# Privacy, Security and Professional Ethics

UBCO Master of Data Science – DATA 553

SOME SLIDES FROM DATA 553 – DR FATEMEH FARD – MDS OKANAGAN
INSPIRED FROM DR. JAGADISH DATA SCIENCE ETHICS COURSE

# Data validity

Data validity in statistics

Inappropriate use of Data Science methods leading to erroneous conclusions

Representative sample:

- Opinionated customers, for example, are not necessarily representative of all customers

Validity:

- Attributes and measures
- Model design
- Data processing
- Managing change

# Representative sample

Data we have can obtain may not be the data we wish to have for the analysis

Compensation using statistical techniques
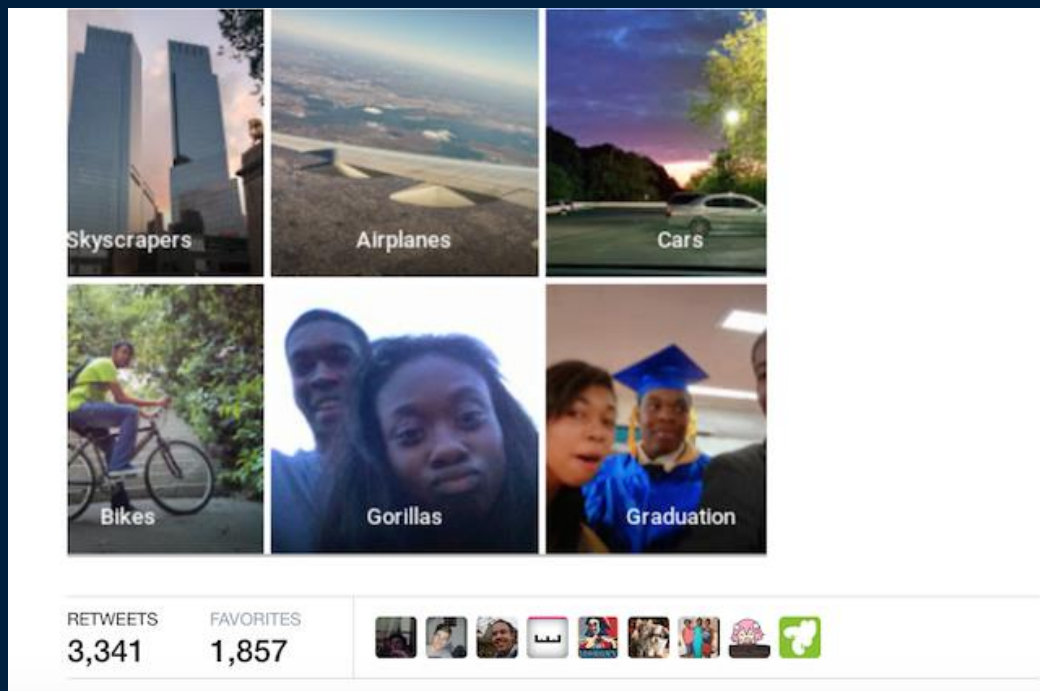- Weighting of the samples

Training data
- Based on current data – enough samples representation – data augmentation
- Challenge when data evaluated is different from the training set.

3

# Sample representation

Google Photo

# Choice of Attribute

Data Analysis
- Choice of attributes
- Metrics

# Choice of Attribute

Data Analysis
- Choice of attributes
- Metrics

Can dramatically impact the results.

# Errors in Data

Errors in data entry are common:

- Typos
- Wrong order in Likert scale
- Wrong code
- Misunderstanding of a question/field
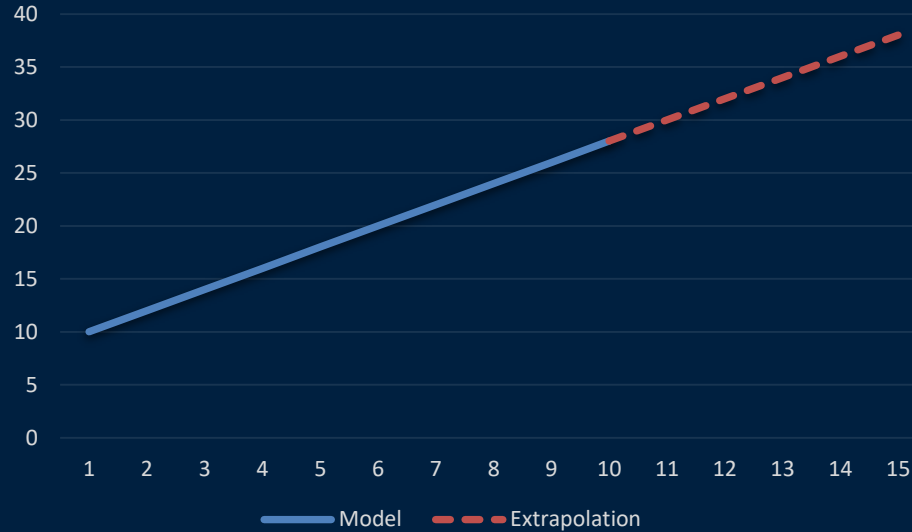
# Errors in Data Processing

