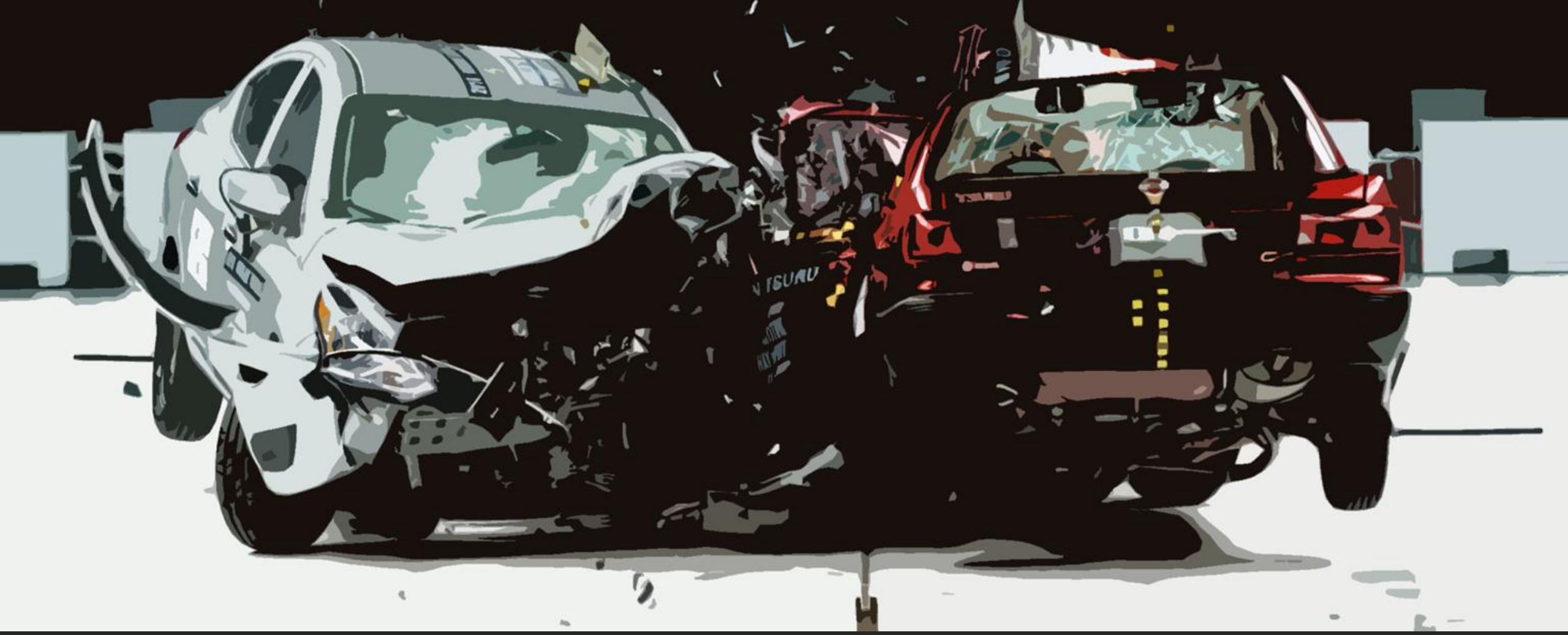


Final Presentation: **Car Safety**

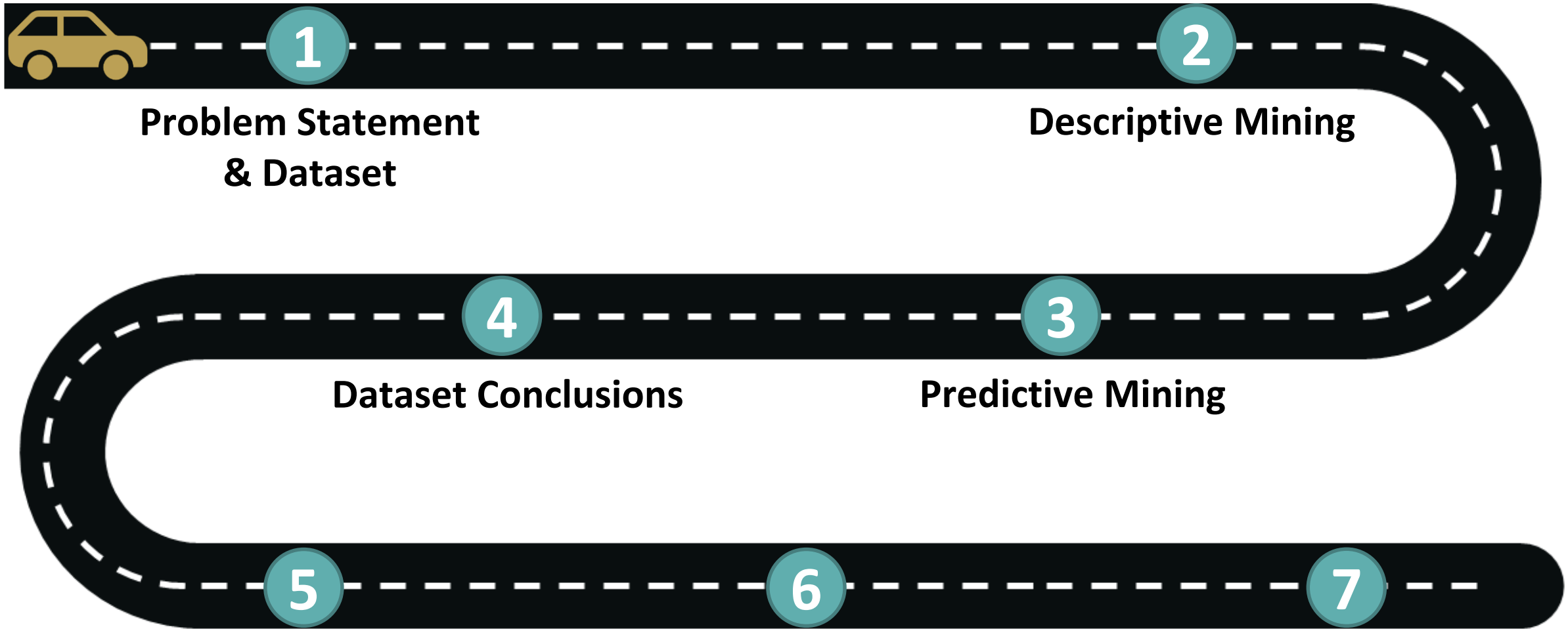
ALI

JYOTI

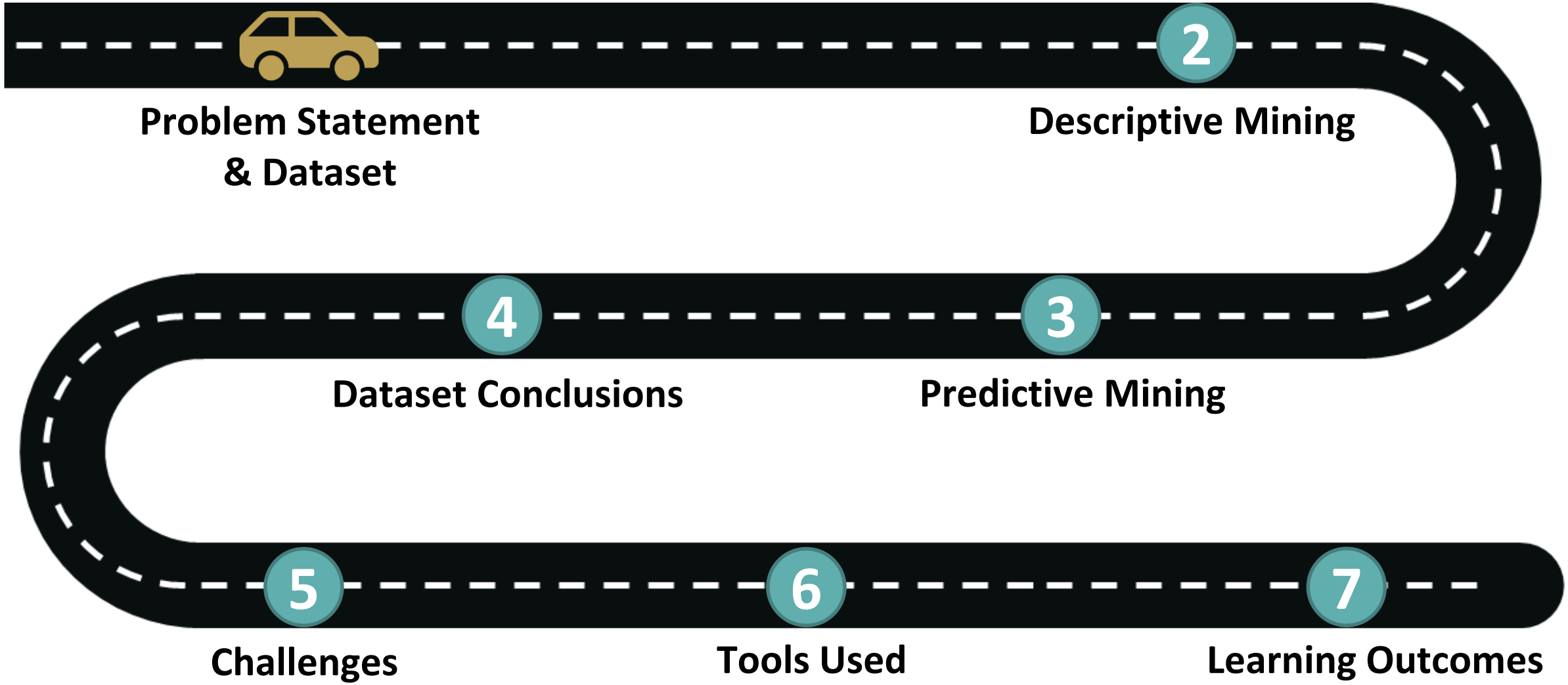
MADHAN

TAREK

Outline


1
Problem Statement & Dataset
2
Descriptive Mining
3
Predictive Mining
4
Dataset Conclusions
5
Challenges
6
Tools Used
7
Learning Outcomes

Problem Statement & Dataset



Problem Statement

- What factors lead to severe road accidents?
- How relevant should standardized crash tests be?
  - How do they relate to the real world?
- Generally pursue anything interesting relating to car safety.

UK Road Safety

- Consists of 3 tables:
  - Vehicles
  - Accidents
  - Casualties
- Attributes *Vehicle_Make* and *Vehicle_Model* retrieved from Kaggle
  - *Accident_Index* and *Vehicle_Reference* for mapping

Department for Transport

NHTSA

- 4 tables out of 6 were used:
  - Test data
  - Vehicle data
