Data Science Campus

# Ethical AI in Government -Validation / Maintenance

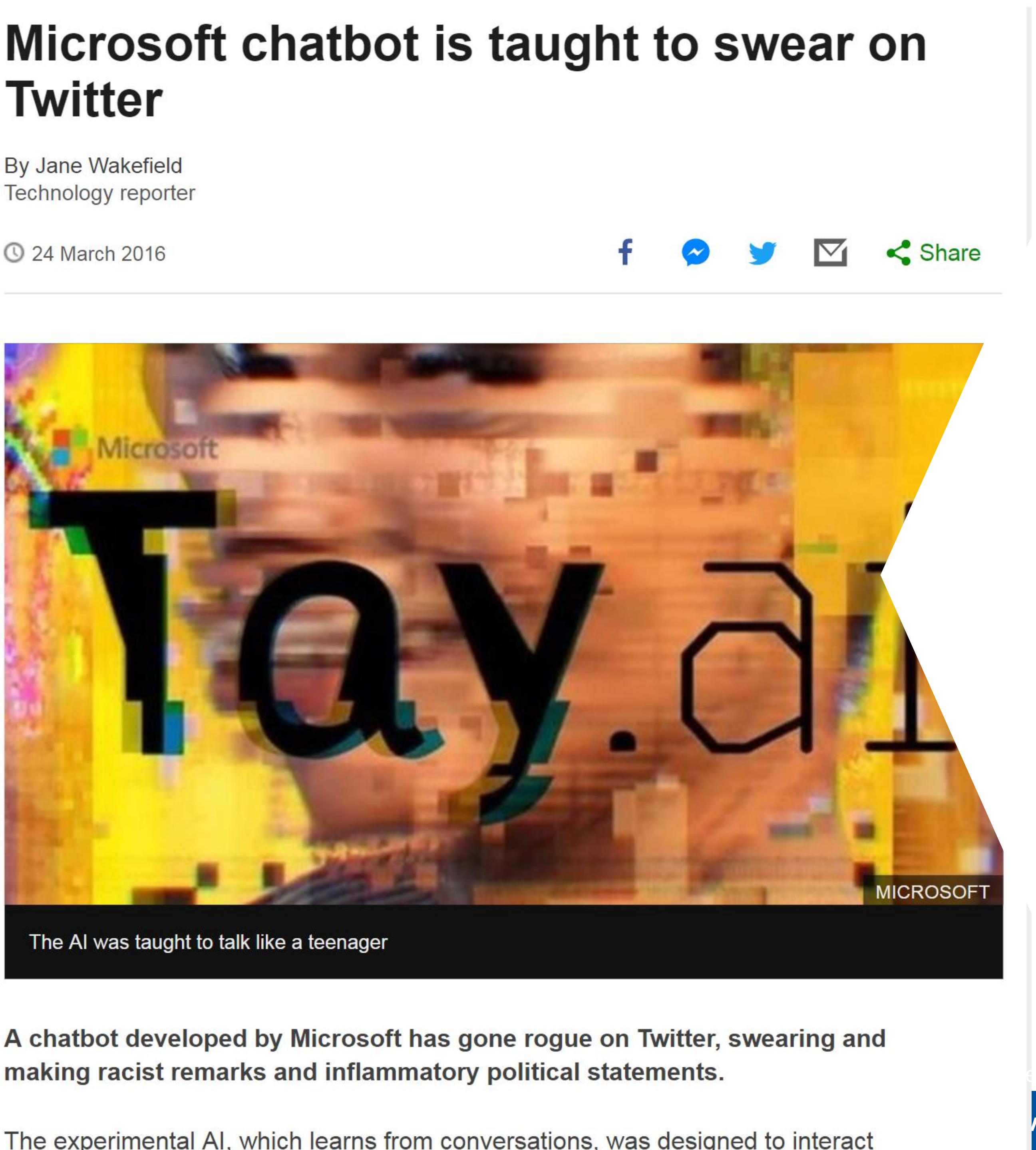
**Jasmine Grimsley**
Senior Data Scientist
Data Science Campus, ONS, UK

Agent to Agent interactions

# Even good bots fight: The case of Wikipedia

https://www.theguardian.com/technology/2017/feb/23/wikipedia-bot-editing-war-study

