Applied AI and ML Data Science Lab Ryerson University

datasciencelab.ca

Dr. Ayse Bener

DAA Toronto 2018
April 19, 2018

Ryerson University

Data Science Laboratory

Outline

- Machine Learning
- Application of Machine Learning
    - News Article Recommendation Based on the Analysis of Article's Content-Time Matters
- Data Science Lab

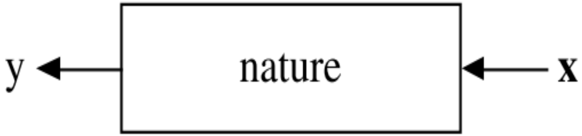
Statistics and Machine Learning

Statistics and Machine Learning

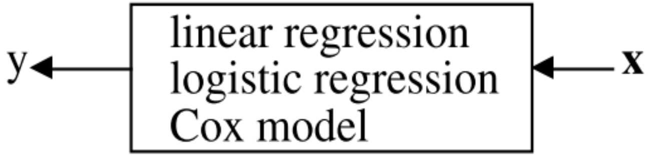
- Statistics and ML today both
  - use data sets
  - create data models
  - analyse relations within data
  - predict unseen input values

- What are the differences?
  - Statistics is a well established science, dates back to 1749
  - ML is field of Computer Science, dates back to 1959

Statistical data analysis



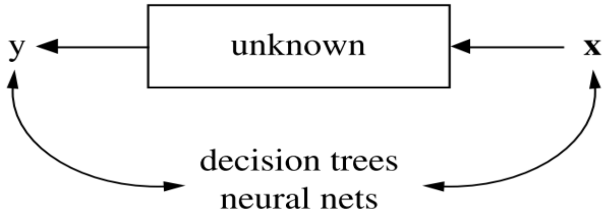
- Independent input vector **x**
- Dependent output vector **y**
- Nature functions to associate **x** with **y**



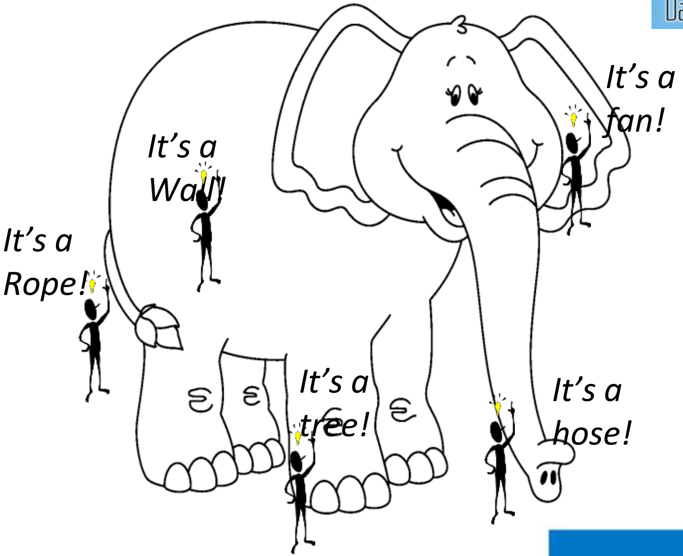
- Assume a stochastic data model
- Parameters are estimated from sampled data
- Goals: prediction, extract information about the nature
- Model validation: yes-no using goodness-of-fit tests and residual examination

[1] Breiman, L. (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3)