  - Barrier data
  - Occupant data
- Removed unrelated technical attributes
- Removed outliers
- API calls to get overall and crash specific star rating
  - Done for each test
  - Returned as JSON text.

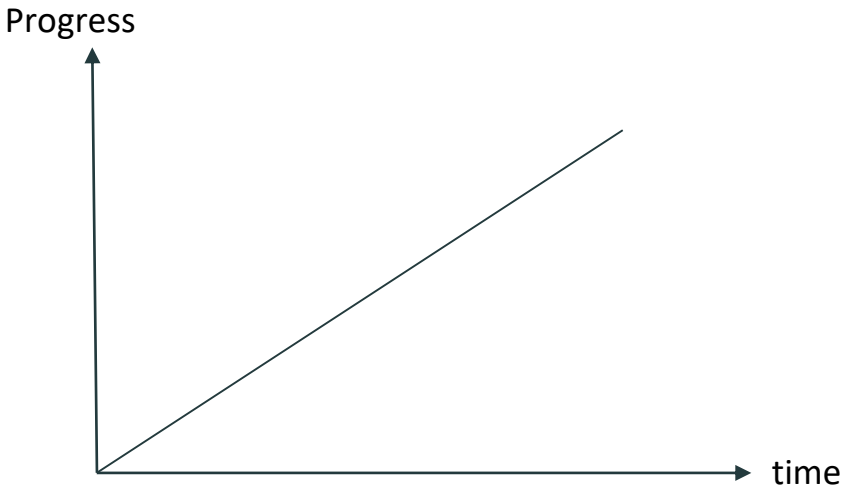
Cars in advertisements

- Premise was all about car safety quality.

https://www.youtube.com/embed/92sXWVxRr0g?start=0&end=77

Expectation

- Straightforward direct relationship between car ratings and average accident severeness.



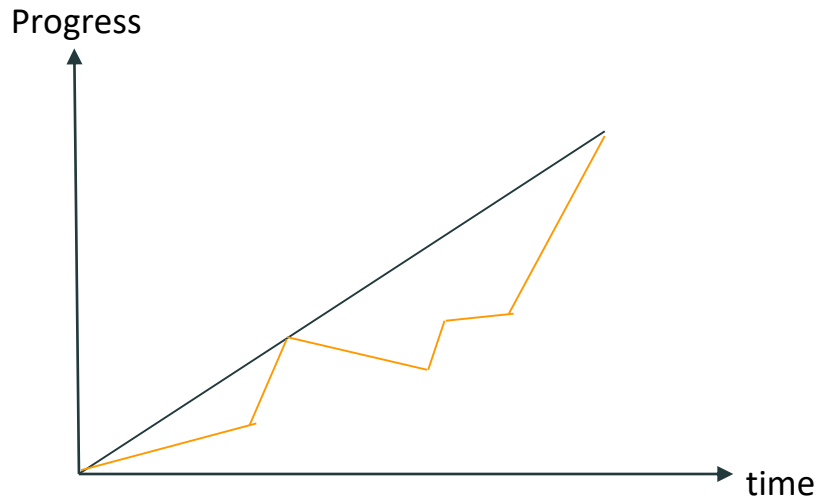
TECHNOLOGY NEWS DECEMBER 4, 2019 / 11:20 AM / 2 MONTHS AGO

Tesla Model X gets 5-star rating from European safety agency

1 MIN READ

Reality

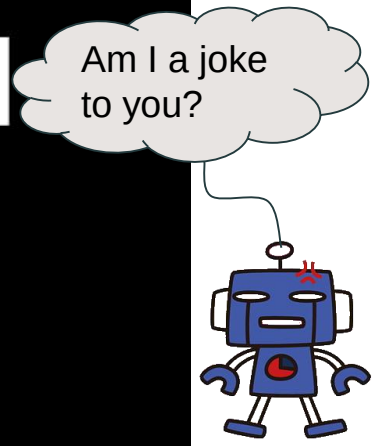
- Lots of subtle factors affecting outcome.



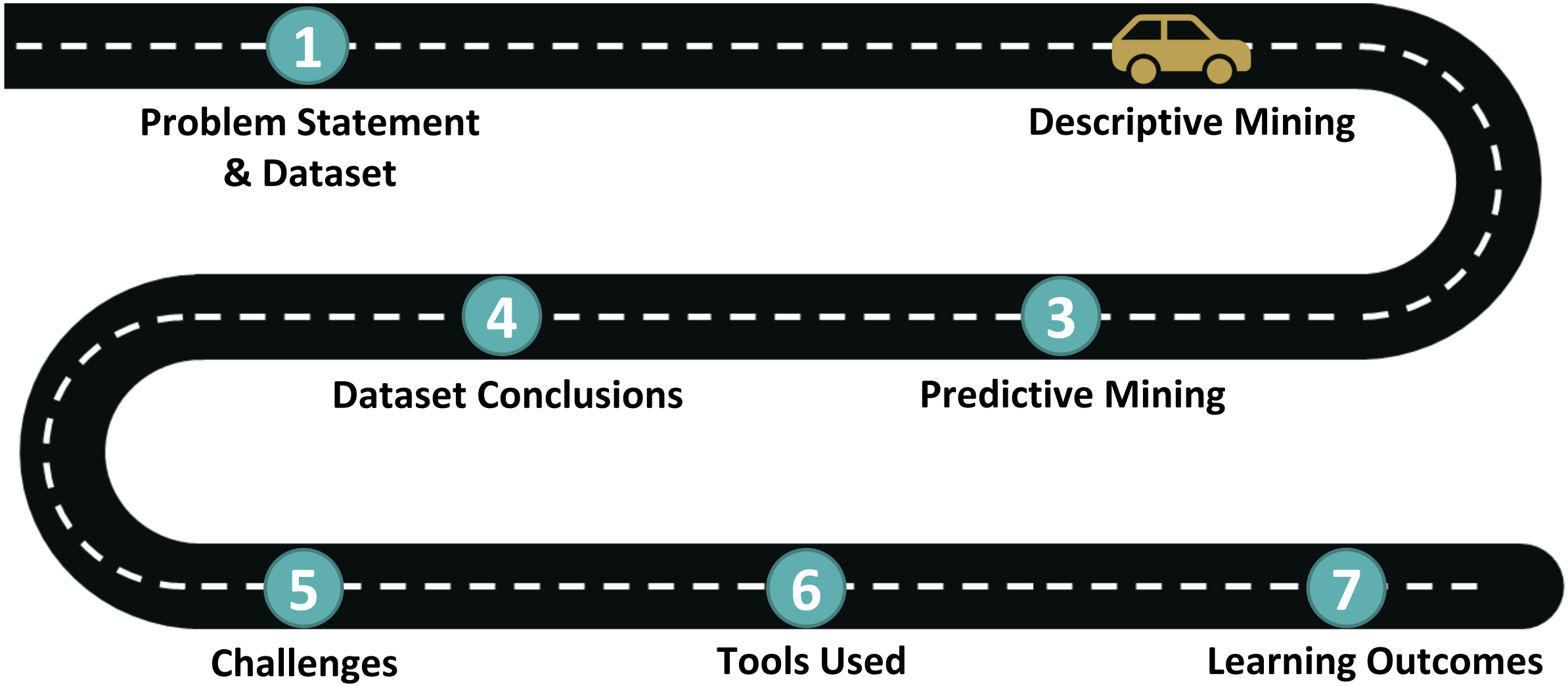
SELF-DRIVING TESLA 'KILLS' ROBOT IN LAS VEGAS CRASH, RAISING SUSPICIONS ABOUT RUSSIAN FIRM

