## Key Ideas

### Machine Learning

*Class-imbalance problem:*

* when one class outweighs another i.e. anomaly detection there will be many more false values than true values
* How to mitigate? -> (see <http://www.chioka.in/class-imbalance-problem/>)
  + Cost-based functions (i.e. false postive == 100 false negatives; thus the model tries to make fewer false positives)
  + Sampling based (i.e. over- or under-sample to emphasise the smaller class)
  + … or SMOTE (Synthetic Minority Over-Sampling Technique) combination of over and under-sampling
  + Thresholding -> i.e. to create a binary image

*Hand-crafted features:*

* Feature extraction/selection chosen by human rather than the model
* hand-crafted is the opposite of automatic feature extraction which is available in a CNN
* common in SVM, Naïve Bayes and Decision trees

*Thresholding:*

* used in image classification (thus CNNs)
* used to create a binary image (i.e. black and white) e.g. if above 0.5 then 1 (used in conjunction with oversampling in Zhang & Cheng (2019) to account for class-imbalance

*Pooling Layers:*

* used for dimensionality reduction and extraction of useful features

*Limitations of Machine Learning:*

* problem with SVM, Decision tree and Naïve Bayes is these such classifiers usually need hand-crafted features as input for training (Zhang & Cheng, 2019).

*STARIMA:*

* built on the assumption that you can combine

*Training:*

* ﻿Cross- validation is performed to select the optimal number of hidden nodes

*Anisotropy:*

* \*\* Directional dependent, different properties in different directions
* Isotrophy or anistrophy can be determined using a variogram
* *﻿Are there spatial pockets of out or in movement? (from Murray et al. 2012)*

## **Data Preperation**

- [On outlier removal]: Although transformation of data could be used to accommodate the anomaly, removing outliers from a data set is often one of the options when outliers lie far outside the range of the remaining data. Outlier removal can be achieved using histograms and descriptive stats (Cheng et al., 2011)

### Limitations with this research

*Data limitations:*

* Is it representative of the whole city of Montréal?

### Similar Research:

*Similar datasets:*

* NY Taxi-cab movements
* Uber-movements

*Similar Methods:*

* Maybe classify mode and where people are going? (after Bantis & Haworth, 2017)

*Data:*

* MTL2016:
  + 5452 subjects
  + GPS functionality is turned off and is only reactivated once the phone has moved 100 meters from the last GPS location (using the approximate location from Wi-Fi)
  + ﻿Respondents were prompted to “validate” their movements between locations by providing information (travel mode and trip purpose) about﻿ their latest trip when phones were detected to have stopped (the application ran 24 hours a day)
  + MTL ﻿Trajet provides block level spatial accuracy and is thus suitable for mapping subject locations to air pollution grids derived from existing exposure surfaces.
* MTL2017:

# Ideas:

* Overlay the Canadian MSOA with demographic with the linestrings (after Bantis & Haworth, 2017)
* Maybe have a summary stats section in data section discussing the data (after Eluru *et al.*, 2012)

# General Plan & Timetable

Word Limit: 12,000 words

Due: Friday 30th August

## Rough order of activities

## Readings

~100 papers (currently 17)

## Literature Review

~2000 words

Write up subheadings for these i.e. past studies, methodological, network data, spatio-temporal datasets and ML

## Data Pre-processing

Clean data

Find and Merge to a socio-demographic data source and POI or land use dataset

Write notes for the methodology

Create a note of limitations in the data and its preprocessing

## Data Analysis

Create an exploratory spatio-temporal data analysis

Try a few classification models

## Methodology Write-up

Write Data source and processing part

Create a note of limitations in the methods

Create subheadings

~1500-2000 words

## Results Write-up

~4000 words

Write up subheadings i.e. ESTDA, model comparisons and hyper parameter tuning

## Introduction Write-up

~1200 words

Motivation

## Discussion Write-up

~2000 words

Think of initial limitations (i.e. from data, methods)

See key ideas and think of more to write about

## Conclusion Write-up

~1000 words

Main ideas for each section

## Introduction:

## Literature Review:

## Methodology

## Results:

## Discussion:

## Conclusions:

## Structure of Thesis

*Starting Pages:*

Title Page

Abstract: 300 words max

Declaration

Table of Contents

List of Figures and Tables

List of acronyms and abbreviations

Acknowledgements

*Around one chapter (4450)*

Abstract: 200 words

1. Introduction: 1200 words

(Motivation: 300 words)

2. Literature Review: 2000 words

Statement of Ethics: 150 words

3. Methodology: 1500 words

*Several Chapters (6000 words)*

4. Results: 4000 words

5. Discussion: 2000 words

*One chapter (800)*

6. Further Work & Conclusions: 1000 words

*Other:*

References

Appendices: