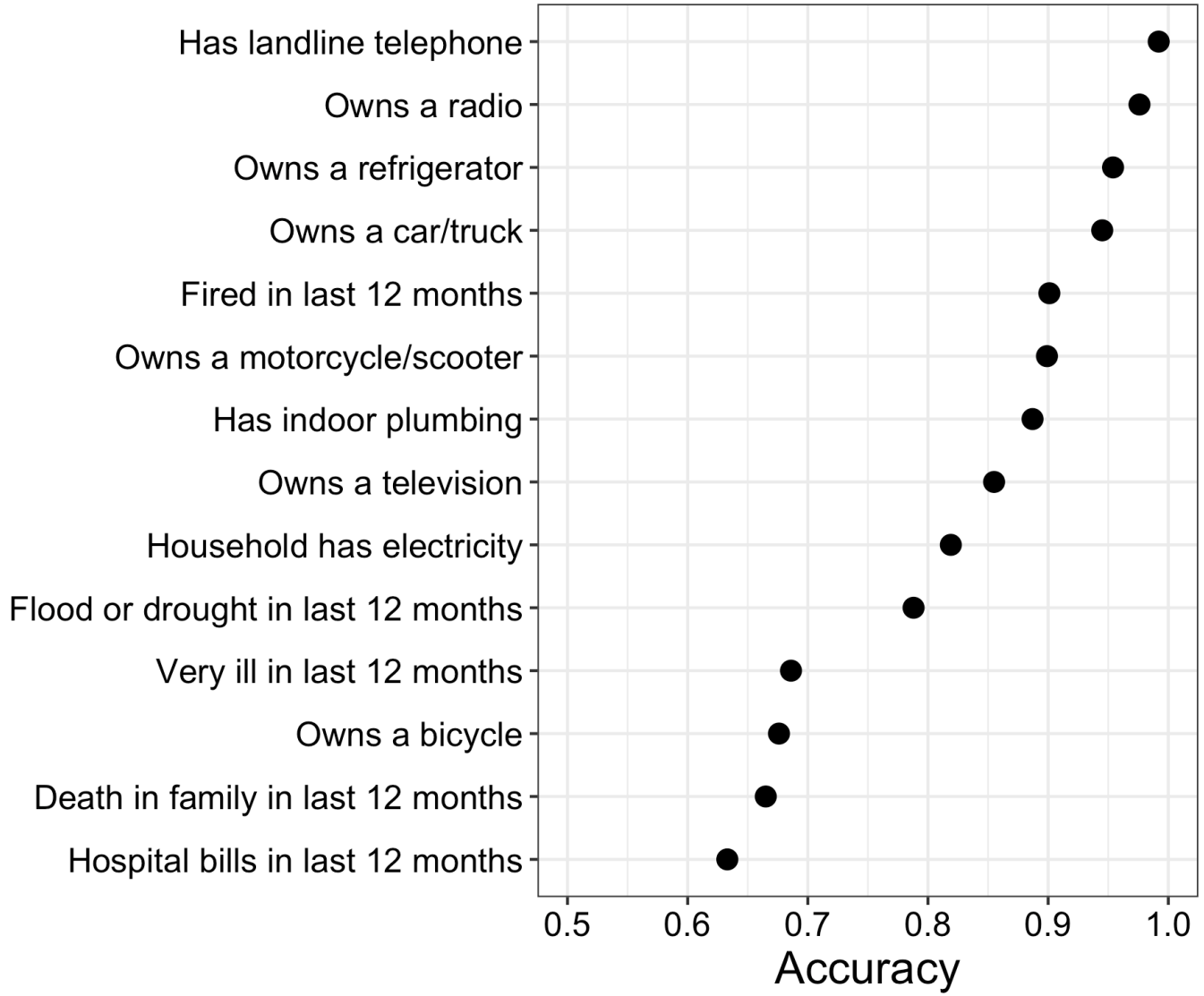
Step 1: Feature engineering step:

- Call records converted into a set of characteristics about each person.
- Features / variables
- Examples:
    - Number of days with activity
    - Number of distinct people a person has been in contact with
    - Amount of money spent on airtime

