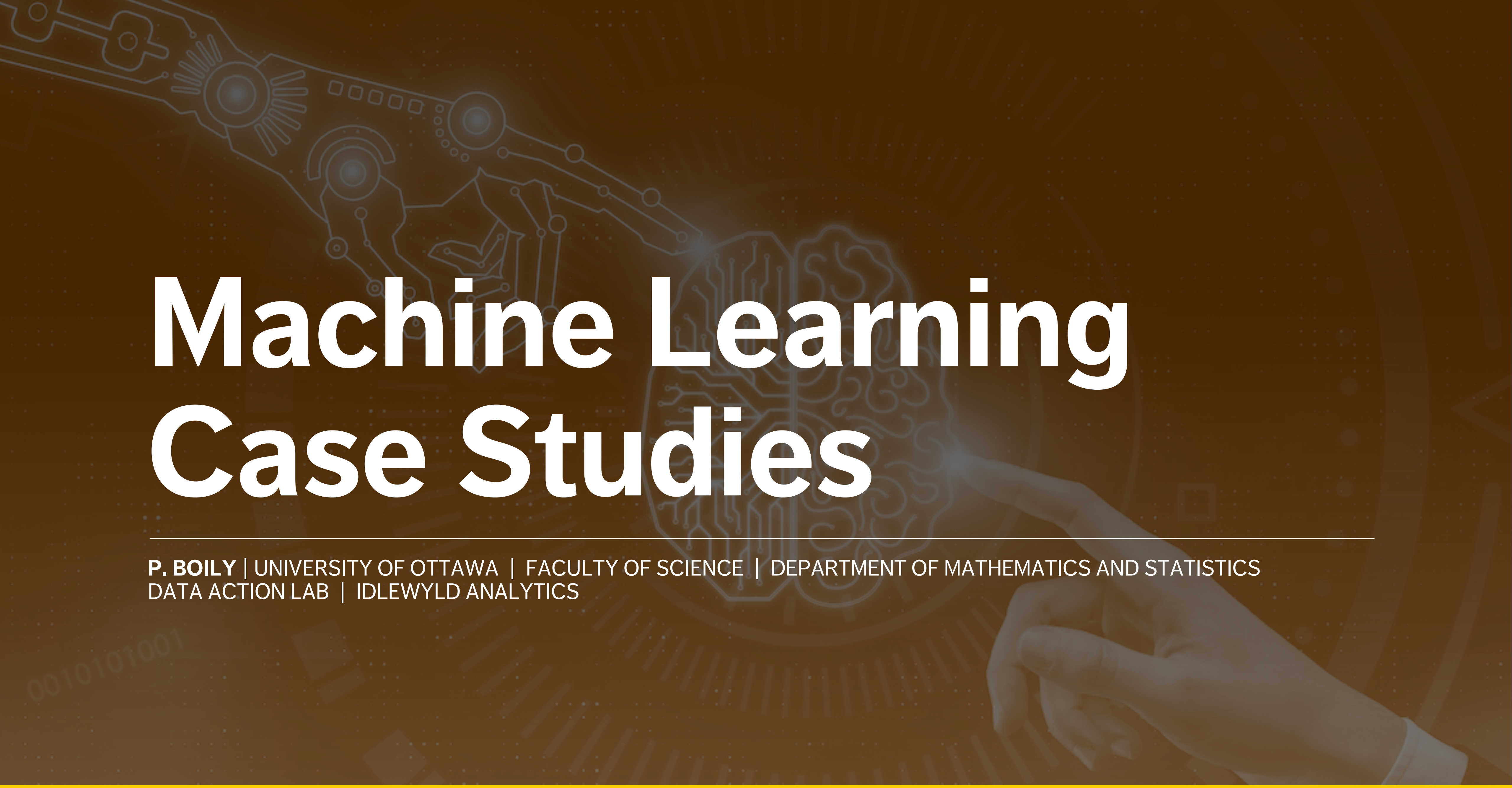


Machine Learning Case Studies

P. BOILY | UNIVERSITY OF OTTAWA | FACULTY OF SCIENCE | DEPARTMENT OF MATHEMATICS AND STATISTICS
DATA ACTION LAB | IDLEWYLD ANALYTICS

Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Objective

Using data from the *Danish National Patient Registry*, the authors sought connections between different **diagnoses:** how does a diagnosis at some point in time allow for the prediction of another diagnosis at a later time?

Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*
[Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients]
Nature Communications, vol. 5, 2014

Methodology

1. compute the **strength of correlation** for pairs of diagnoses over a 5 year interval (on a representative subset of the data)
2. test diagnoses pairs for **directionality** (one diagnosis repeatedly occurring before the other)
3. determine reasonable **diagnosis trajectories** (thoroughfares) by combining smaller (but frequent) trajectories with overlapping diagnoses
4. **validate** the trajectories by comparison with non-Danish data
5. **cluster** the thoroughfares to identify a small number of **central medical conditions** (key diagnoses) around which disease progression is organized

Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Data

The *Danish National Patient Registry* is an electronic health registry containing administrative information and diagnoses, covering the whole population of Denmark, including private and public hospital visits of all types:

- inpatient (overnight stay)
- outpatient (no overnight stay)
- emergency visits.