Russian-made Promobot robot has previously met and shaken hands with Vladimir Putin

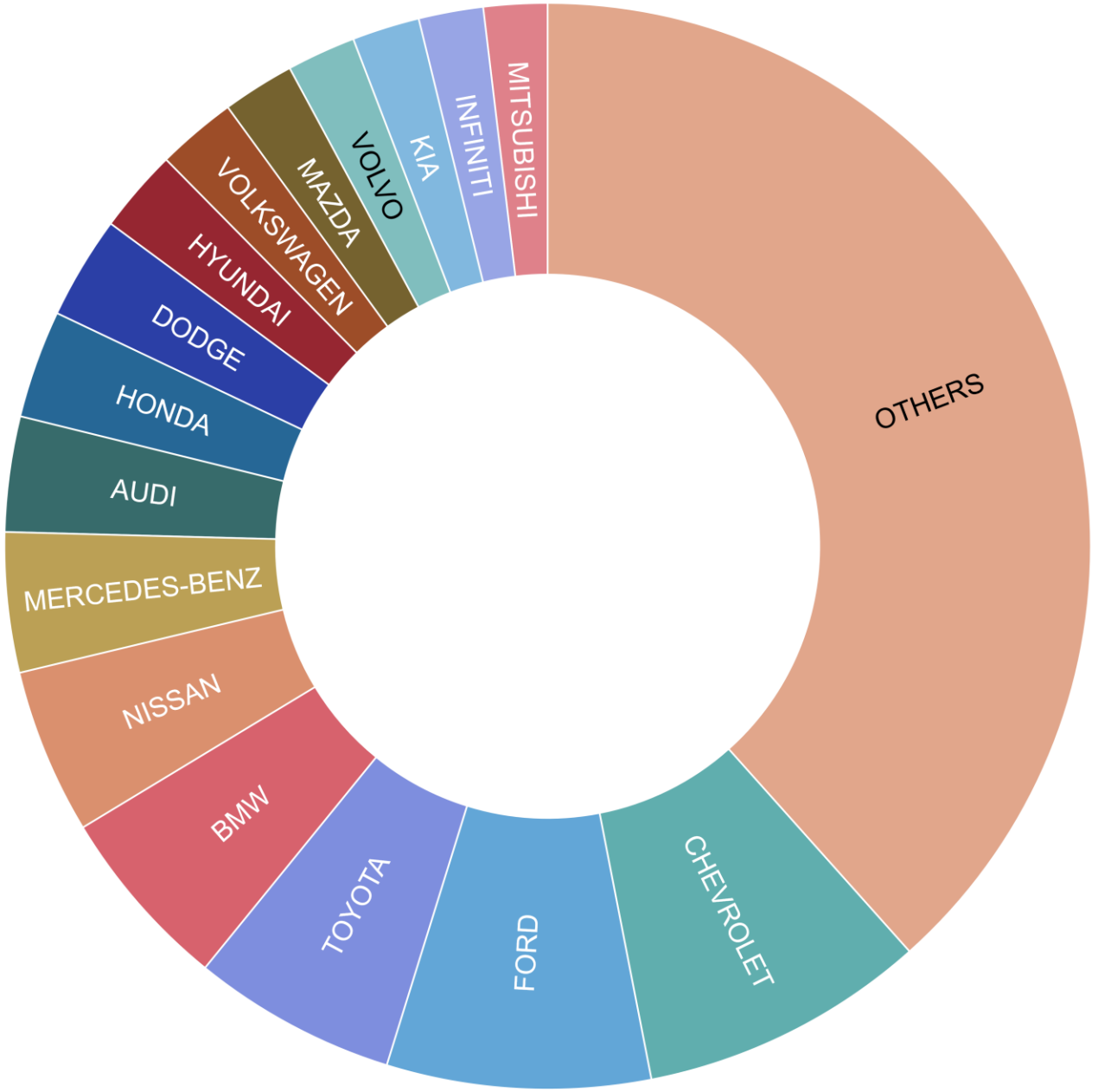
Anthony Cuthbertson | @ADCuthbertson |
Wednesday 9 January 2019 13:07 |



