

# Garbage in, garbage out?

Measuring error rates and data quality to get ready for AI

#Rosenverse

Caroline Jarrett

@cjforms.bsky.social

# My book is about Total Survey Error



# I got interested in better forms because of errors

The image displays three examples of data entry forms, each with the label 'Amount' followed by a series of input boxes. The first example shows a negative value '-10' with four empty boxes preceding it. The second example shows the value '12578163' with each digit in its own box. The third example shows the word 'NONE' in the last four boxes, with three empty boxes preceding it.

Amount     - 1 0

Amount 1 2 5 7 8 1 6 3

Amount    N O N E

# For “AI”, data quality matters

[Assessing if artificial intelligence is the right solution | Centre for Digital Public Services](#)

## Consider your current data state

For your AI model to work, it needs access to a large quantity of data.

Work with specialists who know your data to assess your data state, such as data scientists.

You can assess whether your data is high enough quality for AI using a combination of:

- accuracy
- completeness
- uniqueness
- timeliness
- validity
- sufficiency
- relevancy
- representativeness
- consistency

# Agenda

Think about fudging forms

Share what we consider to be an error

Try using “Six ways to think about errors”

Think about error rates

Meet the data quality framework

Think about what we might do differently

# I've observed that people “fudge” through forms

To fudge a form:  
answer one or more questions  
in a way that's not entirely truthful



# Let's try this question

## 2. Work and education

About how many miles is it between your home and your work?

☒

Enter number of miles

Miles

☐

Less than one mile

☐

Do not know

☐

Prefer not to answer



[How to answer this question](#)

# Share an example

I had to give a wrong answer to that question.  
If you've ever given an inaccurate answer on another question, please tell me about it in this EasyRetro.

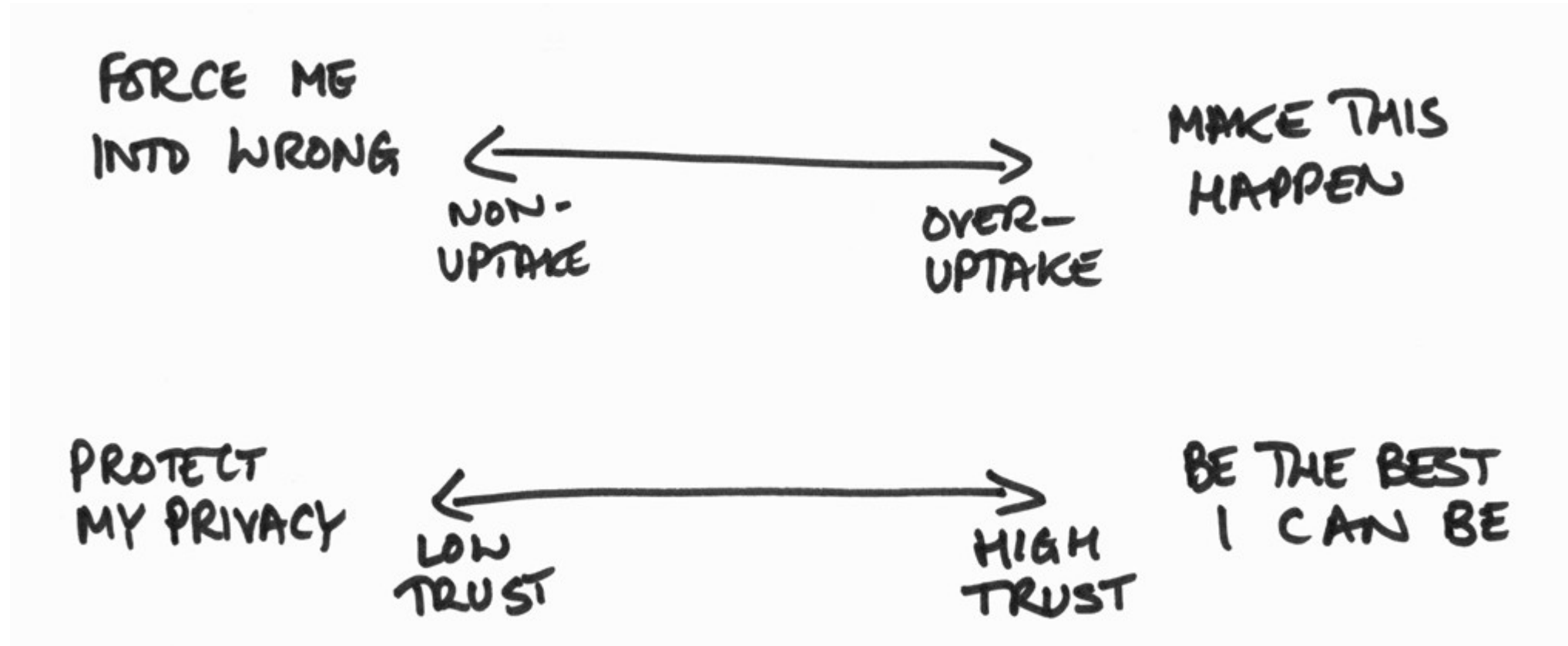
[bit.ly/gigo25](https://bit.ly/gigo25)