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171774



# Microsoft chatbot is taught to swear on Twitter

By Jane Wakefield
Technology reporter

24 March 2016

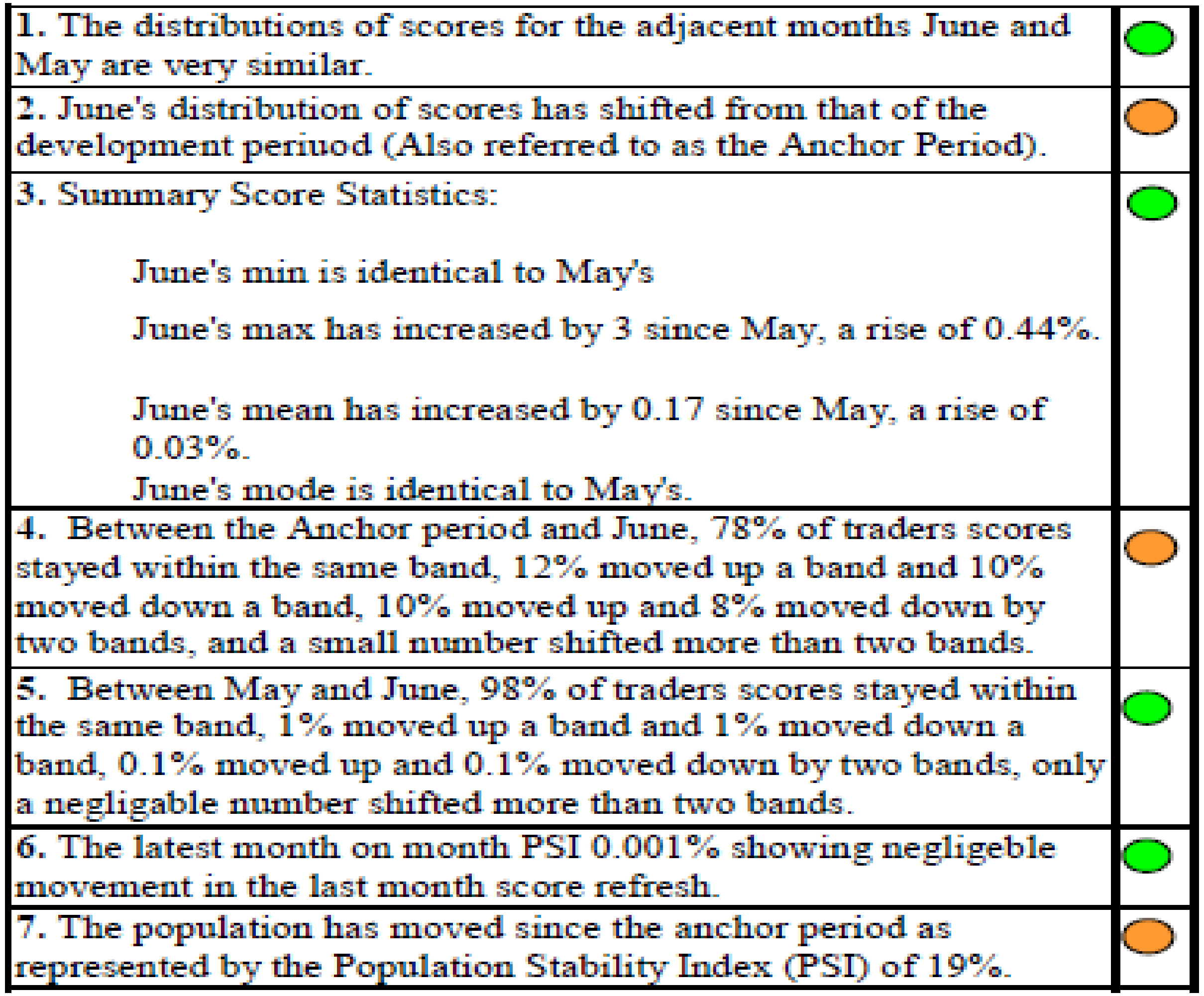
The AI was taught to talk like a teenager

**A chatbot developed by Microsoft has gone rogue on Twitter, swearing and making racist remarks and inflammatory political statements.**