Descriptive Mining

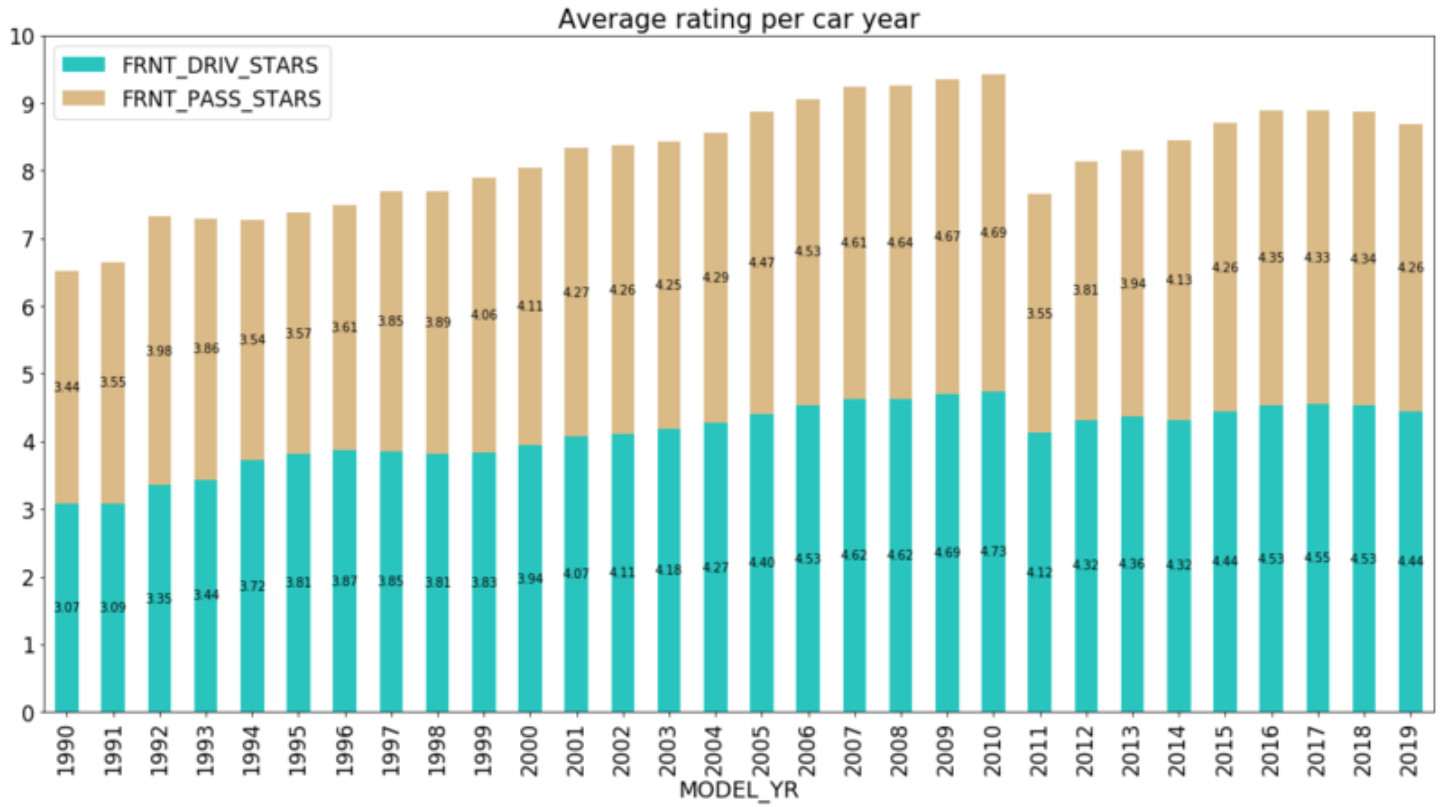


Top tested car models

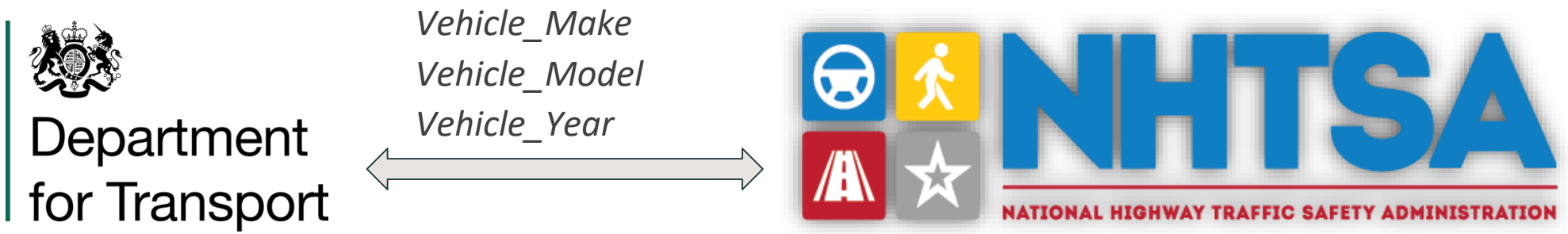


Analysis of car ratings

- NCAP ratings not on the same scale across the years
- Rating system revised in 2011
- Two solutions:
  - Chunk data to pre-2011 & post-2011
  - Normalize the car ratings



Linking UK Road Safety and NHTSA



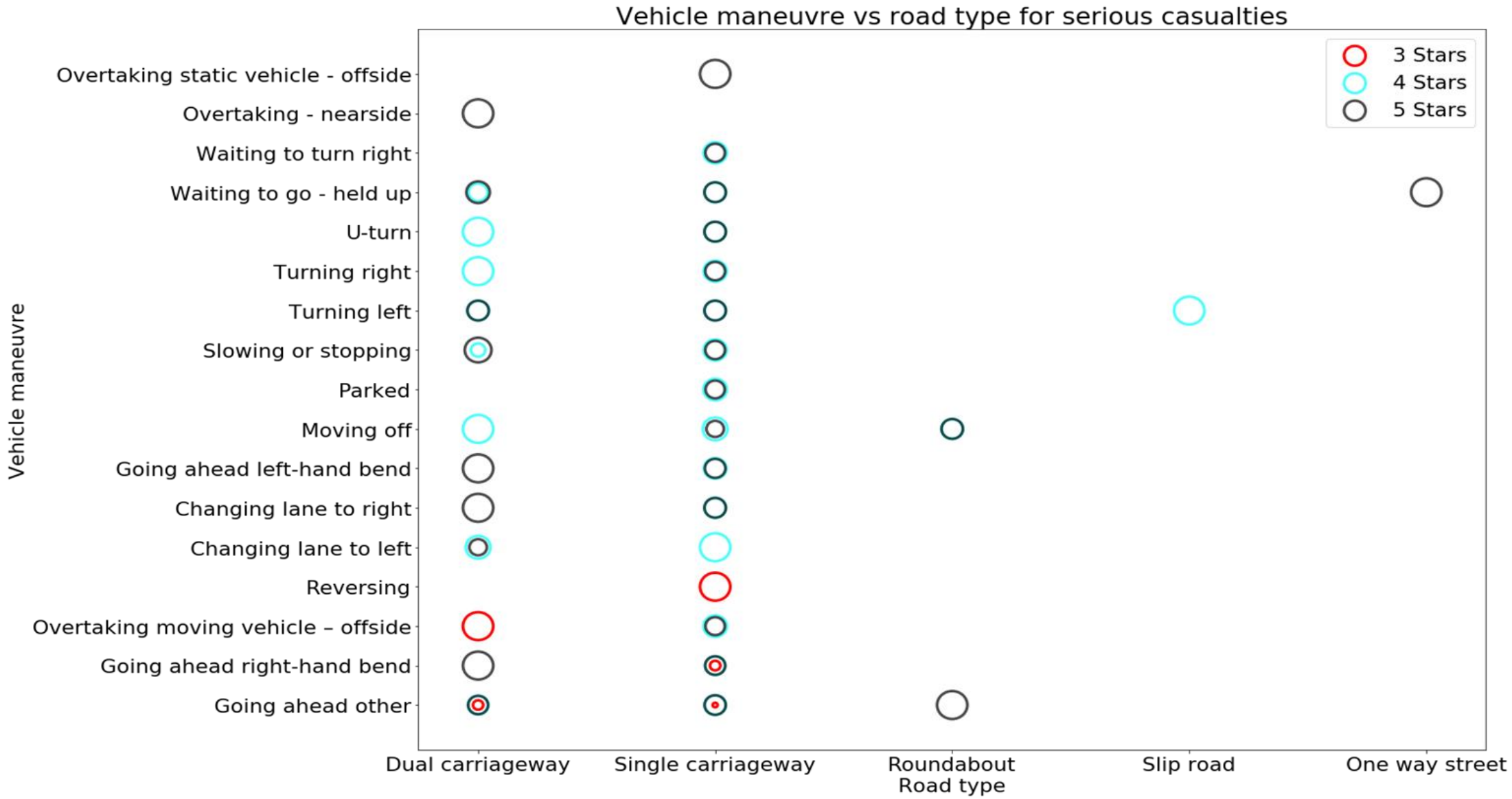
- Out of 108,668 distinct vehicle (car and van) model/year combinations in UK data, 15,308 (14%) unique model/year were matched with an NCAP rating.

Severe casualties vs NCAP ratings

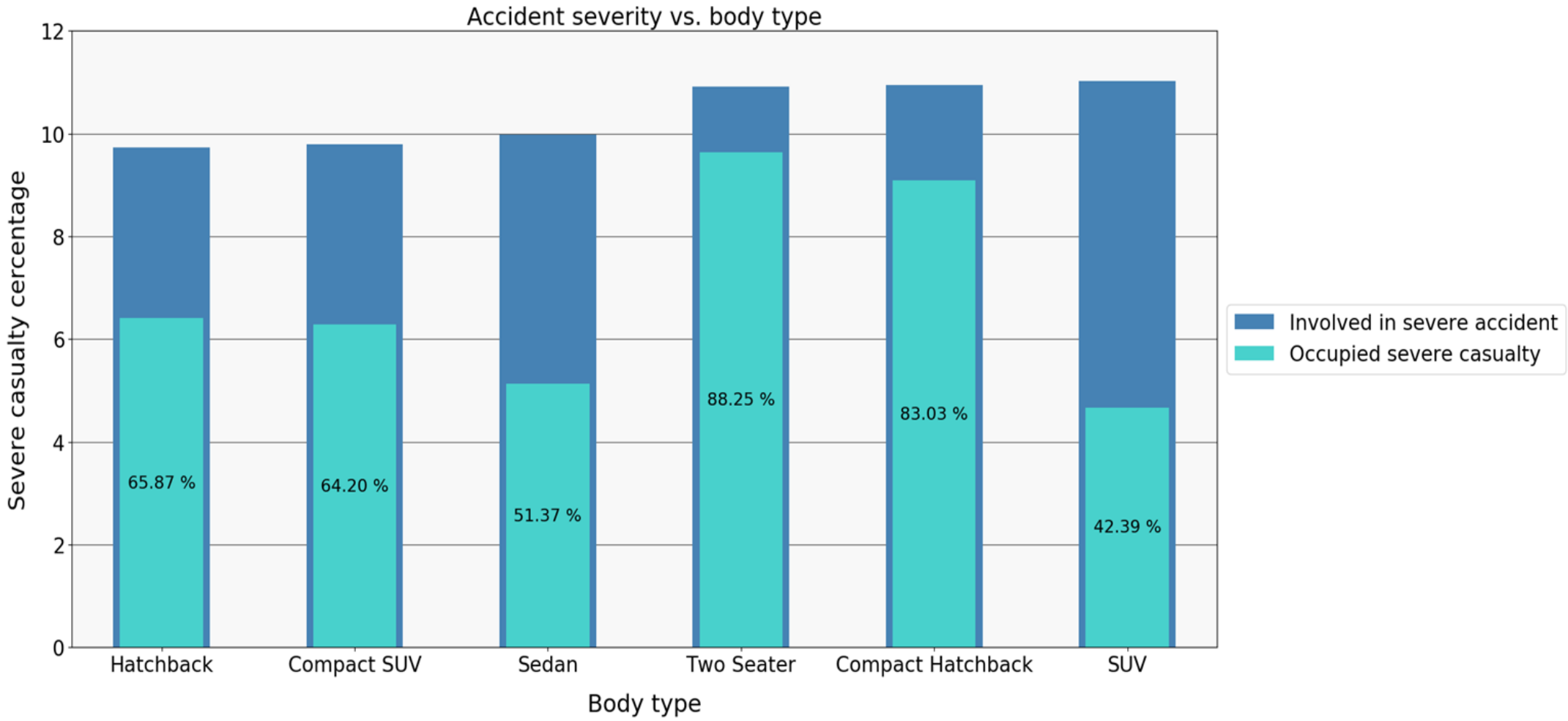
- No direct correlation between % of severe casualties and car ratings in the real world.
- Driver demographics and car type are the most influencing factors to safety in a vehicle.
- To determine the ‘correctness’ of the safety rating, accidents were grouped based on these factors.
- Significant negative correlations obtained between % of severe casualties and NCAP ratings upon grouping.