The processing of the data can generate errors:

- Extracting subjective categories from text

- Merging records for the same person

- Recognizing faces from photos

# Errors in Model Design

Extrapolation



Can be dangerous to apply if you don't know that the model is correctly chosen

# Group Analysis

Simpson's paradox

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| Total | 8442 | 44% | 4321 | 35% |

https://en.wikipedia.org/wiki/Simpson%27s_paradox

# Group Analysis

## Simpson's paradox

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| A | 825 | 62% | 108 | **82%** |
| B | 560 | 63% | 25 | **68%** |
| C | 325 | **37%** | 593 | 34% |
| D | 417 | 33% | 375 | **35%** |
| E | 191 | **28%** | 393 | 24% |
| F | 373 | 6% | 341 | **7%** |

https://en.wikipedia.org/wiki/Simpson%27s_paradox

# Group Analysis

## Clinical / Pharmaceutical studies

- **Clinical trials**
  - **Not carefully balanced until the 90s**

- **Pharmaceutical trials**
  - **Still arguments on whether the use of female mice is useful or not**
    - 80% of rodent drug studies are conducted only on males
    - Bias towards male subjects found in surveys of pre-clinical animal research on pain, cardiovascular disease, diabetes and surgical methods.
  - **Male bias has led to some major health issues for women**
    - Only the difference in weight in taken into account
    - Not the differences in women's body may metabolize drugs differently
    - Women are 1.5 times more likely to develop adverse reaction to prescriptions than men

# Group analysis

When doing group analysis:

- Think about the population you are studying
- Collect the data on the right population
- Think carefully how you are doing the segmentation of the population

# Managing change

Is the analysis still valid over time?

- Some changes may impact the analysis
- Most may not

Example of system change: "Google Flu Trends"

- Google algorithm was vulnerable to overfitting to seasonal terms unrelated to the flu
- Changes in search behaviour changed overtime was not taken into account

Example of human behaviour change

- Protecting privacy by entering erroneous information (email, dob,…)

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

# Summary

Data can be corrupted

Model don't fit the data perfectly

Both can lead to bad results

Those can have an impact on society, with a risk of real harm.

# Algorithm fairness

AI algorithm have shown how human bias can be introduced on an otherwise 'neutral' algorithm

- Training set
- Algorithm development
- Deployment

# Example of Bias

In 2019, Robert Downey Jr. was the third highest paid actor in Hollywood. *Avengers: Endgame* dominated the box office for much of the year. His co-star in that film, Scarlett Johansson was the highest paid actress in 2019 for the second year in a row.

Google search:

- Robert Downey Jr
  - most related searches they include *Iron Man*, *Sherlock Holmes*, *Avengers*, *Tropic Thunder*, and *Civil War*
- Scarlett Johansson
  - search include "body," "cute," "bed," "photoshoot," "makeup " and Vanity Fair.

Man search focused on work, woman on physical appearance

# Algorithm Fairness

Learning algorithms are inherently bias

- data size
- Unconscious bias in selecting data sets

Algorithm development

- Robustness of algorithm to react to bias

Deployment

- Use of AI technique should be done with full understanding

# Bias

## Interaction Bias

- Bias introduced by the users, driving the interaction
  - Datasets containing the community biases
    - Potential harm

## Latent Bias

- Bias in historical data or stereotype in society
  - Dataset include prejudice bias
    - Facilitate proxy discrimination
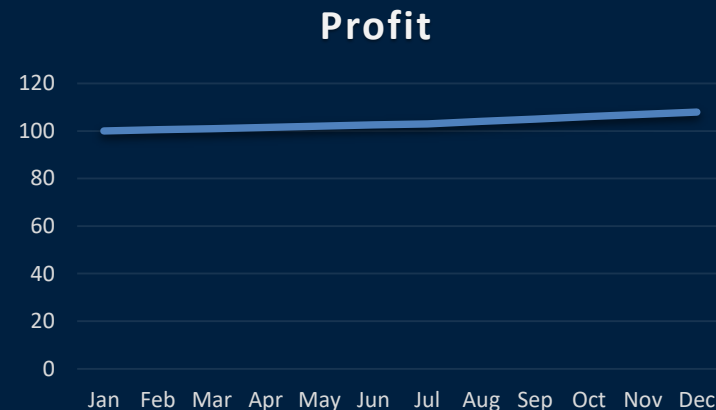    - Reinforce societal differences

## Selection Bias

- Dataset over represents one group and under represents another
  - Dataset build with targeted representation

# Misleading results

## Unfair visualization

**Profit**

**Profit**

## Presenting a single score

- Important data may be suppressed
- Diversity suppression might not always be visible