Two step procedure on phone data

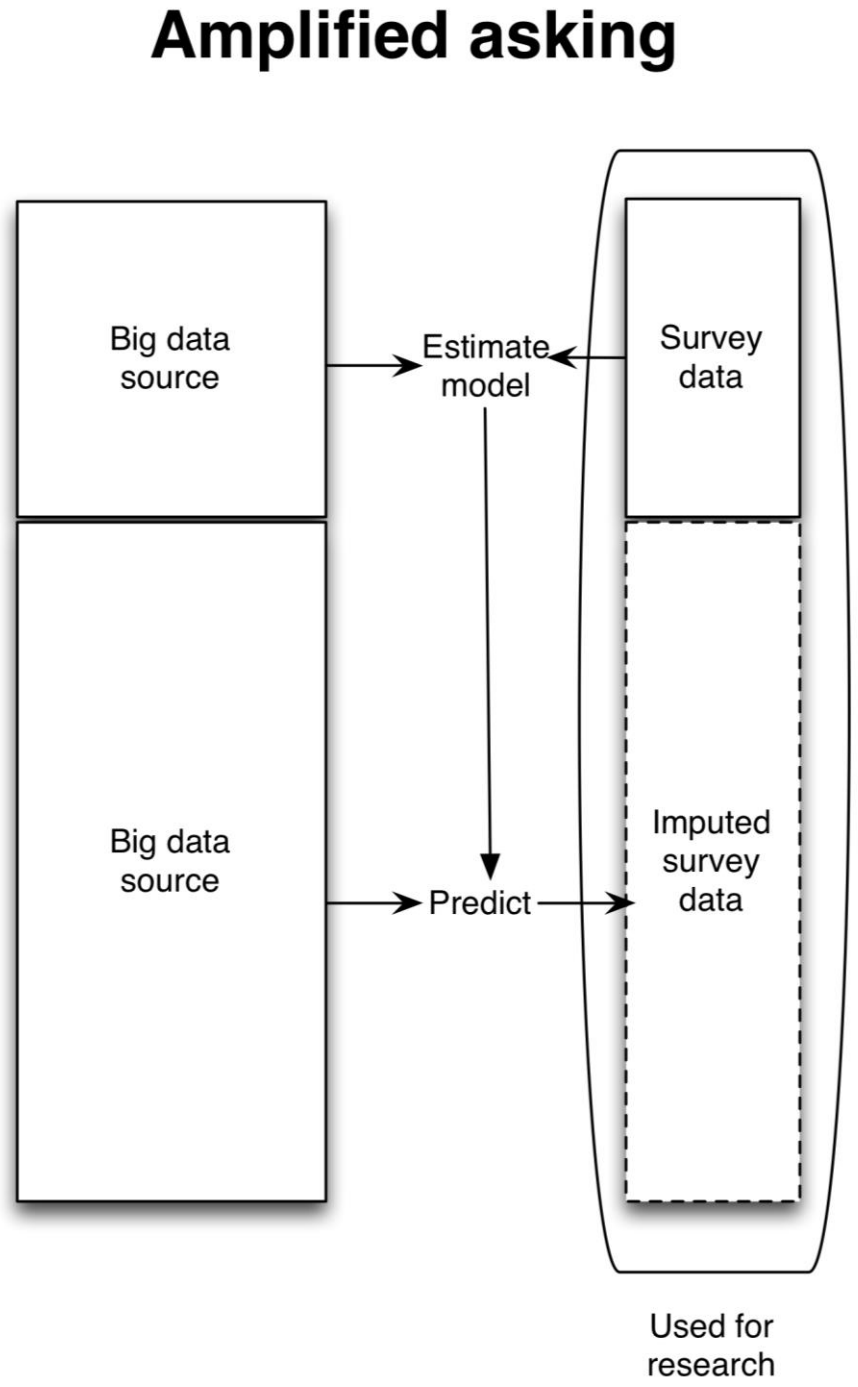
Step 2: Supervised learning step:

- Model to predict the survey response for each person based on their features.
- Used cross-validation to evaluate the performance of the model.
    - How well did the model perform beyond just making a baseline prediction?
    - Make 10 groups of persons in the data, train model on 9 and evaluate performance on 10, do this 10 times.