ML data analysis

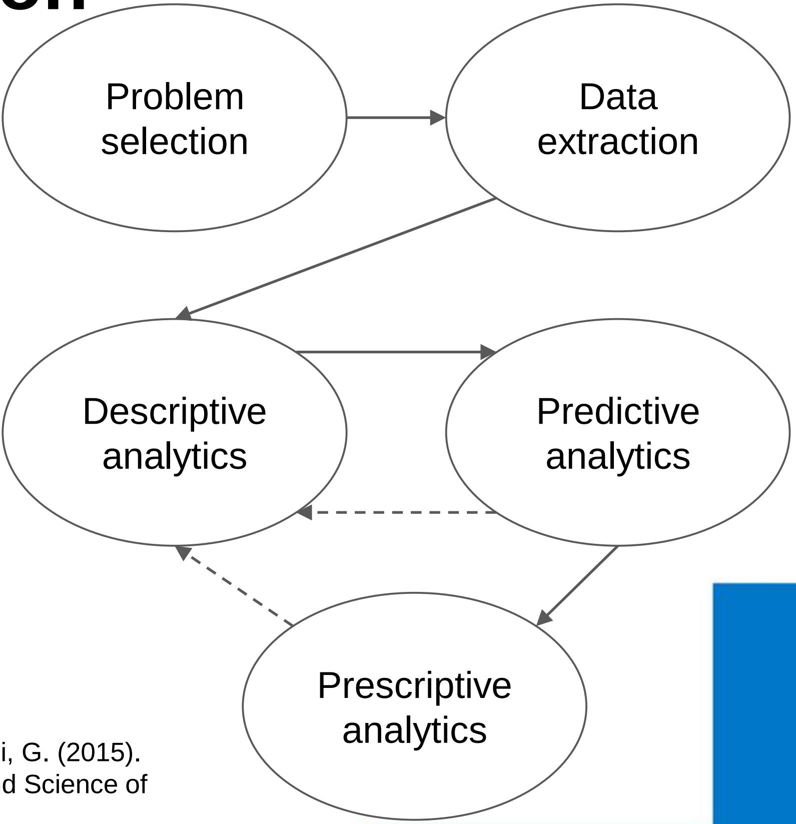


- Consider nature as unknown
- Relate **x** with **y** using an algorithm
- Model validation: measured by predictive accuracy
- Develop algorithms that learn model from data
- Repeat automated learning when needed



Stages of ML Application

1. Problem selection
2. Data extraction
3. Descriptive analytics
4. Predictive analysis
5. Prescriptive analytics



[4] Bener, A., Misirli, A. T., Caglayan, B., Kocaguneli, E., & Calikli, G. (2015). Lessons Learned from Software Analytics in Practice. The Art and Science of Analyzing Software Data

What Machine Learning is / is not?

- ML is not
  - a single software package
  - a one fits all tool
  - an easy task that can be mastered in a few years
- ML is
  - an algorithmic modelling approach
  - multiple customized solutions for each individual problem
  - requires expertise in
    - Problem domain
    - Statistics, Probability
    - Computer Science, Algorithm Design
  - often a teamwork of experts