# Misleading results

Reviews for cell phones

- Restaurant A    3.5  ★★★☆
- Restaurant B    3.4  ★★★☆

# Misleading results

Reviews for cell phones

- Restaurant A      3.5 ★★★☆ with mostly 3 and 4

- Restaurant B      3.4 ★★★☆ with mostly 1 and 5
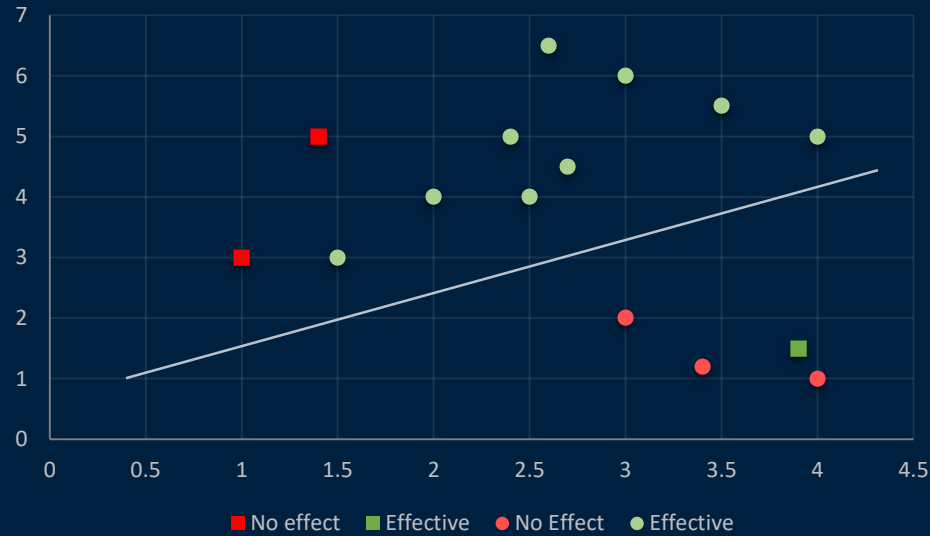
# Misleading results

Reviews for cell phones

- Restaurant A     3.5 ★★★☆ with mostly 3 and 4 (10 reviews)

- Restaurant B     3.4 ★★★☆ with mostly 1 and 5 (200 reviews)
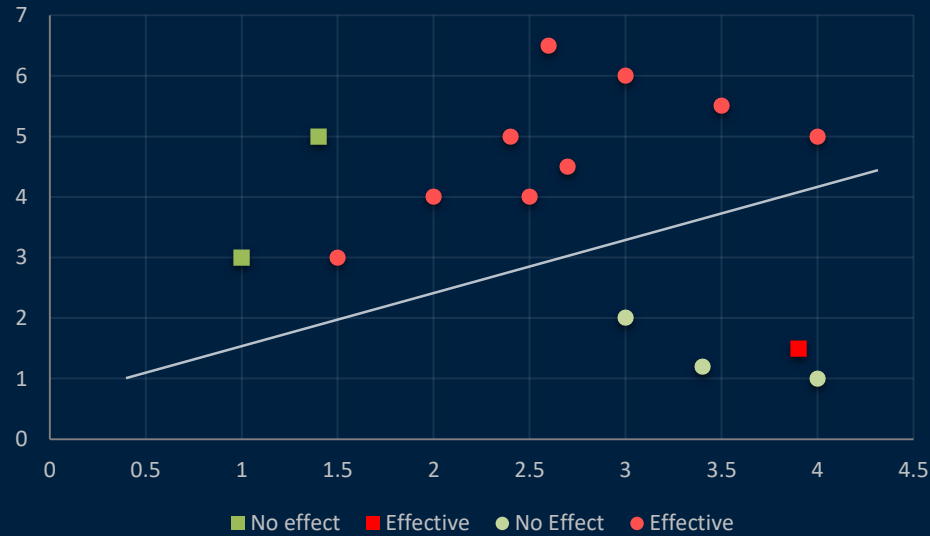
# Misleading results

Minority losses

# Misleading results

Minority losses

# Misleading results

It is not sufficient to have the right results

Doing analysis in an overly simplistic manner can lead to significant societal consequences.

# P Hacking

Hacking the P-value

- Distribution is not bell shaped
    - Must be careful to use appropriate P-value if the shape is different
- Multiple independent tests
    - Careful to use appropriate statistical tools to correct for the multiple tests
- Sequential hypotheses
    - Each slightly different from the previous one
    - Harder to express it mathematically
    - Might not perform multiple tests correction
- Create hypothesis after viewing the data
    - Very common in data science
    - Keep training set separate from evaluation set

# Summary

Human have many biases.

- Even with best intentions!

However, we can minimize risks of bias :

1. Ensure the data used is valid
2. Watch for potential misleading results
3. Use the appropriate mathematical tools during analysis
4. Watch for bias in machine learning algorithms
5. Make algorithms more robust to bias

Never forget: It impacts real people !