Maneuvre vs Severeness

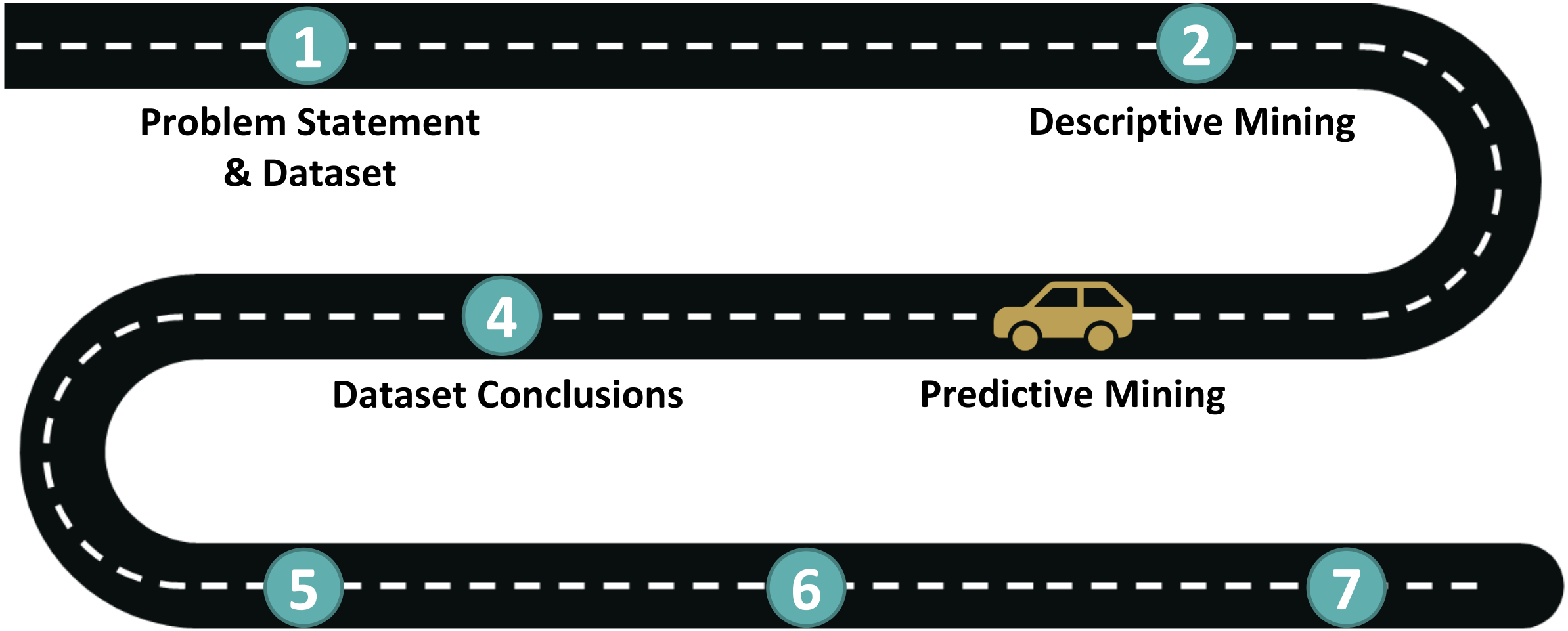
Rating distribution



Accident Severity vs Body type



Predictive Mining

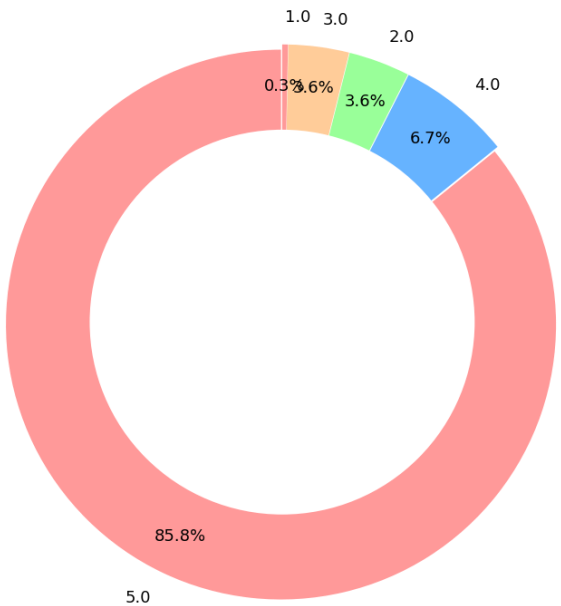


Predictive Mining

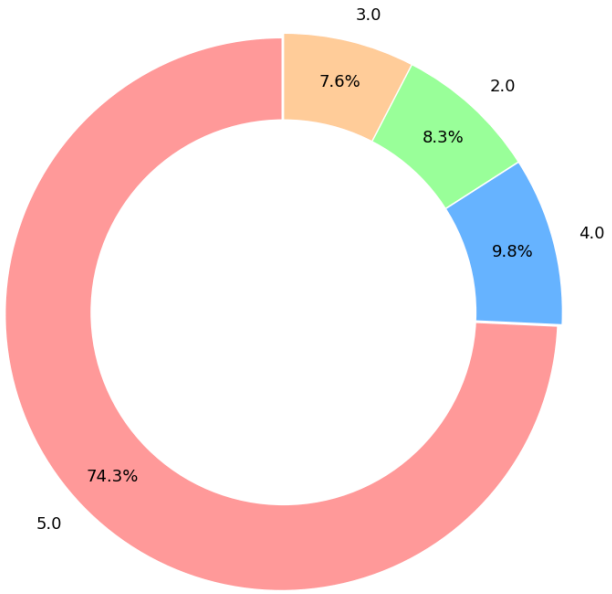
- Focus was on 3 distinct predictions:
  - Predicting NCAP star ratings from crash test data.
  - Predicting number of casualties for a given accident.
  - Predicting casualty/accident severities for given accidents.
- Multiple machine learning methods applied.

Predicting the NCAP Star Ratings

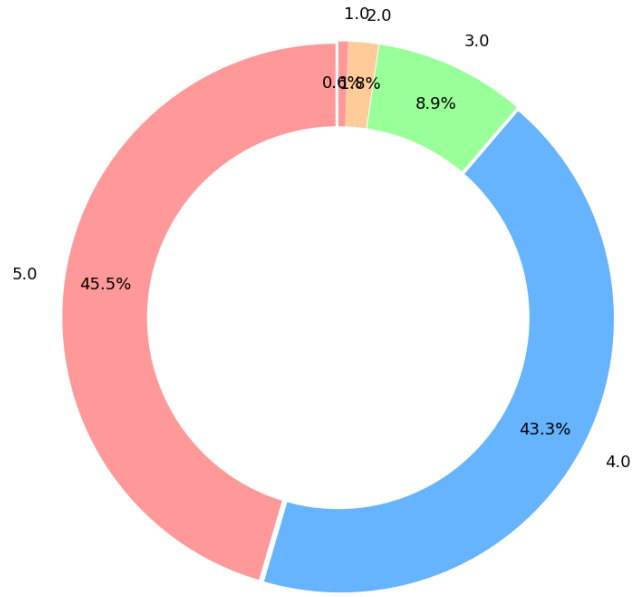
- On what basis is a star rating given?
  - Can we predict the missing star ratings from the NCAP tests?
- Multiple challenges:
  - NCAP rating system change – chunk the data or use the normalization method?
  - Is it a linear or non-linear relationship?
  - Heavy class imbalances:



Class distribution for SIDE_PASS_STARS. Total count: 2670



Class distribution for OVERALL_SIDE_STARS. Total count: 276



Class distribution for FRNT_PASS_STARS. Total count: 4655

Predicting the NCAP Star Ratings

Before 2011

	Overall stars	Overall front stars	Front driver stars	Front passenger stars	Overall side stars	Side driver stars	Side passenger stars	Rollover stars	Side pole stars
Log Regr	0.69	0.28	0.17	0.2	0.47	0.19	0.19	0.26	0.28
Decision Tree	0.73	0.9	0.34	0.38	1.0	0.33	0.3	0.51	0.72
Random Forest	0.73	0.96	0.38	0.46	1.0	0.42	0.38	0.62	0.72
SVM	0.73	0.96	0.34	0.46	1.0	0.4	0.34	0.53	0.72