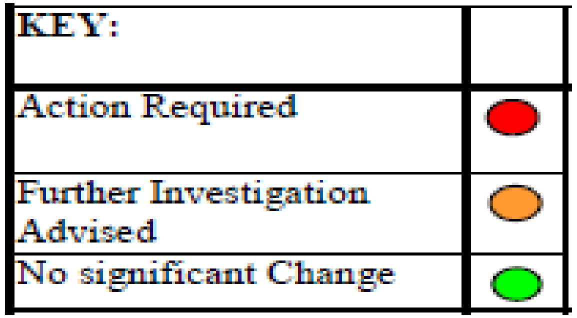
The experimental AI, which learns from conversations, was designed to interact

Models can become Unethical over time

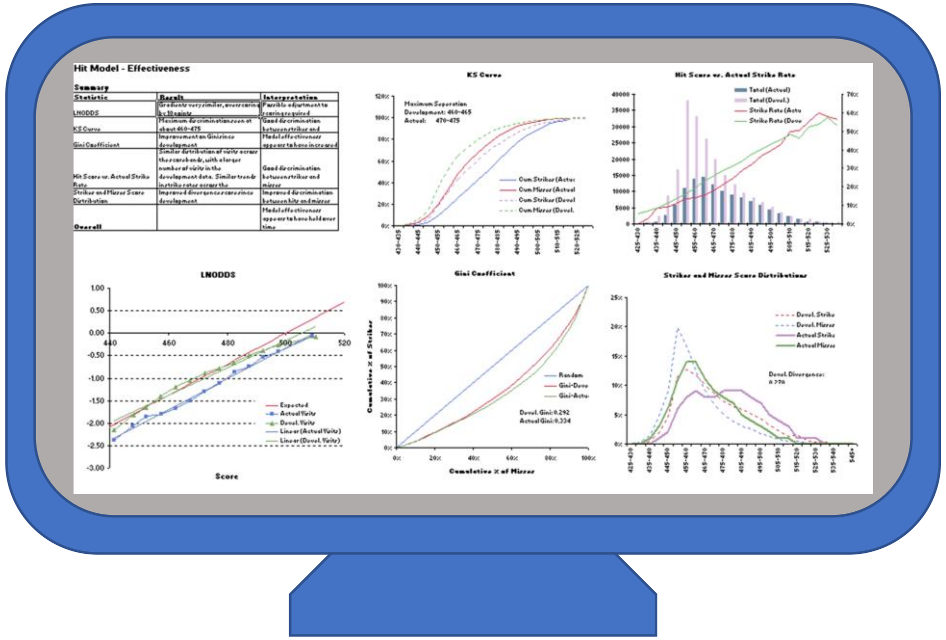
# Dashboard Approach



| | |
|---|---|
| 1. The distributions of scores for the adjacent months June and May are very similar. | 🟢 |
| 2. June's distribution of scores has shifted from that of the development periuod (Also referred to as the Anchor Period). | 🟠 |
| 3. Summary Score Statistics:<br>June's min is identical to May's<br>June's max has increased by 3 since May, a rise of 0.44%.<br>June's mean has increased by 0.17 since May, a rise of 0.03%.<br>June's mode is identical to May's. | 🟢 |
| 4. Between the Anchor period and June, 78% of traders scores stayed within the same band, 12% moved up a band and 10% moved down a band, 10% moved up and 8% moved down by two bands, and a small number shifted more than two bands. | 🟠 |
| 5. Between May and June, 98% of traders scores stayed within the same band, 1% moved up a band and 1% moved down a band, 0.1% moved up and 0.1% moved down by two bands, only a negligable number shifted more than two bands. | 🟢 |
| 6. The latest month on month PSI 0.001% showing negligeble movement in the last month score refresh. | 🟢 |
| 7. The population has moved since the anchor period as represented by the Population Stability Index (PSI) of 19%. | 🟠 |



| KEY: | |
|---|---|
| Action Required | 🔴 |
| Further Investigation Advised | 🟠 |
| No significant Change | 🟢 |