The data set covers 15 years of such visits, from January ’96 to November ’10, and consists of 68 million records for 6.2 million patients.

Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Challenges and Pitfalls

- Access to the **patient registry** was protected and could only be granted after approval by *the National Board of Health*.
- There are gender-specific differences in diagnostic trends, but many diagnoses were made predominantly in different sites, suggesting the stratifying by **site** as well as by **gender**.
- In the process of forming small diagnoses chains, they had to compute the correlations using **large groups** for each pair of diagnoses (1 million diagnosis pairs = 80+ million samples) to compensate for **multiple testing** (1000s years' worth of CPU run time) – pre-filtering steps were used to avoid this pitfall.

Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

The dataset was reduced to **1,171 significant trajectories**.

These thoroughfares were clustered into patterns centred on 5 key diagnoses for disease progression:

- **diabetes**
- **chronic obstructive pulmonary disease** (COPD)
- **cancer**
- **arthritis**
- **cerebrovascular disease**

Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*
Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients
Nature Communications, vol. 5, 2014

Project Summary and Results

Early diagnoses for these central factors can help reduce the risk of adverse outcome linked to future diagnoses of other conditions.

Among the specific results, the following “surprising” insights were found:

- a diagnosis of anemia is typically followed months later by the **discovery of colon cancer**
- a diagnosis of gout was identified as **a step on the path** toward eventual diagnosis of cardiovascular diseases
- COPD is **under-diagnosed** and **under-treated**