After 2011

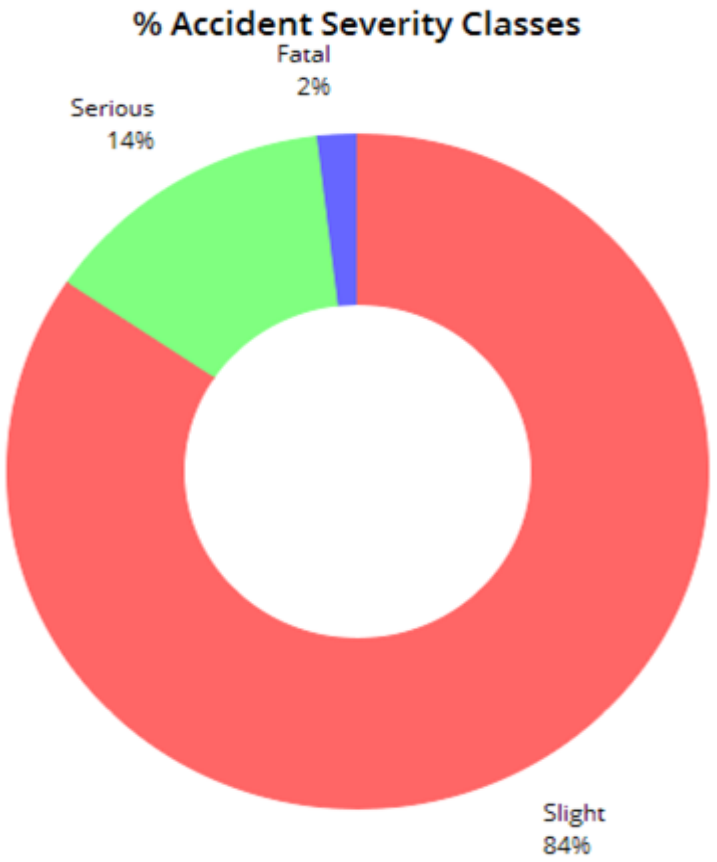
	Overall stars	Overall front stars	Front driver stars	Front passenger stars	Overall side stars	Side driver stars	Side passenger stars	Rollover stars	Side pole stars
Log Regr	0.25	0.22	0.17	0.24	0.24	0.3	0.31	0.28	0.17
Decision Tree	0.83	0.69	0.81	0.68	0.72	0.7	0.8	0.77	0.53
Random Forest	0.8	0.72	0.81	0.69	0.8	0.75	0.78	0.75	0.64
SVM	0.79	0.66	0.75	0.68	0.8	0.75	0.77	0.72	0.64

Normalized ratings

	Overall stars	Overall front stars	Front driver stars	Front passenger stars	Overall side stars	Side driver stars	Side passenger stars	Rollover stars	Side pole stars
Log Regr	0.17	0.14	0.14	0.16	0.15	0.16	0.27	0.2	0.15
Decision Tree	0.43	0.42	0.41	0.38	0.38	0.37	0.44	0.64	0.7
Random Forest	0.48	0.47	0.44	0.39	0.44	0.47	0.51	0.69	0.74
SVM	0.45	0.47	0.41	0.35	0.41	0.47	0.46	0.65	0.74

Prediction of Accident Severity

- Given the attributes of a particular accident:
  - Can we predict whether the it was slight, severe, or fatal?
- Multiple machine learning methods applied:
  - Skope rules
  - Decision trees
  - Random forests
- Sampling techniques:
  - Group fatal & serious accidents → under sample the slights



Prediction of Accident Severity

- Random forest performed the best:

	Precision	Recall	F1 score
Severe	0.70	0.71	0.71
Slight	0.71	0.70	0.70

- Most interesting findings in the feature importance.

Feature	Importance
Engine_Capacity_(CC)	0.164214
Age_of_Vehicle	0.125550
Day_of_Week	0.085836
Age_Band_of_Driver	0.077059
Vehicle_Manoeuvre	0.071134
1st_Point_of_Impact	0.062993
Number_of_Casualties	0.054869
Number_of_Vehicles	0.049742
1st_Road_Class	0.037799
Speed_limit	0.037605

Prediction of Casualty Severity

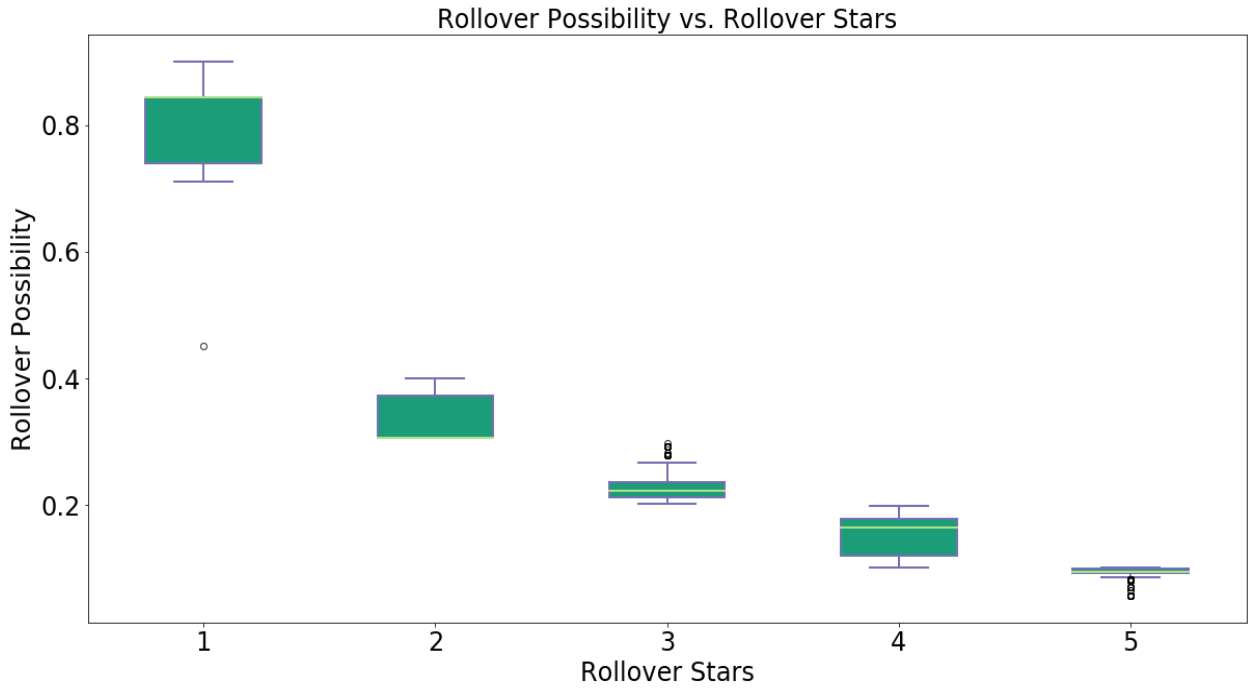
- Now able to include significant attributes:
  - Age of the casualty
  - Sex of the causality
  - Whether they where a pedestrian, car occupant, cyclist, etc.
- Two classes: severe casualty or slight casualty
- Model used: Logistic Regression
- After some additional preprocessing and using SMOTE sampling → 88% accuracy reached