Hit Model - Effectiveness

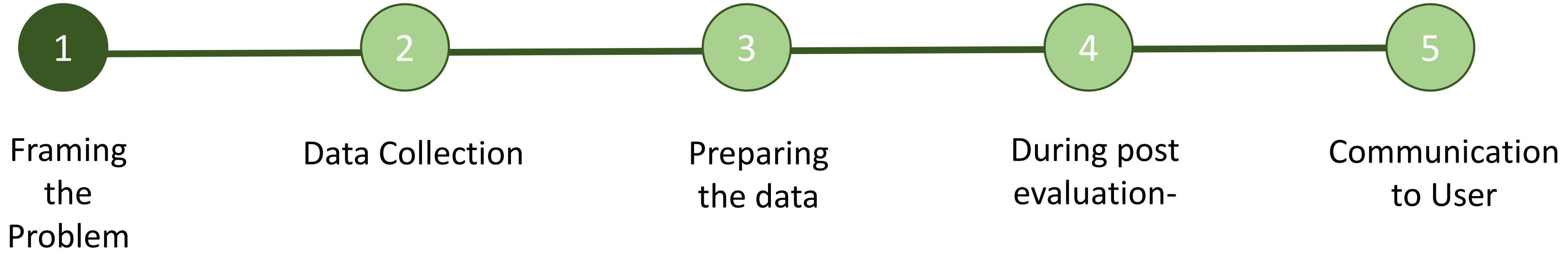
Summary

| Statistic | Result | Interpretation |
|---|---|---|
| LNODDS | Gradients very similar, over scoring by 10 points | Possible adjustment to scoring required |
| KS Curve | Maximum discrimination seen at about 460-475 | Good discrimination between strikes and |
| Gini Coefficient | Improvement on Gini since development | Model effectiveness appears to have increased |
| Hit Score vs. Actual Strike Rate | Similar distribution of visits across the score bands, with a larger number of visits in the development data. Similar trends in strike rates across the | Good discrimination between strikes and misses |
| Strikes and Misses Score Distribution | Improved divergence scores since development | Improved discrimination between hits and misses |
| Overall | | Model effectiveness appears to have held over time |

Shakeel Khan. PA Handbook http://asd.inrev.gov.uk/Content/PDF/PredictiveAnalysisHandbook_V3.3.pdf

# Why might Santa only be distributing coal this winter?

# When to Incorporate Ethical Maintenance



1 Framing the Problem

2 Data Collection

3 Preparing the data

4 During post evaluation-

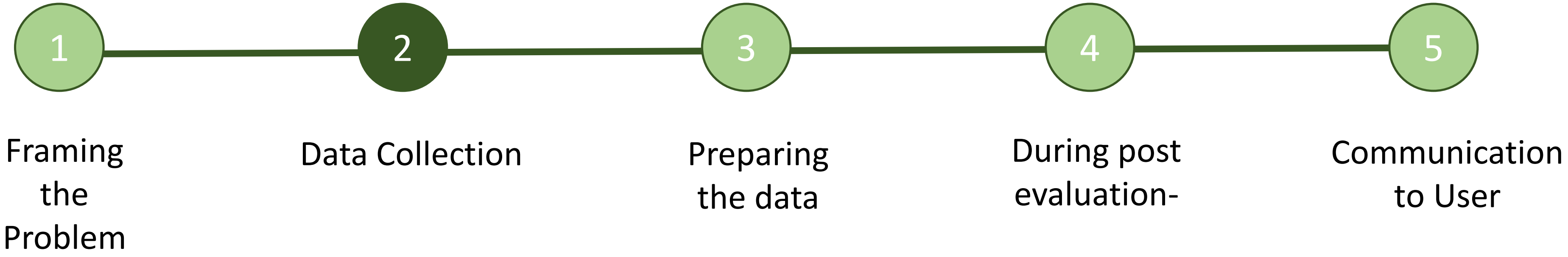
5 Communication to User

Is it my job as a Data Scientist or the Stakeholders?

Knowledge of future changes in the field?

Risky variables?

# When to Incorporate Ethical Maintenance



1 Framing the Problem

2 Data Collection

3 Preparing the data

4 During post evaluation-

5 Communication to User

**Stakeholder engagement**

Do I have fair representation of groups?

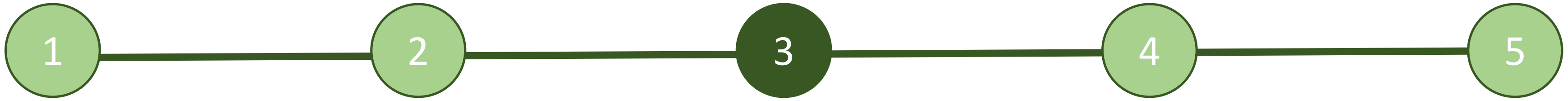
Can I generalise my data set to the population of interest?

Does this need a periodic review?

Would the user like this in a visualisation or report?