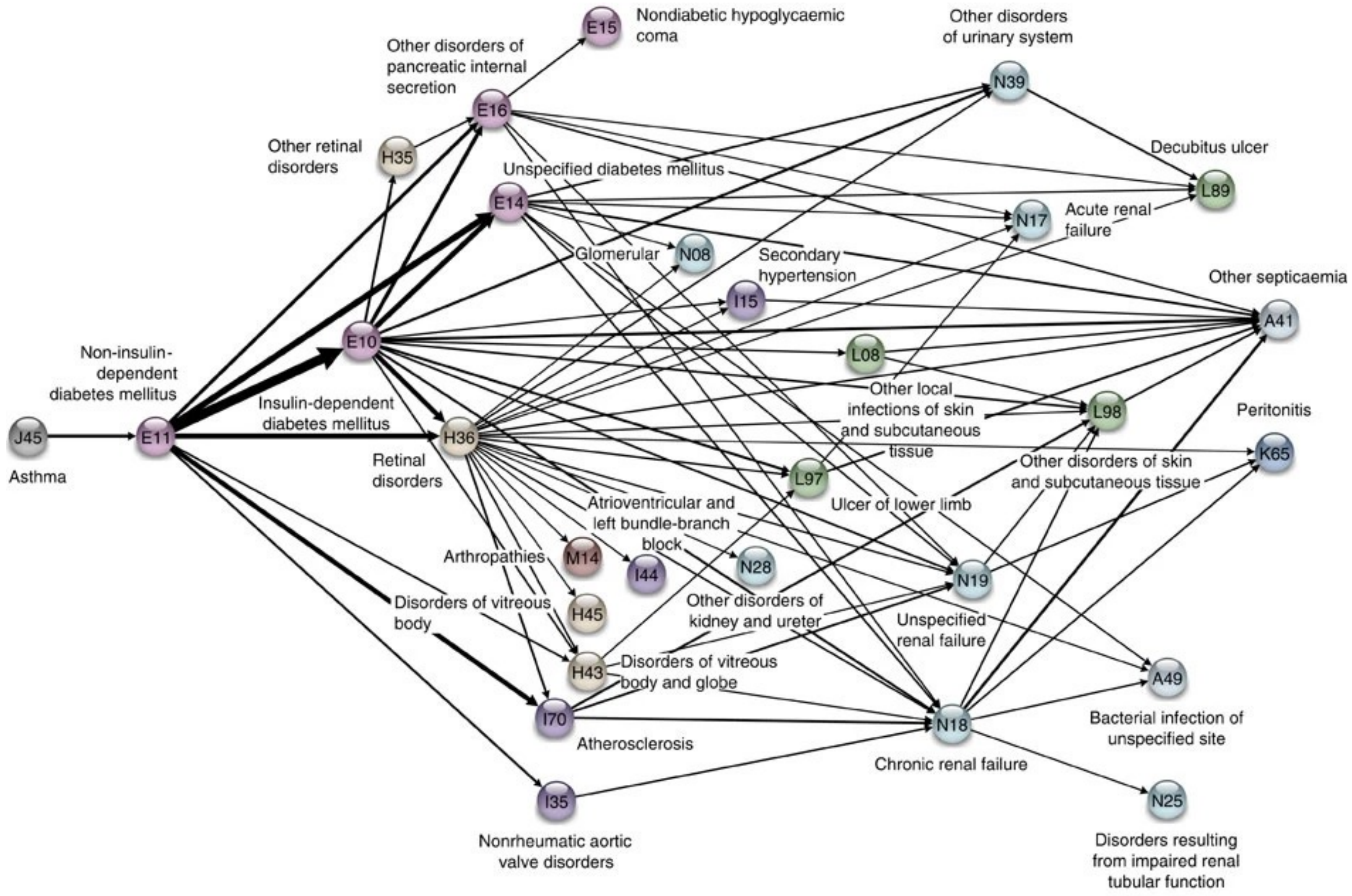
Association Rules Mining Case Study

Danish Medical Data

Jensen *et al.*

[Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients](#)

Nature Communications, vol. 5, 2014



E15 Nondiabetic hypoglycaemic coma
Other disorders of urinary system
N39
Other disorders of pancreatic internal secretion
E16
Other retinal disorders
H35
Unspecified diabetes mellitus
Decubitus ulcer
L89
E14
N17
Acute renal failure
Glomerular
N08
Secondary hypertension
Other septicaemia
I15
A41
E10
L08
Non-insulin-dependent diabetes mellitus
Other local infections of skin and subcutaneous tissue
Insulin-dependent diabetes mellitus
L98
Peritonitis
J45
E11
H36
K65
Asthma
Retinal disorders
L97
Other disorders of skin and subcutaneous tissue
Ulcer of lower limb
Atrioventricular and left bundle-branch block
Arthropathies
M14
I44
N28
N19
Disorders of vitreous body
H45
Other disorders of kidney and ureter
Unspecified renal failure
H43
Disorders of vitreous body and globe
A49
I70
N18
Bacterial infection of unspecified site
Atherosclerosis
Chronic renal failure
I35
N25
Nonrheumatic aortic valve disorders
Disorders resulting from impaired renal tubular function

Classification Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Objective

The U.S. Internal Revenue Service (IRS) estimated that there were large gaps between **revenue owed** and **revenue collected** for 2001 and for 2006.

Using DoR data, the authors sought to increase **efficiency** in the audit selection process and to **reduce the gap** between revenue owed and revenue collected.

Classification Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue](#)
Real World Data Mining Applications, 2015

Methodology

1. **data selection and separation:** experts selected several hundred cases to audit and divided them into training, testing and validating sets
2. **classification modeling** using MultiBoosting, Naïve Bayes, C4.5 decision trees, multilayer perceptrons, support vector machines, etc.
3. **evaluation of all models** on the testing set – models performed poorly until the size of the business being audited was recognized to have an effect, leading to two separate tasks (large/small businesses).
4. **model selection/validation** compared the estimated accuracy between different classification model predictions and the actual field audits (MultiBoosting with Naïve Bayes was selected as the final model; suggesting improvements to increase audit efficiency).

Classification Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Data

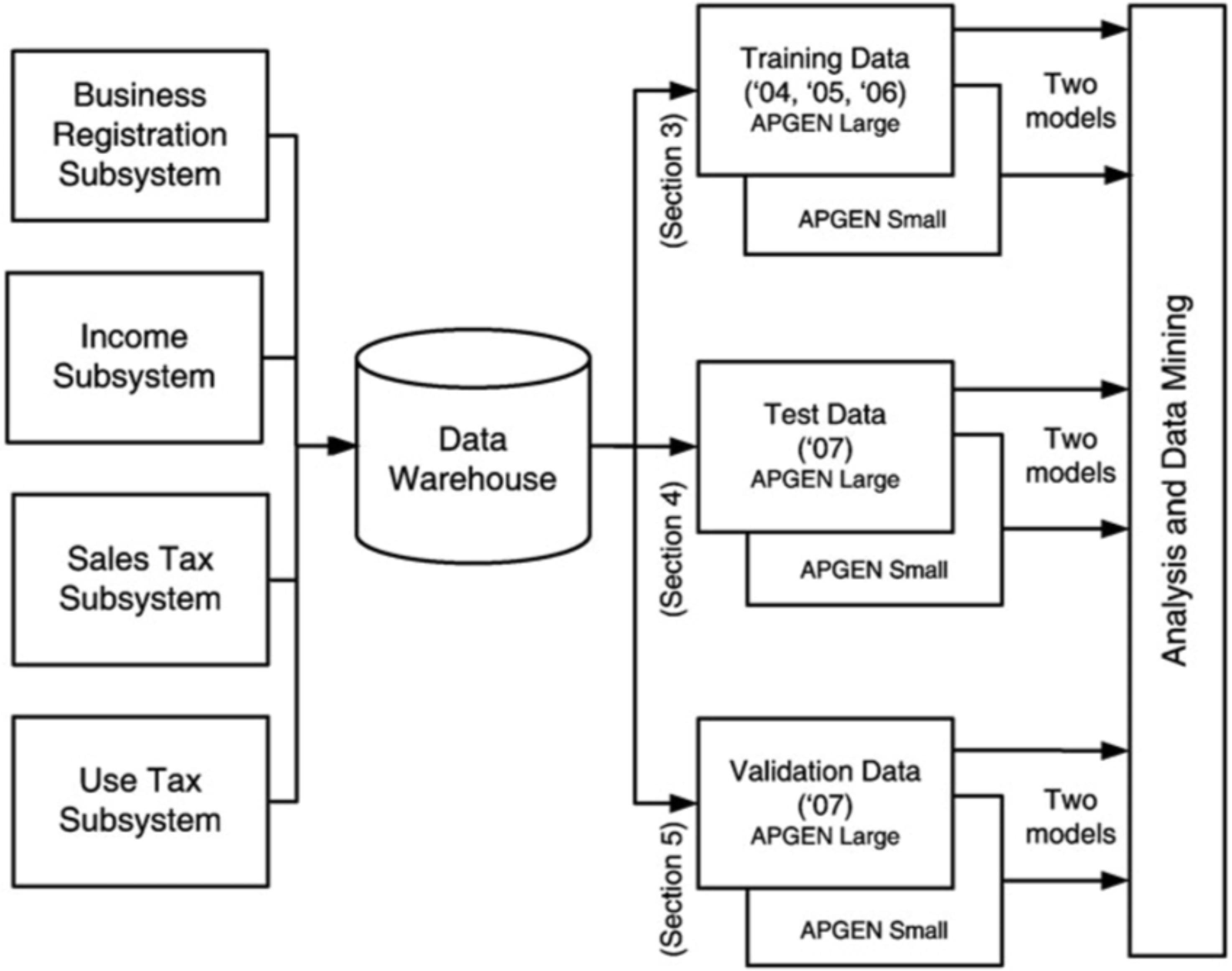
Selected tax audit cases from 2004 to 2007, collected by the audit experts, which were split into training, testing and validation sets:

- the **training data** set consisted of *Audit Plan General* (APGEN) *Use Tax* audits and their results for the years 2004-2006
- the **testing data** consisted of APGEN Use Tax audits conducted in 2007 and was used to test or evaluate models (for Large and Smaller businesses) built on the training dataset
- while **validation** was assessed by actually conducting field audits on predictions made by models built on 2007 Use Tax return data processed in 2008.

Classification Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue](#)
Real World Data Mining Applications, 2015



Classification Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

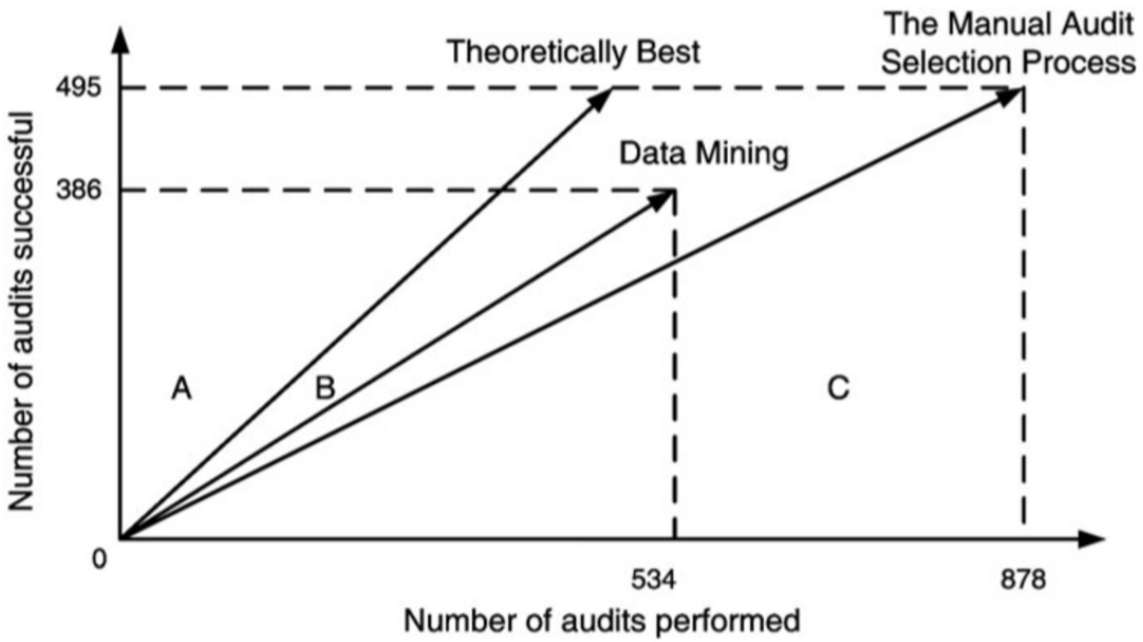
Strengths and Limitations of the Algorithms