2 minutes





# I'm thinking about why questions get wrong answers



# Takeaway

There are many ways in which the data we collect isn't entirely accurate

# Agenda

~~Think about fudging forms~~

Share what we consider to be an error

Try using “Six ways to think about errors”

Meet the data quality framework

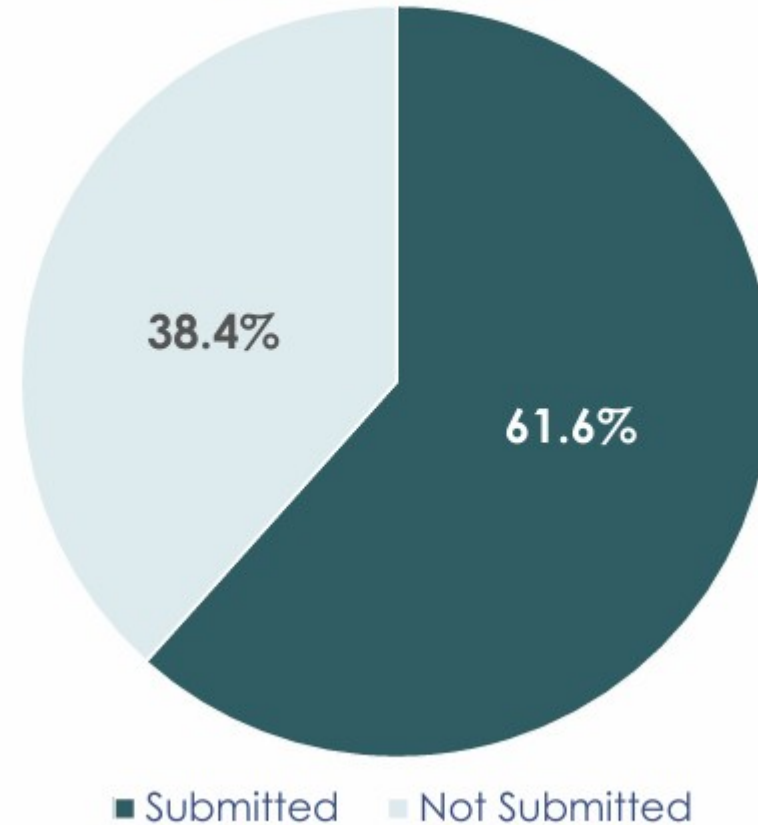
Think about what we might do differently

# Share what we consider to be an error

# Some people don't get to the end of their application

People on low incomes or no income in California can apply for government money to buy fresh food (CalFresh).

PROPORTION OF APPLICATIONS SUBMITTED  
(NOV – DEC)



# The result matters more than the journey

People rarely fill in forms for fun

We're aiming to achieve some result



# Let's learn from our experiences with errors

Add a card to the EasyRetro

- a brief description of a service that your team is working on
- an example of something that you consider is an error in that service

Please write one card for each error



2 minutes



# Agenda

~~Think about fudging forms~~

~~Share what we consider to be an error~~

Try using “Six ways to think about errors”

Think about error rates

Meet the data quality framework

Think about what we might do differently



# Here are some ways to think about errors

Thanks to the team at HMRC who inspired this

Errors	Examples
Problems along the way	Tried to buy 2Kg carrots, ordered 2 carrots
Wrong result	Ordered pork pie, got lettuce
Unnecessary action	No delivery arrives, called to find out why
Delayed-impact problem	Credit card OK when placed order, failed later
Non-uptake	Decide to use a different supermarket
Over-uptake	Accidentally placed order twice
Technology issue	Supermarket goes offline due to cyber-attack
Something else	?

# There's a template you can use

Problems along the way	Unnecessary action	Wrong result	Delayed-impact problem
Non-uptake	Over-uptake	Technology problem	Something else?