Clear agreements with data providers regarding format changes or data discontinuation

# When to Incorporate Ethical Maintenance



1 Framing the Problem

2 Data Collection

3 Preparing the data

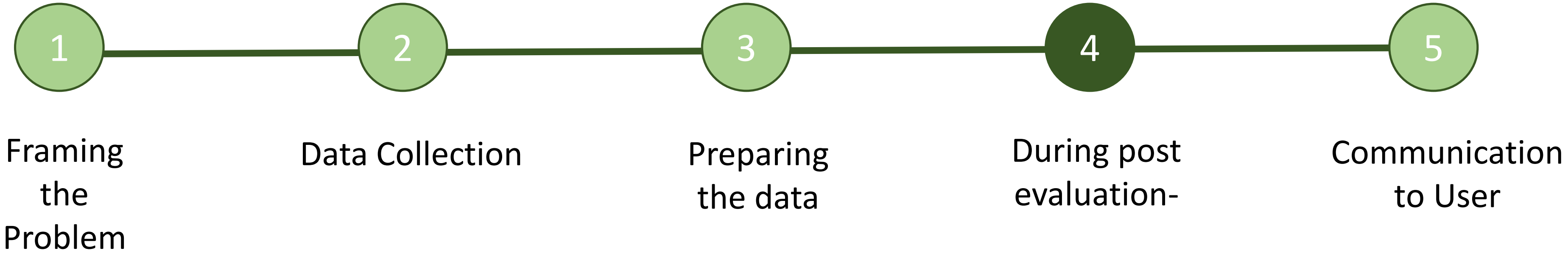
4 During post evaluation-

5 Communication to User

Carful of making a new variable highly correlated with a sensitive one when using dimension reduction- Correlation report?

How much data is missing or being imputed?

# When to Incorporate Ethical Maintenance



1 Framing the Problem

2 Data Collection

3 Preparing the data

4 During post evaluation-

5 Communication to User

Are their changes in proportional classification, is a change threshold needed?

How distinct are groups?

Have the maintenance processes for your AI model adequately taken into account the possibility of distributional shifts in the underlying population?

Has the model been properly retuned and re-trained to accommodate changes in the environment?

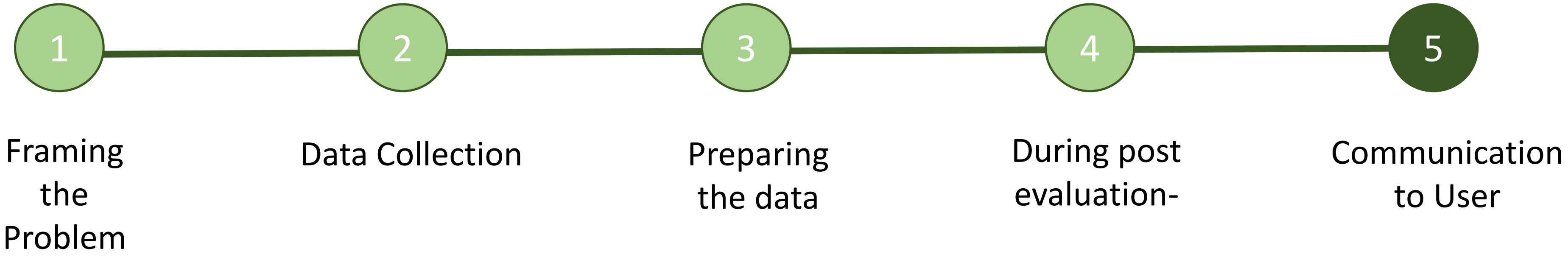
Data Science Campus

# We are safe. Santa has a risk management strategy.

| | |
|---|---|
| 🟥 | Action required |
| 🟨 | Further investigation required |
| 🟩 | No Significant change |

| | **Monthly Model Report** |
|---|---|
| 🟨 | > 5% of children are on the good list |
| 🟥 | No new names were added in Nov 2019 |
| 🟩 | Gift storage is 80% full |
| 🟥 | Annual review of new data sources 100 months overdue |