- Naïve Bayes classification assumes independence of the features, which rarely occurs in real-world situations. This approach is also known to potentially introduce bias to classification schemes. In spite of this, classification models built using it have a successfully track record.
- MultiBoosting is an **ensemble technique** that uses committee (i.e. groups of classification models) and "group wisdom" to make predictions; unlike other ensemble techniques, it is different from other ensemble techniques in the sense that it forms a committee of sub-committees, which has a tendency to reduce both bias and variance of predictions.

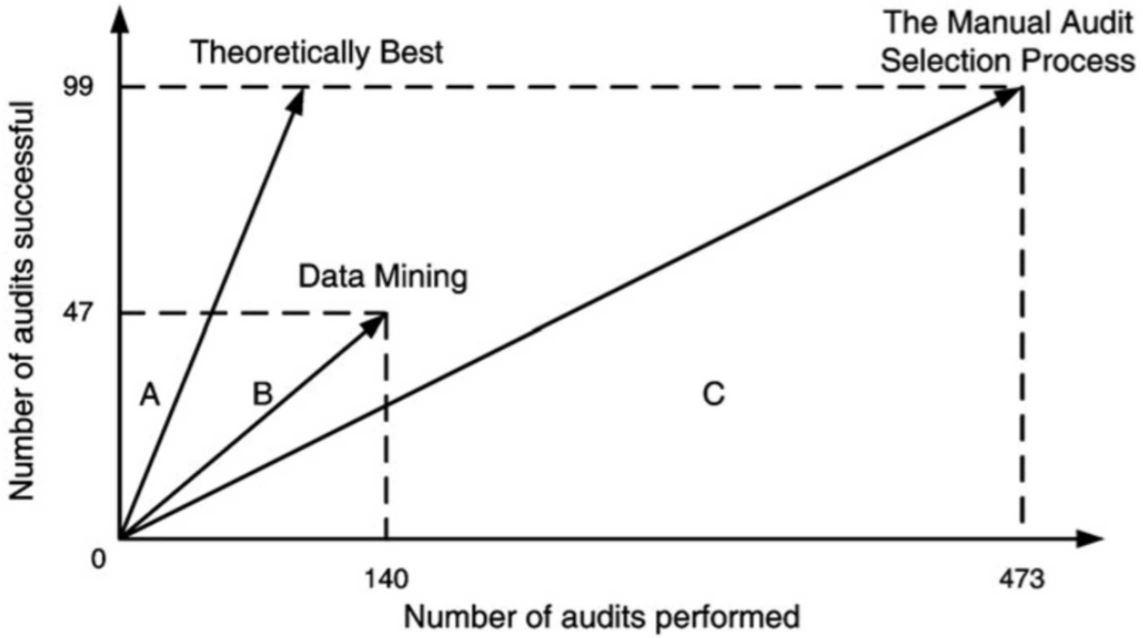
Classification Case Study

Minnesota Tax Audits

APGEN Large



APGEN Small

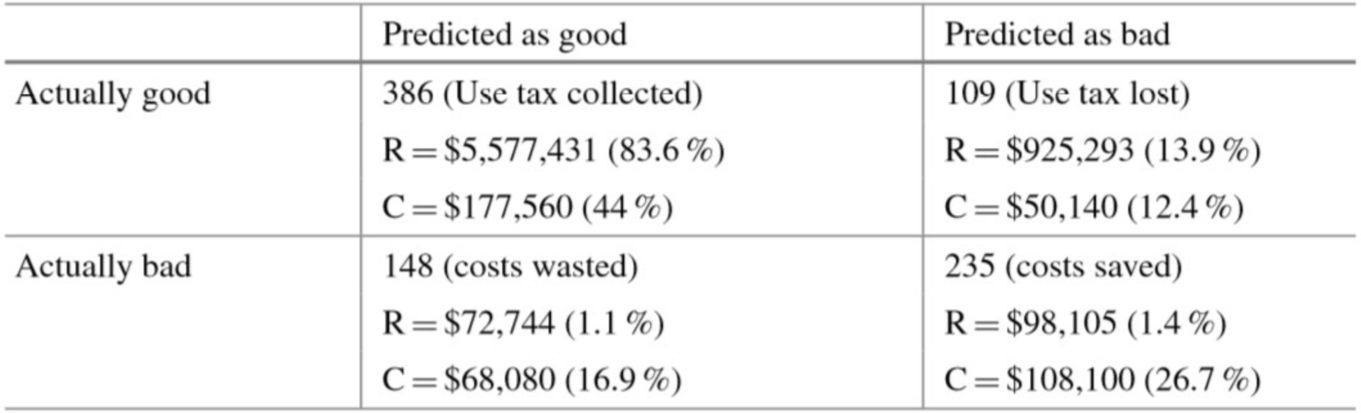


Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue](#)
Real World Data Mining Applications, 2015