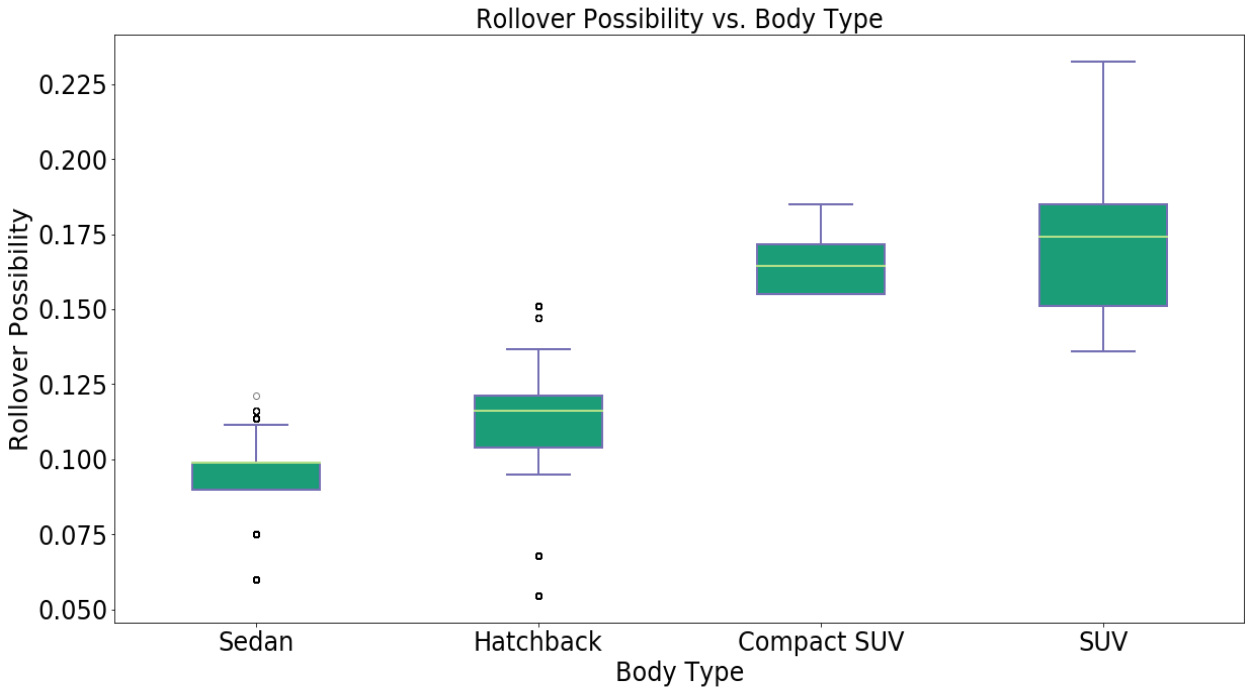
One Additional Predictive Task – Casualty Severity

- → Highest casualty prediction accuracy: 88%.
  - → Augment NCAP ratings with the accident data.
    - → Accuracy rose to 92%.

- → Most important feature: *Rollover Stars*

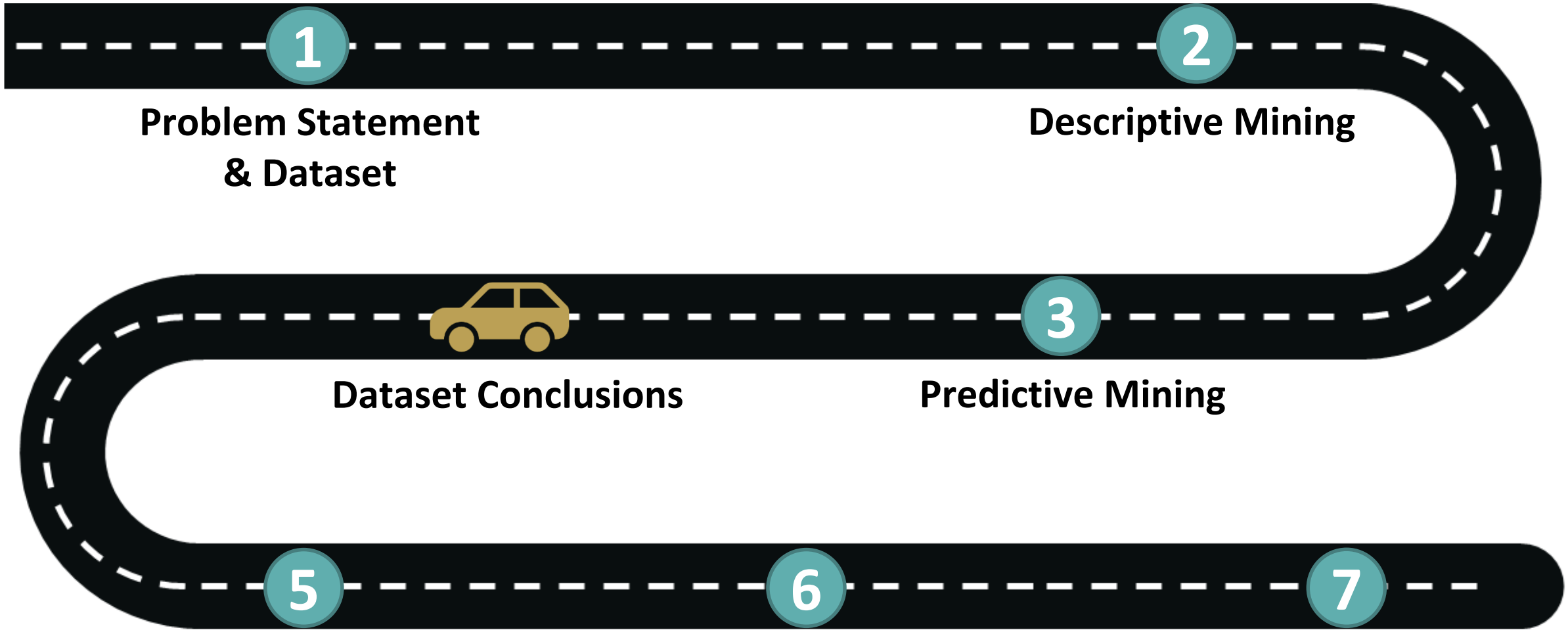
One Additional Predictive Task – Casualty Severity





→ *Rollover Stars ≅ Body Type*

Dataset Conclusions



Main Conclusions

- Most important accident severity predictors:

• Driver age • Driver sex • Performance car vs. family car	Driving style
• Speed limit • Road type	Collision speed
• Occupant car's body type • Opponent car's body type • Point of Impact	Collision type

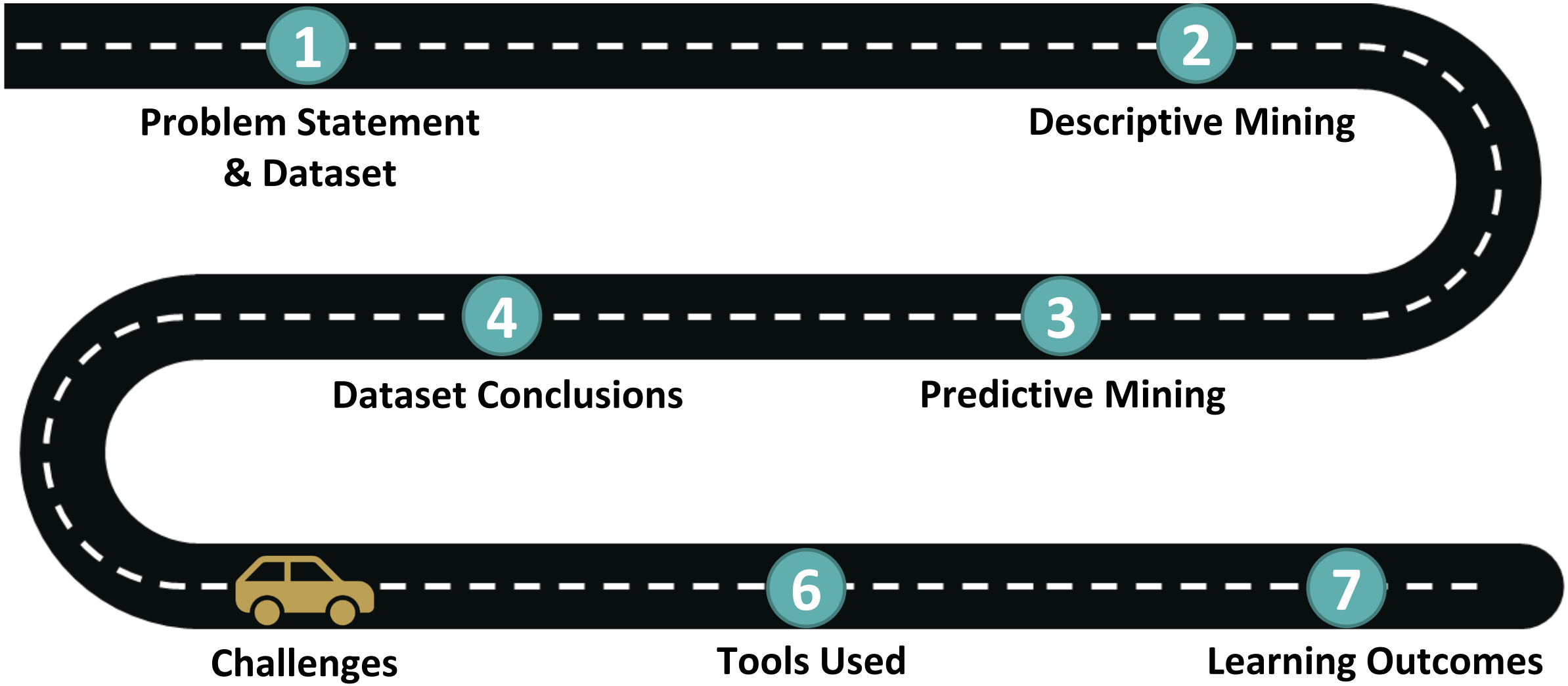
- Factor in: age of vehicle, casualty age, weather conditions, etc.