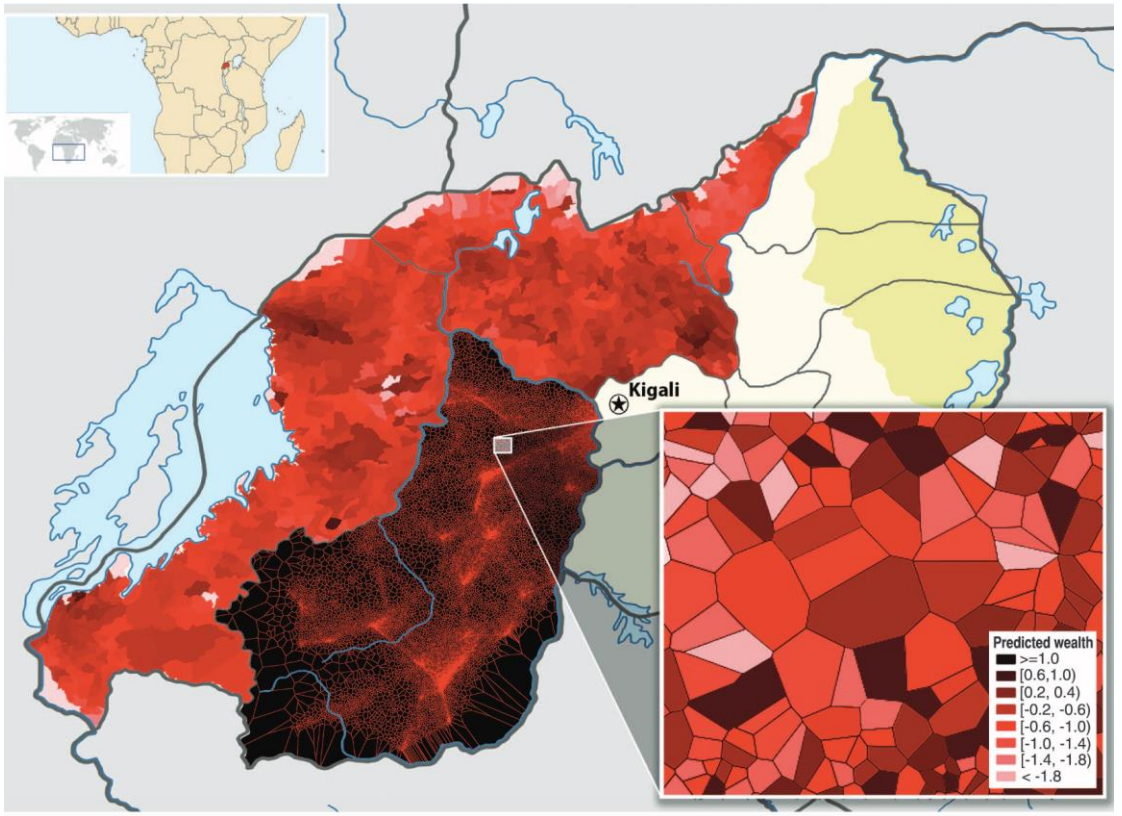
Next: Prediction model

- Instead of all separate survey variables, create composite wealth index.
- (again cross-validate)
- Predict the wealth index score of all 1.5 million people in the call records.



Next: Prediction model

- Predict the wealth of all 1.5 million people in the call records.
- Geospatial information from the call data gives an estimate of the geographic distribution of wealth at an extremely fine spatial granularity.
- Can estimate wealth of each of Rwanda's **neighborhood**.



Why be skeptical?

What types of errors are introduced by this procedure?

Why be skeptical?

- Predictions at individual level are noisy.
- People with mobile phones are systematically different from people without mobile phones.
    - Especially when it comes to wealth → **coverage** error
- Errors have been introduced during the feature selection and supervised learning step.
    - **Algorithmic** error for the measurement of wealth.

Comparison to high quality survey

- Survey considered the gold standard.
- Estimates were very similar.
- But this method: 10x faster and 50x cheaper.
- With the budget of the survey, which is done every couple of years, you can do this every month.