Classification Case Study

Minnesota Tax Audits

APGEN Large

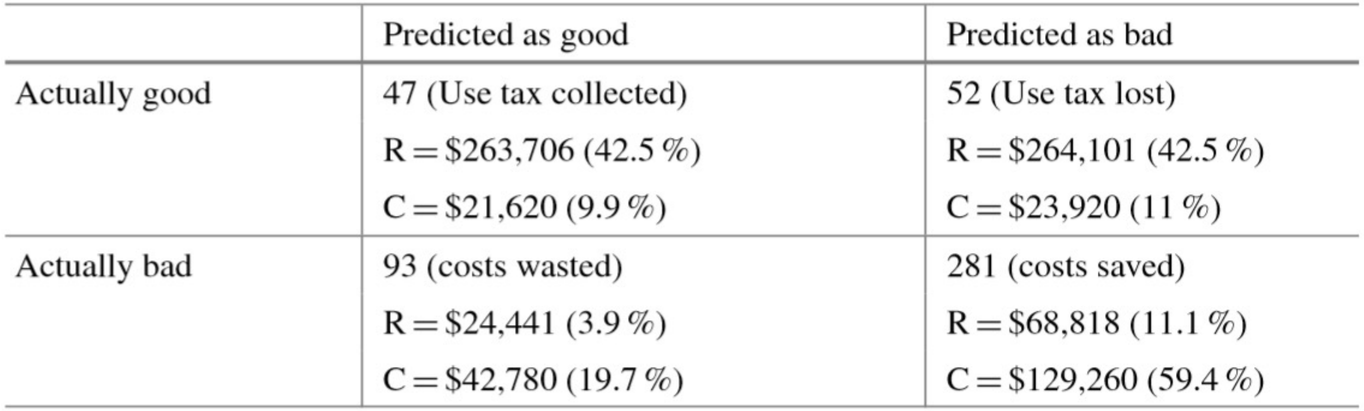


	Predicted as good	Predicted as bad

Actually good	386 (Use tax collected)
	R = $5,577,431 (83.6 %)
	C = $177,560 (44 %)
		109 (Use tax lost)
		R = $925,293 (13.9 %)
		C = $50,140 (12.4 %)

Actually bad	148 (costs wasted)
	R = $72,744 (1.1 %)
	C = $68,080 (16.9 %)
		235 (costs saved)
		R = $98,105 (1.4 %)
		C = $108,100 (26.7 %)

APGEN Small



	Predicted as good	Predicted as bad

Actually good	47 (Use tax collected)
	R = $263,706 (42.5 %)
	C = $21,620 (9.9 %)
		52 (Use tax lost)
		R = $264,101 (42.5 %)
		C = $23,920 (11 %)

Actually bad	93 (costs wasted)
	R = $24,441 (3.9 %)
	C = $42,780 (19.7 %)
		281 (costs saved)
		R = $68,818 (11.1 %)
		C = $129,260 (59.4 %)

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Classification Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Take-Aways

- Many models were churned out before the team made a final selection.
- Past performance of a model family in a previous project can guide the selection, but remember the *No Free Lunch (NFL) Theorem*: nothing works best all the time!
- The feature selection process could very well require a number of visits to domain experts before the feature set yields promising results.
- Data analysis teams should seek out individuals with a good understand of both data and context.
- Domain-specific knowledge has to be integrated in the model in order to beat random classifiers, on average.
- Even slight improvements over the current approach can find a useful place in an organization – data science is not solely about Big Data and disruption!

Clustering Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Objective

When we think of similarity at the urban level, we typically think in terms of neighbourhoods. Is there some other way to identify similar parts of a city?

The researchers aims to draw the boundaries of **livehoods**, areas of similar character within a city, by using clustering models. Unlike **static** administrative neighborhoods, the livehoods are defined based on the **habits** of their inhabitants.

Clustering Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Methodology

The authors use **spectral clustering** to discover **distinct geographic areas** of the city based on collective **movement patterns**.

Livehood clusters are built as follows:

1. a **geographic distance** is computed based on pairs of check-in venues' coordinates;
2. a **social similarity** is computed between each pair of **venues** using cosine measurements;
3. spectral clustering produces **candidate livehoods**;
4. interviews are conducted with residents in order to **explore**, **label**, and **validate** the clusters discovered by the algorithm.