# When to Incorporate Ethical Maintenance



1 Framing the Problem

2 Data Collection

3 Preparing the data

4 During post evaluation-

5 Communication to User

## Clear on why an outcome has happened

Dear Jane,

This year you have Coal,

this is due to your never

cleaning the slate after class

Love

Santa

Dear Jane,

This year you have Coal,

this is due to engaging in

trolling behaviour on Twitter

Love

Santa

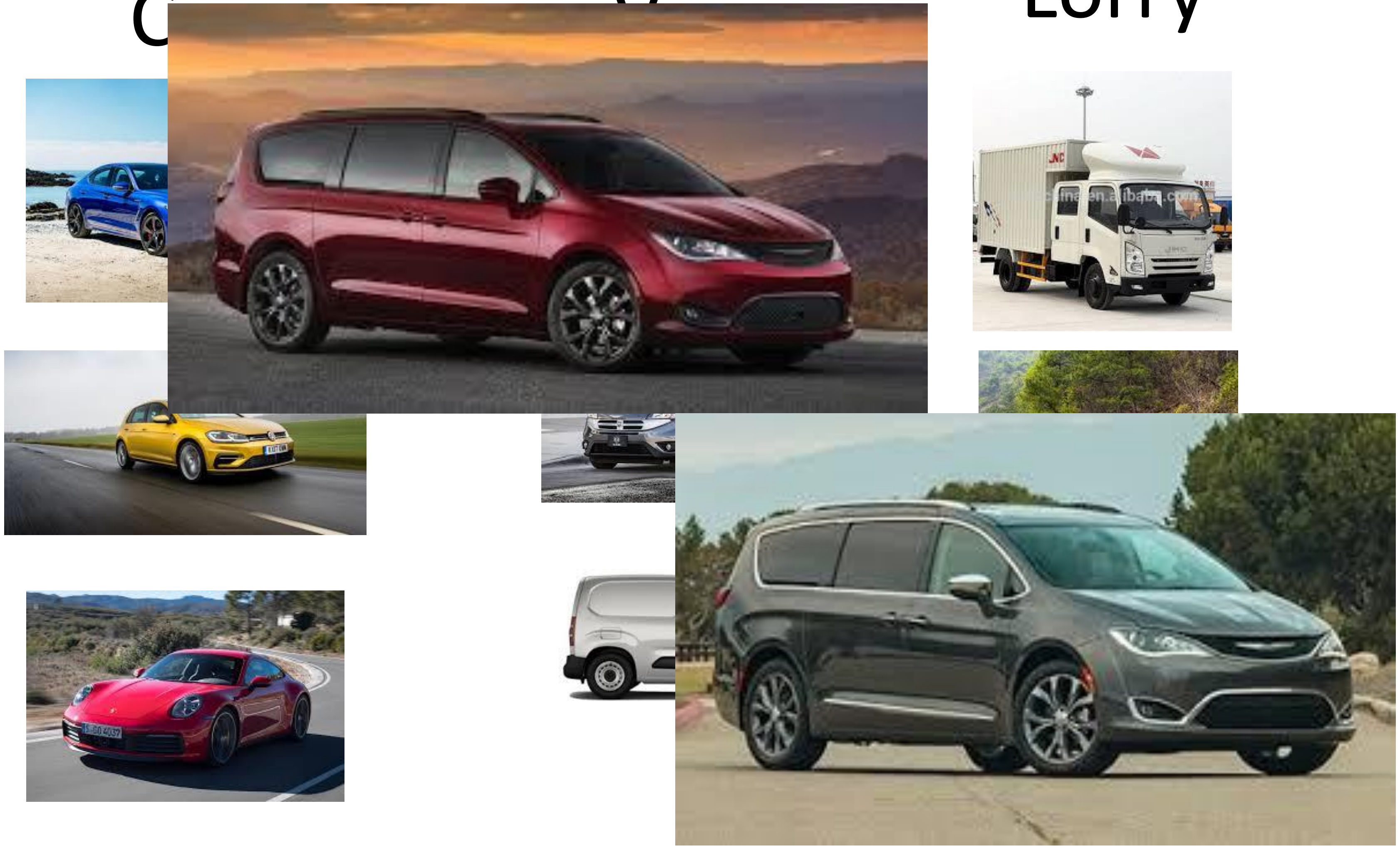
# How might this effect policy?

Vehicle classification

C

V

Lorry



Van use is up and is correlated with increased pollution near schools!

Policy= Light Goods Vehicle (LGVs) with a Euro 5 or earlier engine can not enter zone

Data Science Campus

# Thank you for listening

## Data Science Campus, ONS, UK

Governmen Data Science Conference, London, 2017