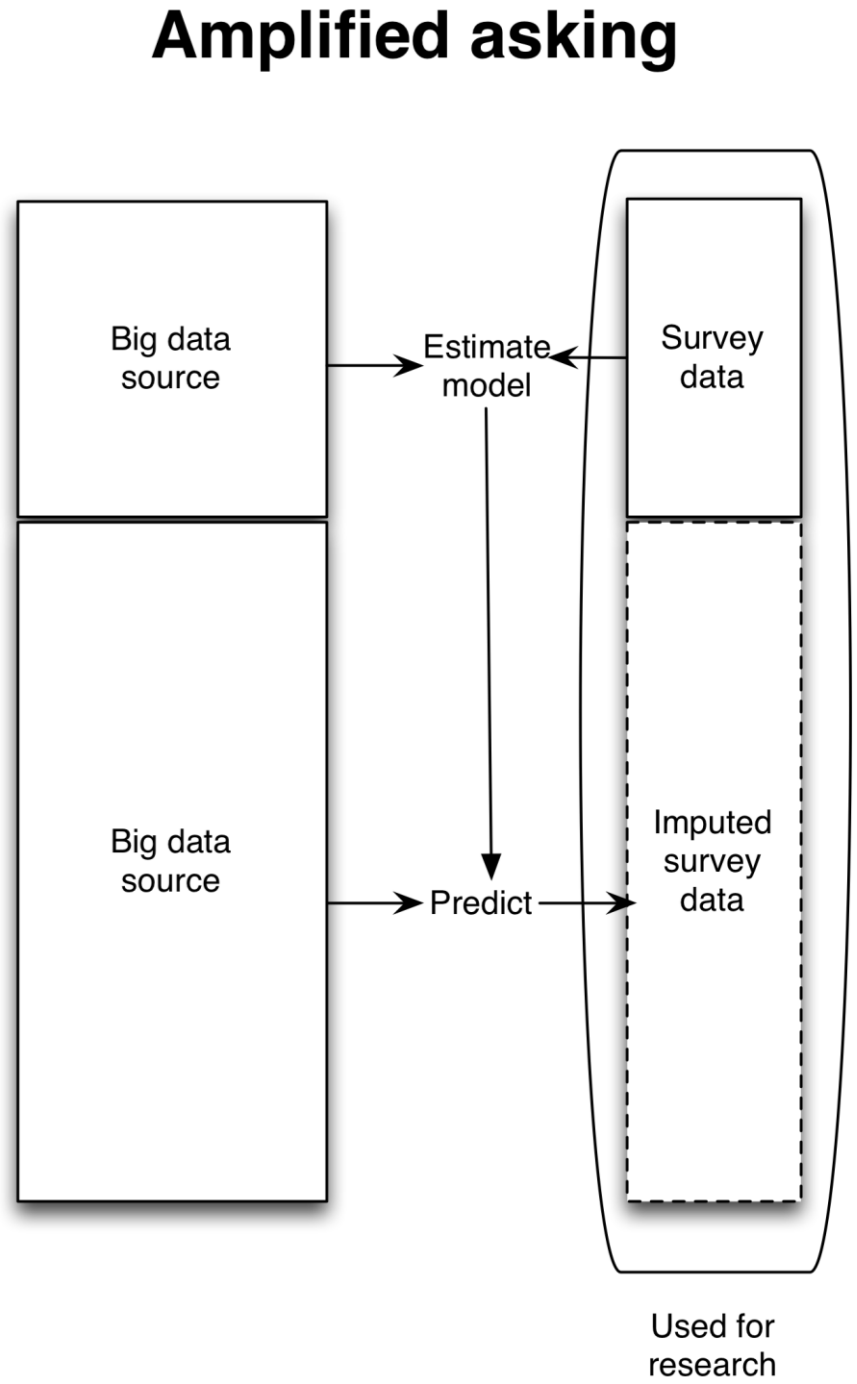
Trade-offs:

- No strong theoretical basis for this kind of approach.
- Do not know when this will work and when not.
- Especially coverage bias is/can be a big issue.

General conclusion

When you have
1. Data source has many variables but of few people
2. And one has few variables of many people

You can do:
1. For the people in both data sources, build a machine learning model that uses digital trace data to predict survey answers
2. Use that model to infer the survey answers of everyone in the big data source.

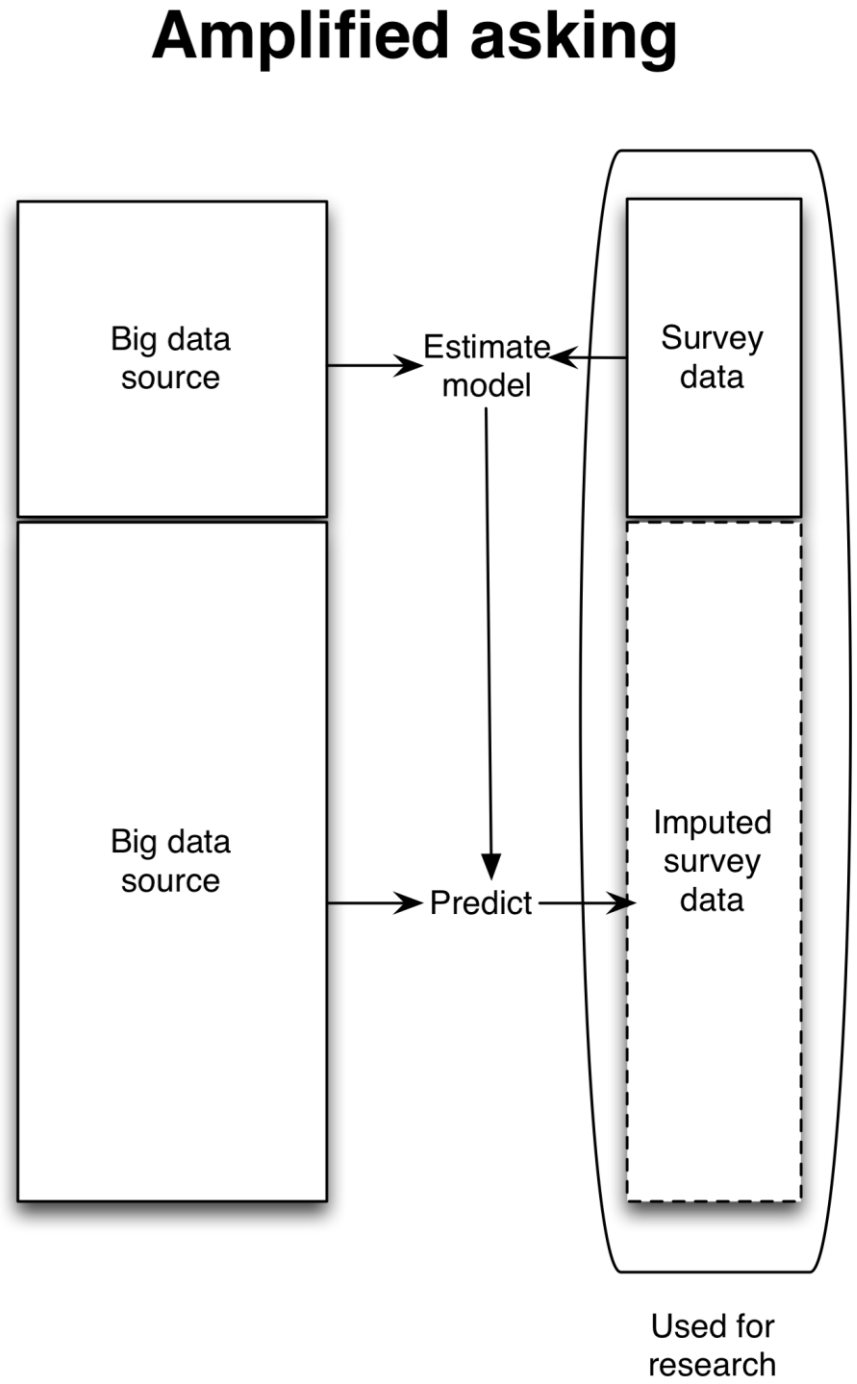


Generalize

Can use this approach for many different types of (combinations of) datasets

E.g. with a user-centric approach, we can:

- Have many people fill in a survey.
- A small group does the app/donation/etc.
- Make a model for the survey part.
- Use that to make predictions for the dtd part.
- Tailor your survey to that it can optimally inform the model.