→ Standardized safety ratings are not so significant in real world collisions.

Main Conclusions

- Insurance companies are well aware of this.
- High safety rating ≠ immunity on the roads.
- Still, do not completely disregard cars safety rating.
  - But put a higher emphasis on your driving style and drive defensively.

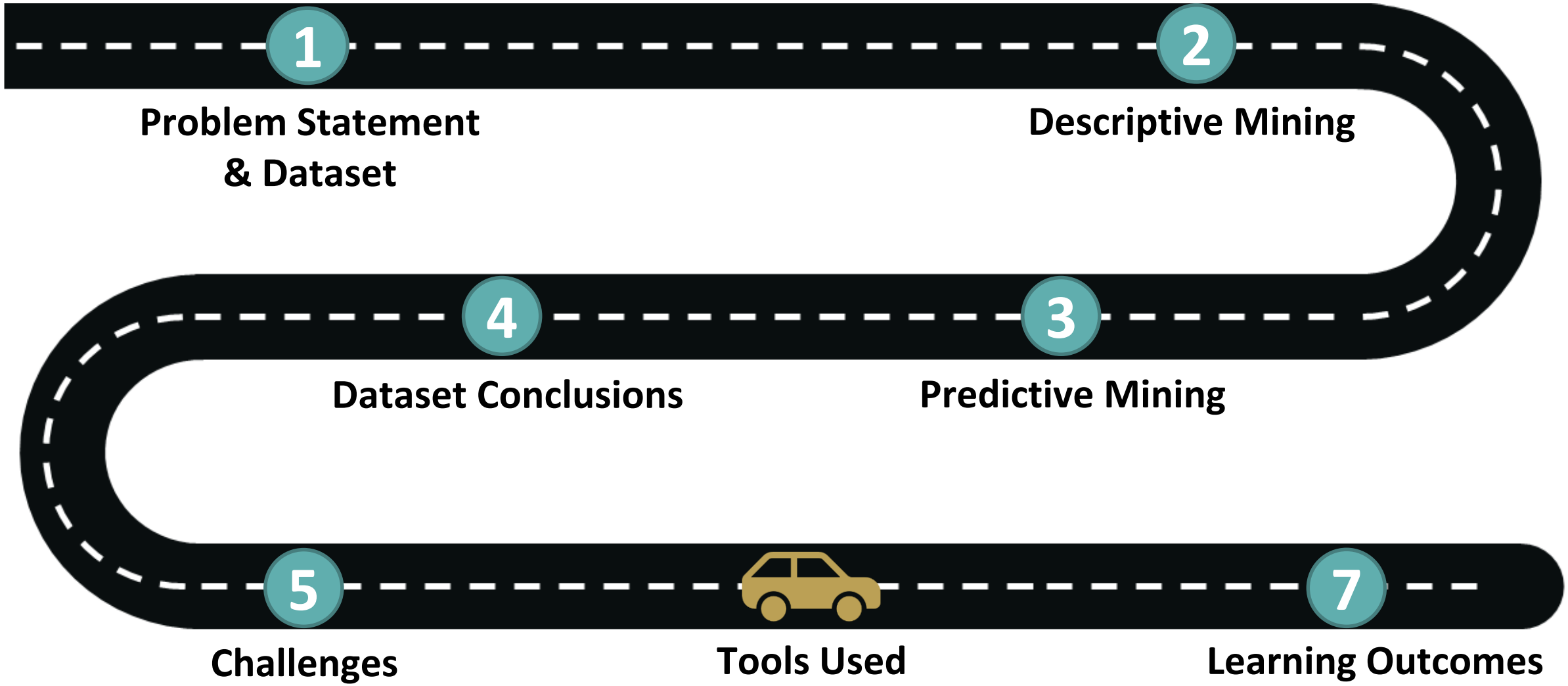
Challenges faced



Challenges faced

- Multiple Datasets
  - Data standardization specific to dataset
  - Revisited the merging of datasets
  - Too much data to discard
- Missing values, mostly categorical
- Failed attempts:
  - Initial hypothesis did not hold true
  - PCA, Clustering
  - Predicting number of casualties

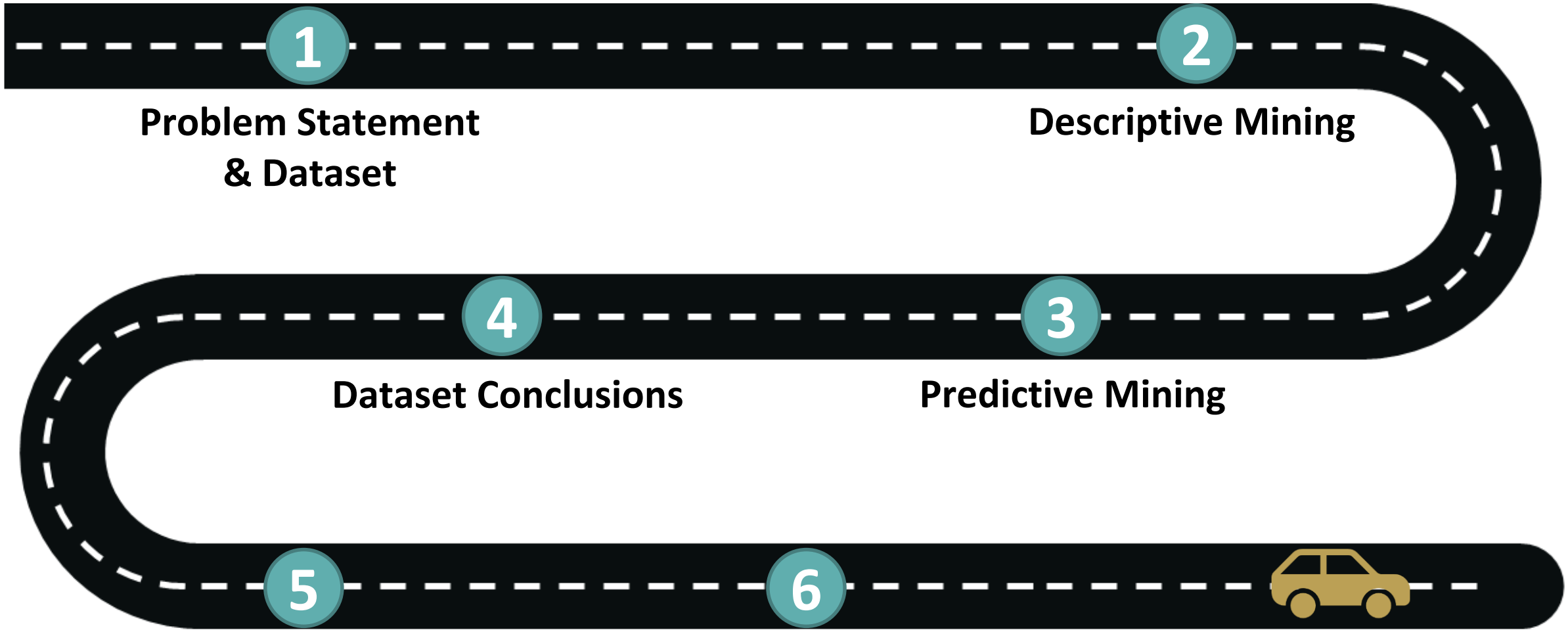
Tools Used


1
Problem Statement & Dataset
2
Descriptive Mining
3
Predictive Mining
4
Dataset Conclusions
5
Challenges
Tools Used
7
Learning Outcomes

Tools Used

- Python
- Jupyter Notebook
- Google Colab
- Matplotlib, Seaborn, Pyplot
- Scikit-learn
- GitLab
- Whatsapp
- Slack?

Learning Outcomes



Learning Outcomes

- Experience with real-world data from multiple sources
- Proper linkages across multiple data sources must be maintained
- Not just finding answers, but asking the right questions
- Not to be biased with an initial hypothesis
- Assessing the raw data and processing it for further descriptive and predictive phases
- Descriptive Mining plays an important role in:
  - Understanding the data
  - Discovering underlying patterns
  - Finding hidden correlations

Learning Outcomes (contd.)

- Selecting and applying proper visualization methods for reporting analytical findings
- Systematically apply supervised and unsupervised algorithms
- Not all algorithms work for every dataset (eg. clustering in our case)
- Points to bear in mind before Predictive Mining:
  - Class imbalance
  - ´Overall Accuracy´ not a true measure
- Accuracy vs. computational cost trade-off
- Accuracy vs. comprehensibility trade-off

Thank You