# Agenda

~~Think about fudging forms~~

~~Share what we consider to be an error~~

~~Try using “Six ways to think about errors”~~

Think about error rates

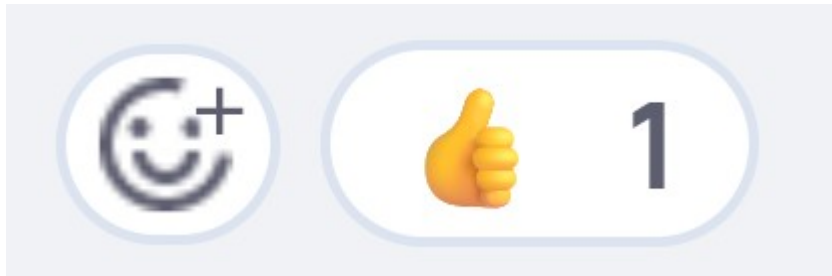
Meet the data quality framework

Think about what we might do differently

# Think about error rates

# And now, what about error rates?

If you know the error rate for any of the errors in your service, add a thumbs-up



1 minute



# Only about a third of errors had a known error rate



# Completion rate is a relatively simple calculation

(also known as conversion rate)

$$\text{Completion rate} = \frac{\text{Number of people who complete}}{\text{Number of people who start}}$$

## Error rate might also be simple

$$\text{Error rate} = \frac{\text{Number of errors}}{\text{Number of people who start}}$$



## Other error rate calculations are available

$$\text{Error rate} = \frac{\text{Number of attempts with an error}}{\text{Number of people who start complete are eligible other?}}$$

# Example: there are at least 3 definitions of turnout

In an election, a failure to vote is an error

$$\text{Turnout} = \frac{\text{Number of votes counted}}{\text{Number of people who}}$$

1. Send in postal votes or go in person
2. Are on the electoral roll
3. Are in the voting-eligible population

# Takeaway

To track your error rates:

- Count the errors
- Think hard about what you divide by

# Agenda

~~Think about fudging forms~~

~~Share what we consider to be an error~~

~~Try using “Six ways to think about errors”~~

~~Think about error rates~~

Meet the data quality framework

Think about what we might do differently

# Previous takeaway

There are many ways in which  
the data we collect isn't  
entirely accurate

# Data quality can deteriorate or change all by itself

## People change

- Lose their credit card
- Move house
- Change jobs
- And many others

## Organisations change

- Delete an old product
- Add a new one
- Are affected by energy costs
- And many others



# Missing data can have serious effects

The Home Office destroyed the documents that migrants later needed to prove their right to remain



## 'Windrush' migrants facing deportation threat

Michael Braithwaite was let go after his employer ruled him an illegal immigrant.

Like many other members of the "Windrush generation" of Caribbean migrants who moved to the UK before the 1971 Immigration Act, Michael had the right to remain but no documents to prove it.

'Windrush generation' deportation threat

UK · 13 April 2018

['Windrush' migrants facing deportation threat - BBC News](#)

# Previous takeaway

There are many ways in which  
the data we collect isn't  
entirely accurate



# New takeaway

There are many ways in which  
the data we use isn't  
entirely accurate

# What is about long-term data quality?

Are there data quality issues that might arise over time?

Add a card for any issue that you think of

3 minutes



# There is a UK Government Data Quality Framework



# For “AI”, data quality matters

[Assessing if artificial intelligence is the right solution | Centre for Digital Public Services](#)

## Consider your current data state

For your AI model to work, it needs access to a large quantity of data.

Work with specialists who know your data to assess your data state, such as data scientists.

You can assess whether your data is high enough quality for AI using a combination of:

- accuracy
- completeness
- uniqueness
- timeliness
- validity
- sufficiency
- relevancy
- representativeness
- consistency

# Takeaway

Data quality is important  
but not easy

# Agenda

~~Think about fudging forms~~

~~Share what we consider to be an error~~

~~Try using “Six ways to think about errors”~~

~~Think about error rates~~

~~Meet the data quality framework~~

Think about what we might do differently

# Add ideas for doing something differently

There's a column for anything that you might decide to do differently

[bit.ly/gigo25](https://bit.ly/gigo25)

5 minutes



# Takeaway

AI may help to start useful conversations about error rates and data quality



# I'm starting to collect some resources

## General advice on measurement

Steve Messer [Metrics, measures and indicators: a few things to bear in mind](#)

## Official / government guidance

ONS [Quality in official statistics - Office for National Statistics](#)

## Headline stories

The Guardian [Windrush victims could have compensation reconsidered](#)

Disability News [DWP helped cause mental distress of claimant who took her own life](#)

## Academic research

The King's Fund [Lost In The System: The Need For Better NHS Admin](#)

# Keep up the conversation: Caroline Jarrett

BlueSky @cjforms.bsky.social

[cj@effortmark.co.uk](mailto:cj@effortmark.co.uk)

[www.effortmark.co.uk](http://www.effortmark.co.uk)