Clustering Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Data

The data comes from two sources, combining approximately 11 million check-ins from the dataset of Chen et al. (a recommendation site for venues based on users' experiences) and a new dataset of 7 million Twitter check-ins downloaded between June and December of 2011.

For each check-in, the data consists of the **user ID**, the **time**, the **latitude and longitude**, the **name of the venue**, and its **category**.

In this case study, data from the city of Pittsburgh, Pennsylvania, is examined *via* 42,787 check-ins of 3840 users at 5349 venues.

Clustering Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Strengths and Limitations of the Approach

- The technique used in this study is **agnostic** towards the particular source of the data: it is not dependent on meta-knowledge about the data.
- The algorithm may be prone to "majority" bias, possibly misrepresenting/hiding minority behaviours.
- The dataset is built from a **limited** sample of check-ins shared on Twitter and are therefore biased towards the types of visits/locations that people typically want to share **publicly**.
- Tuning the clusters is non-trivial: experimenter bias may combine with "confirmation bias" of the interviewees in the validation stage – if the researchers are residents of Pittsburgh, will they see clusters when there were none?

Clustering Case Study

Livehoods

Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

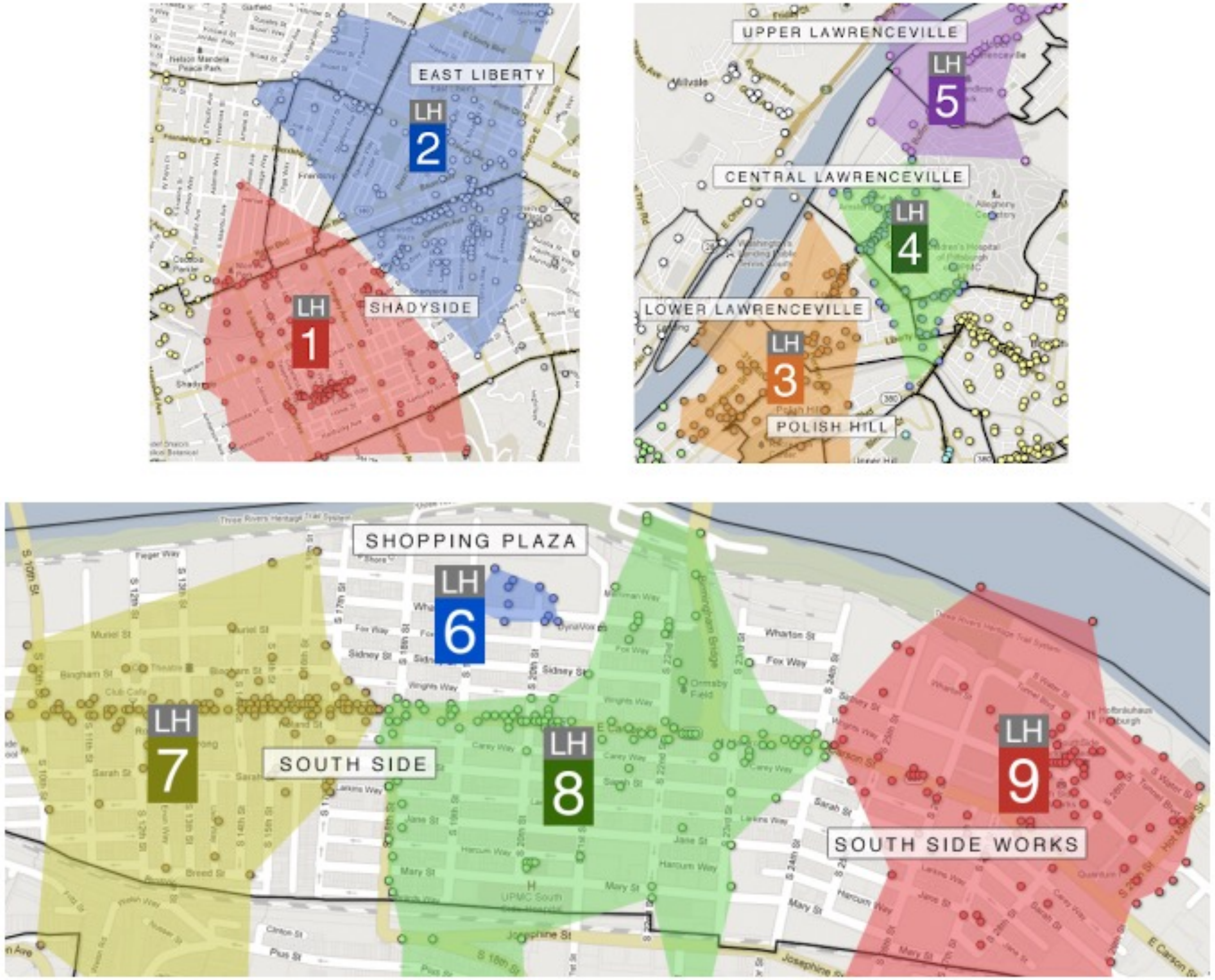
Results, Evaluation, and Validation

Over 3 areas of the city, 9 livehoods have been identified and validated by 27 Pittsburgh residents