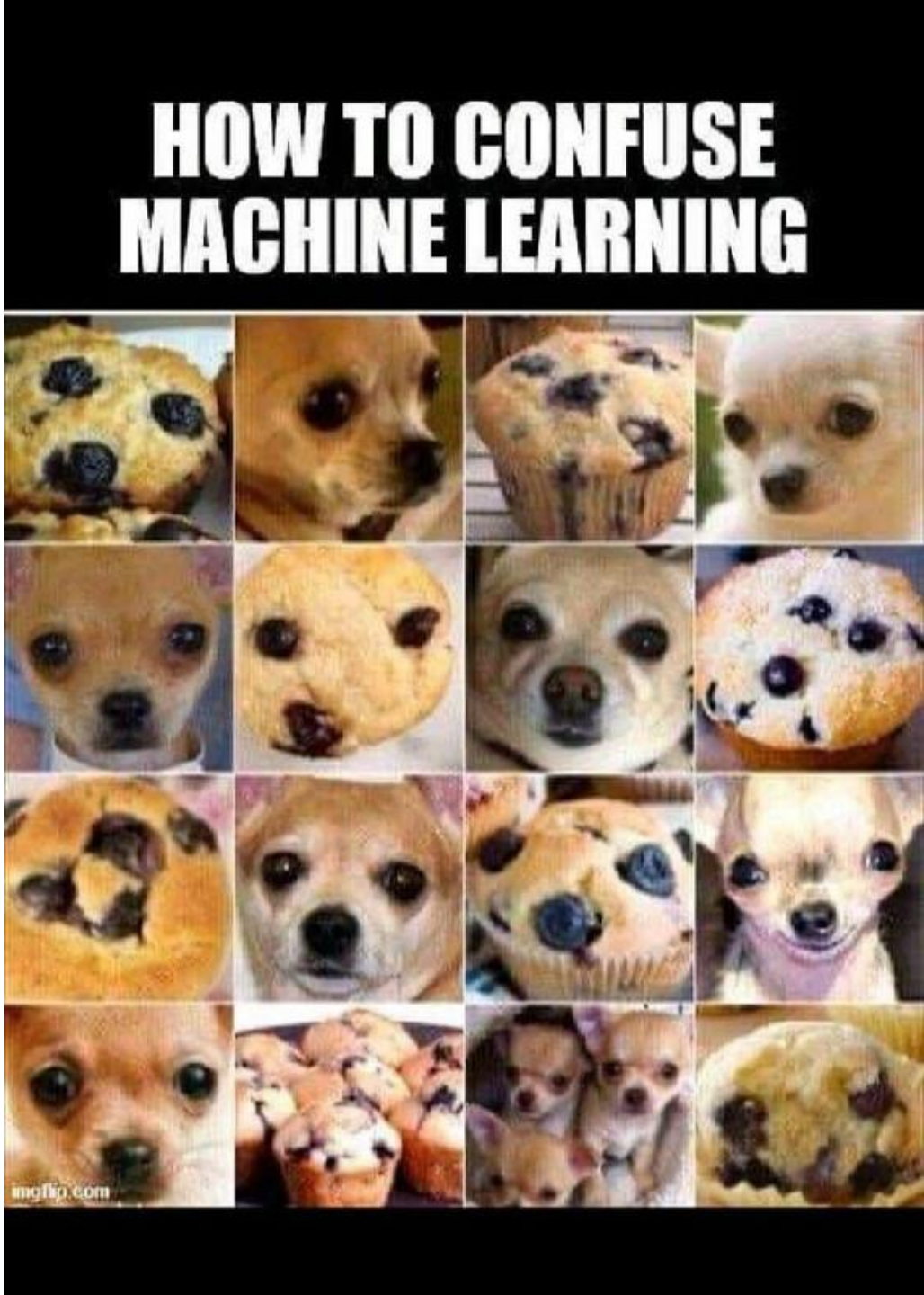
More we can learn from survey methodology

Model building for classification tasks

Consider a model that:

- Classifies whether a website contains “news”
- Classifies whether a message is “happy” or “sad”
- Classifies the content of an image

- Needed when working with digital trace data.
- We need to process all those TikTok videos, messages, browsing histories, etc.



Source

Model building for classification tasks

Researchers use:

- Established models (huggingface)
- Improve existing models using re-enforcement learning
- Build a completely new model.

What can we learn?

You can see this as a missing data problem!