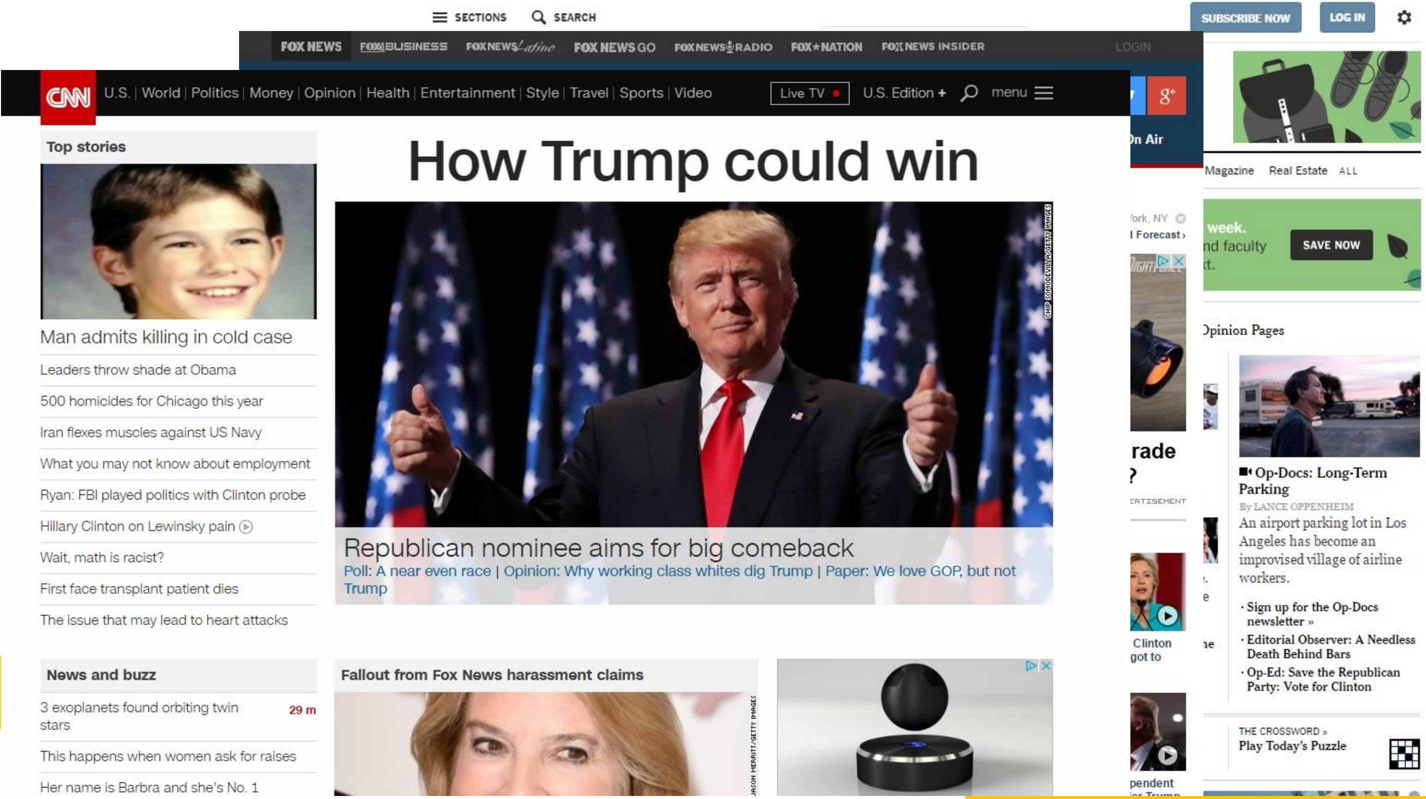
Application of Machine Learning

News Article Position Recommendation Based on the Analysis of Article’s Content- Time Matters

Parisa Lak, Ceni Babaoglu, Ayse Bener and Pawel Pralat, "News Article Position Recommendation Based on The Analysis of Article's Content - Time Matters", CBRecSys 2016, Sept 15-19, Boston, USA

Ryerson University

Introduction- the current problem



Top stories

Man admits killing in cold case

Leaders throw shade at Obama

500 homicides for Chicago this year

Iran flexes muscles against US Navy

What you may not know about employment

Ryan: FBI played politics with Clinton probe

Hillary Clinton on Lewinsky pain

Wait, math is racist?

First face transplant patient dies

The issue that may lead to heart attacks

How Trump could win

Republican nominee aims for big comeback

Poll: A near even race | Opinion: Why working class whites dig Trump | Paper: We love GOP, but not Trump

News and buzz

3 exoplanets found orbiting twin stars 29 m

This happens when women ask for raises

Her name is Barbra and she's No. 1

Fallout from Fox News harassment claims

week.
nd faculty
t.