- **Municipal Neighborhoods Borders:** livehoods are dynamic, and evolve as people's behaviours change, unlike fixed neighbourhoods set by the city government.
- **Demographics:** the interviews displayed strong evidence that the demographics of the residents and visitors of an area play a strong role in explaining the livehood divisions.
- **Development and Resources:** economic development can affect the character of an area. Similarly, the resources provided by a region has a strong influence on the people that visit it, and hence its resulting character.

Clustering Case Study

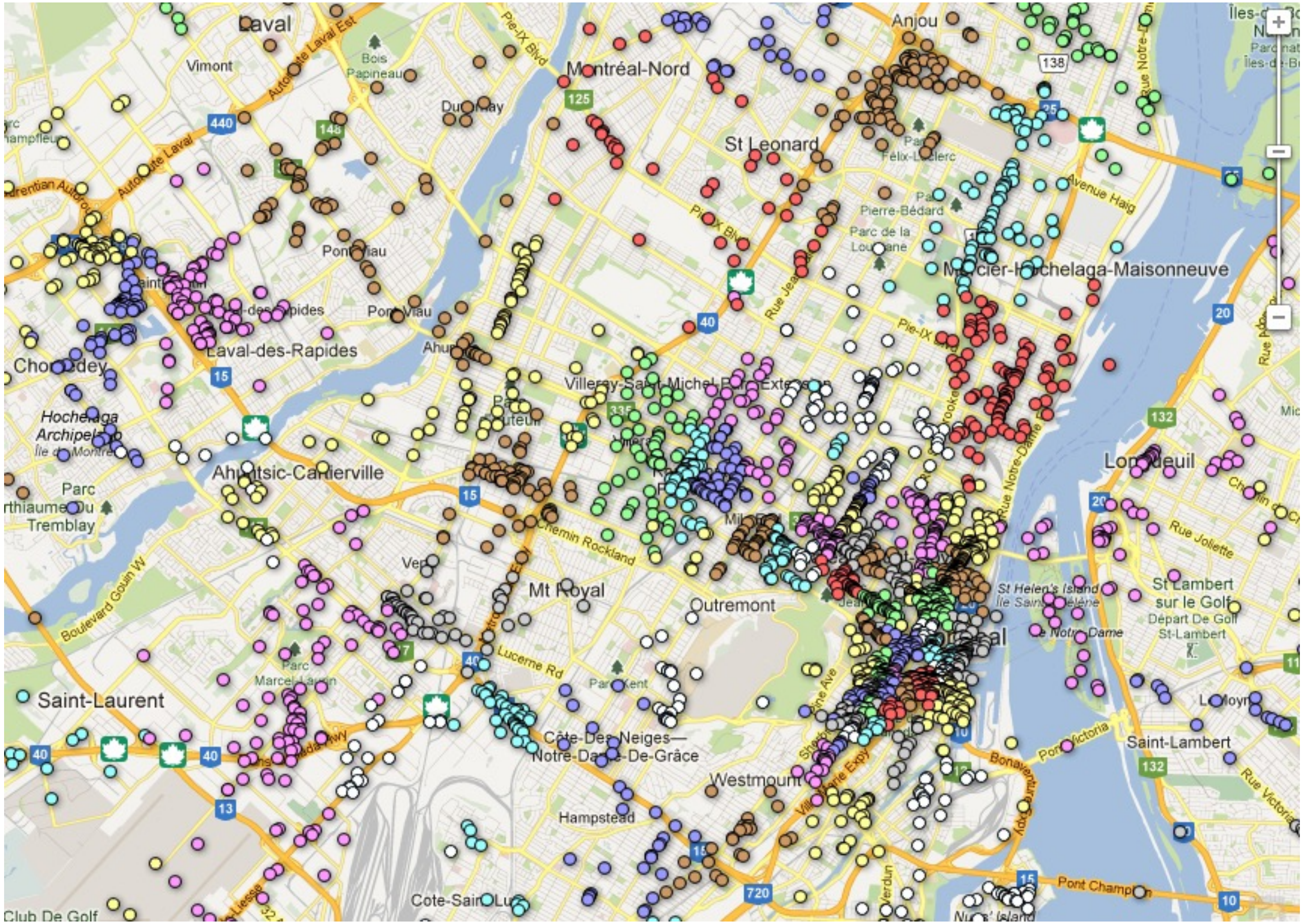
Livehoods



Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012

Clustering Case Study

Livehoods



Cranshaw *et al.*
The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City
ICWSM, 2012