There are many techniques to handle missing data, and they come with assumptions

Models rely on labelled data

Labelling is like filling in a survey

Important:
1. Labels need to be correct (like measurement)
2. Labels need to come from a diverse group of labellers (like representation)

1. Labels need to be correct (measurement)

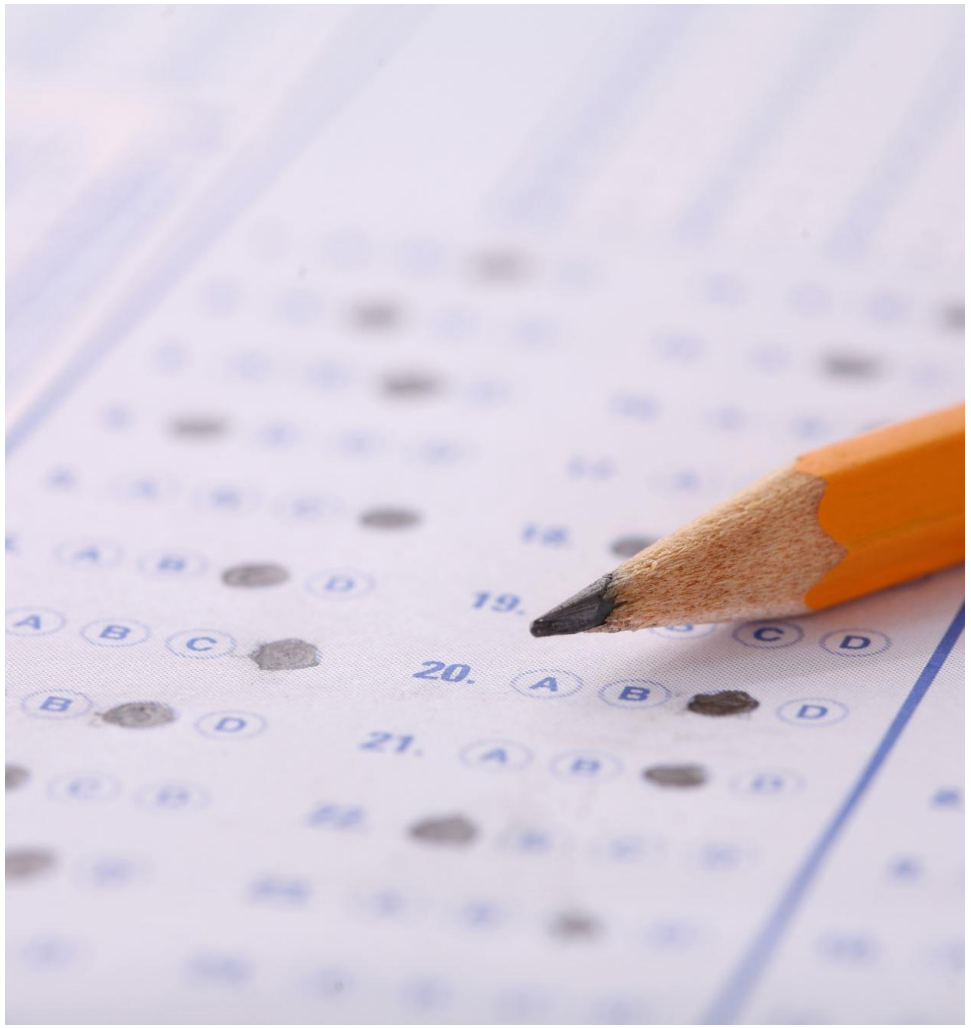
The labeller needs to understand the labelling task, consists of 4 elements:

1. **Comprehension**: Understand the task and the response options
2. **Retrieval**: Search memory for relevant information
3. **Integration**: Integrate the retrieved information to form a response to the task
4. **Mapping**: Map that response onto the provided response choices

What can go wrong?