Op-Docs: Long-Term Parking

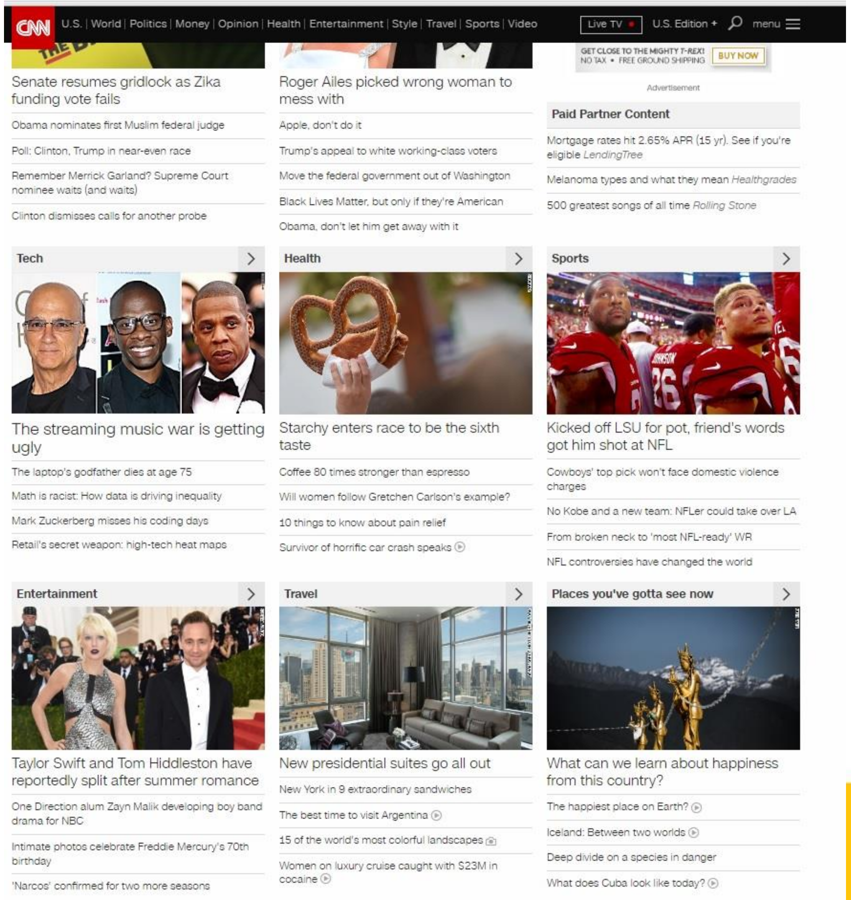
By LANCE OPPENHEIM

An airport parking lot in Los Angeles has become an improvised village of airline workers.

- Sign up for the Op-Docs newsletter »
- Editorial Observer: A Needless Death Behind Bars
- Op-Ed: Save the Republican Party: Vote for Clinton

News Article’s Position

Data Science Laboratory


CNN
U.S. | World | Politics | Money | Opinion | Health | Entertainment | Style | Travel | Sports | Video
Live TV
U.S. Edition +
menu

Senate resumes gridlock as Zika funding vote fails

Obama nominates first Muslim federal judge

Poll: Clinton, Trump in near-even race

Remember Merrick Garland? Supreme Court nominee waits (and waits)

Clinton dismisses calls for another probe

Roger Ailes picked wrong woman to mess with

Apple, don't do it

Trump's appeal to white working-class voters

Move the federal government out of Washington

Black Lives Matter, but only if they're American

Obama, don't let him get away with it

Paid Partner Content

Mortgage rates hit 2.65% APR (15 yr). See if you're eligible *LendingTree*

Melanoma types and what they mean *Healthgrades*

500 greatest songs of all time *Rolling Stone*

Tech

The streaming music war is getting ugly

The laptop's godfather dies at age 75

Math is racist: How data is driving inequality

Mark Zuckerberg misses his coding days

Retail's secret weapon: high-tech heat maps

Health

Starchy enters race to be the sixth taste

Coffee 80 times stronger than espresso

Will women follow Gretchen Carlson's example?

10 things to know about pain relief

Survivor of horrific car crash speaks

Sports

Kicked off LSU for pot, friend's words got him shot at NFL

Cowboys' top pick won't face domestic violence charges

No Kobe and a new team: NFLer could take over LA

From broken neck to 'most NFL-ready' WR

NFL controversies have changed the world

Entertainment

Taylor Swift and Tom Hiddleston have reportedly split after summer romance

One Direction alum Zayn Malik developing boy band drama for NBC

Intimate photos celebrate Freddie Mercury's 70th birthday

'Narcos' confirmed for two more seasons

Travel

New presidential suites go all out

New York in 9 extraordinary sandwiches

The best time to visit Argentina

15 of the world's most colorful landscapes

Women on luxury cruise caught with $23M in cocaine

Places you've gotta see now

What can we learn about happiness from this country?

The happiest place on Earth?

Iceland: Between two worlds

Deep divide on a species in danger

What does Cuba look like today?