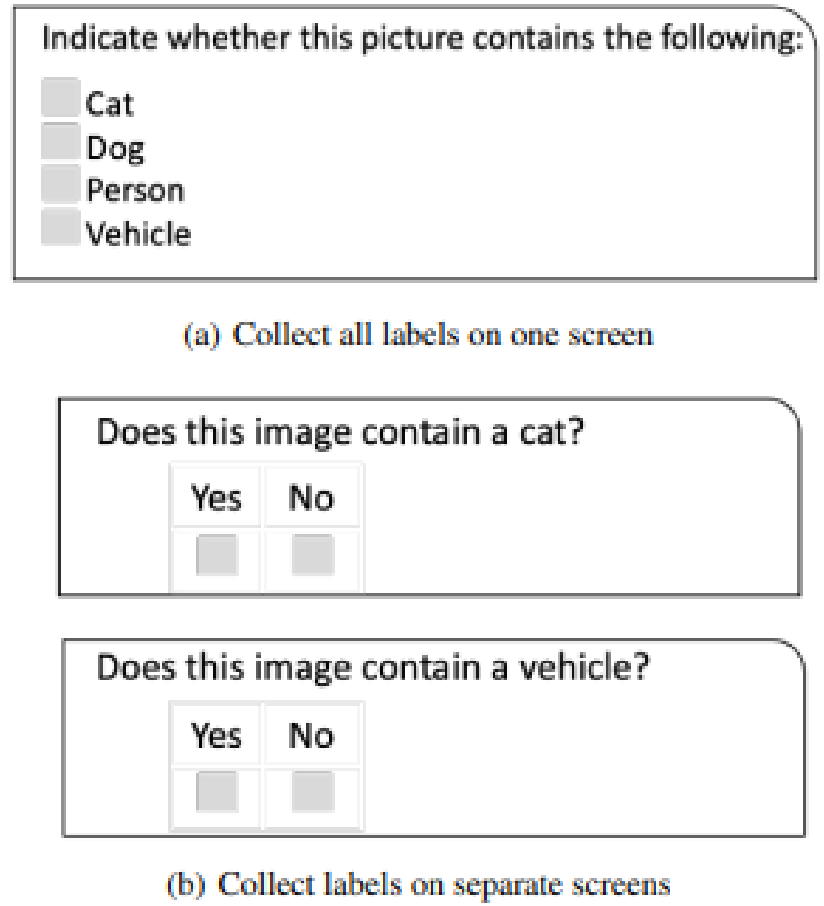
What can go wrong?

- **Satisficing** → take shortcuts
- **Recency bias** → only recall recent info
- **Acquiescence** → "Yes" to all
- **Straightlining** → Same to all
- **Order effects** → Always select first option



Labelling designs that affect label quality

- Wording and reading level
- Including multiple labels →
- Don't know option (what if they really don't know?)
- Order effect (of the tasks)
- Pre-labelling → can stimulate Acquiescence



How to overcome these issues?

- Randomization of observations
- Instrument testing (test label design with a small group)
- Retain paradata
- Feedback to labellers
- Test observations (trick questions)

The labellers (representation)

It can be a problem if the group of labellers differs from the population affected by the models.

- Labels provided may not represent the views of the population
- Can happen if labellers have different characteristics than the population
- Known as **selection bias**.

Labeller characteristics

Is the propensity to participate in labelling correlated to what is being measured by labelling?

Solutions:

- Diversify the labeller pool
- Give more extensive examples and instructions
- Correct for selection issues using weighting (you need to know their characteristics).

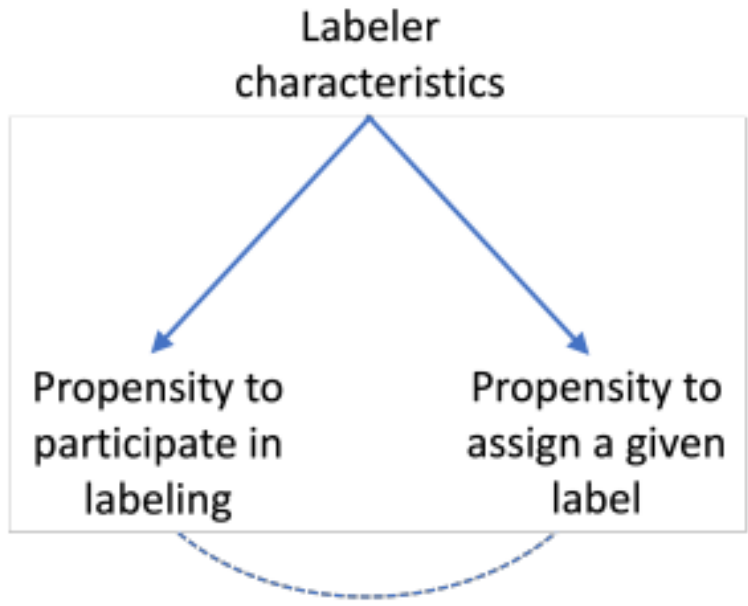


Figure 4. Labeler characteristics induce correlation between propensities (adapted from Groves, 2006)

Conclusion

Errors can be present in every aspect when using DTD to answer research questions.

There is a lot we can learn from survey research:

- When creating labels to train models
- When understanding the quality of digital traces obtained
- To supplement DTD collected through user-centric approaches
- Or to supplement population scale DTD.

Existing statistical techniques can help to account for those errors

More about the upcoming weeks after lunch!