Ryerson University

Problem

- How to predict a new article’s **position** in an online news **Homepage**?
- Assumption
    - New articles are clustered in different sections by editors

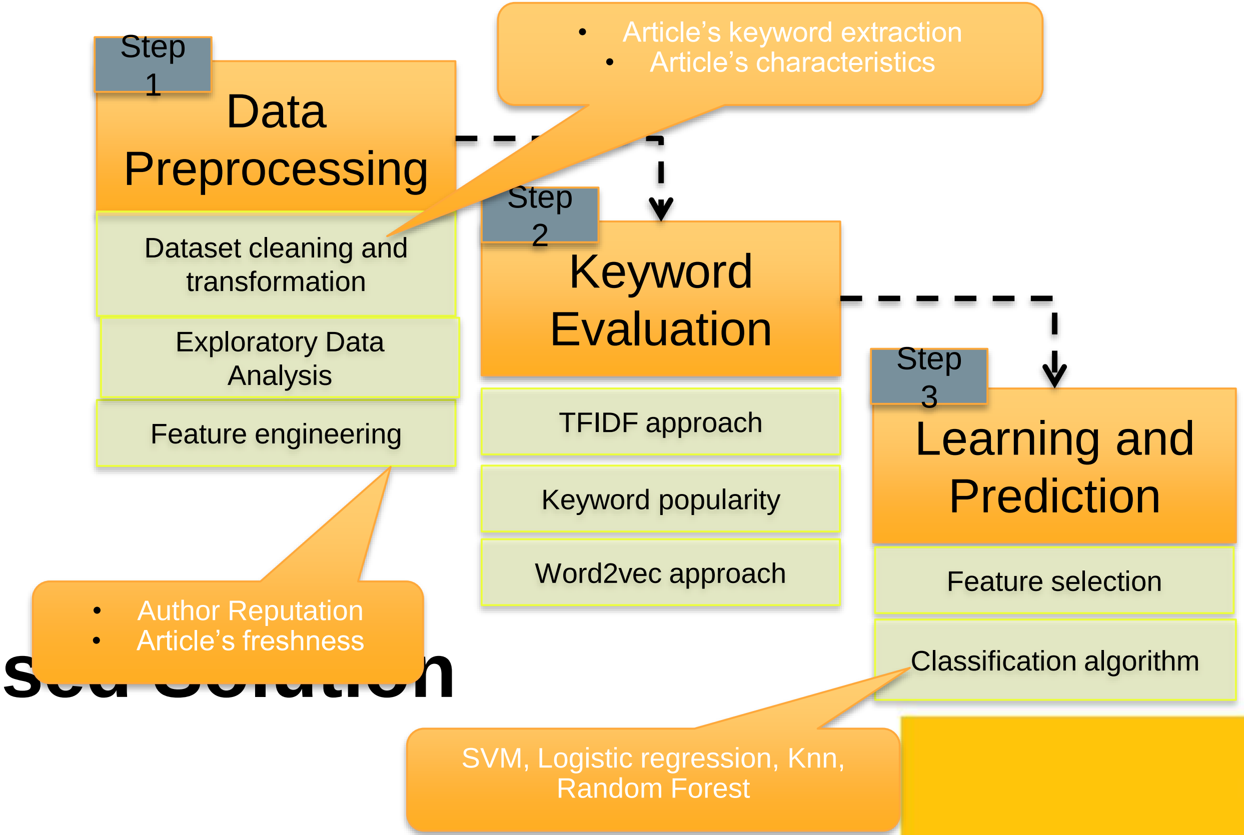
Background

- Article Popularity measure
  - Author’s influence
  - Hot topics
  - Number of visits
  - Duration of stay
  - Comments and likes
- Ranking new articles
  - Based on the similarity between the new article and previous articles

Further Questions

- How many previous articles should be selected?
- How old should the previous articles be?
  - Does time matter?
- Which algorithm(s) should be used?
- What is the best keyword evaluation technique?

Proposed Solution



Step 1

Data Preprocessing

Dataset cleaning and transformation

Exploratory Data Analysis

Feature engineering

- Author Reputation
- Article's freshness

Step 2

Keyword Evaluation

TFIDF approach

Keyword popularity

Word2vec approach

- Article's keyword extraction
- Article's characteristics

Step 3

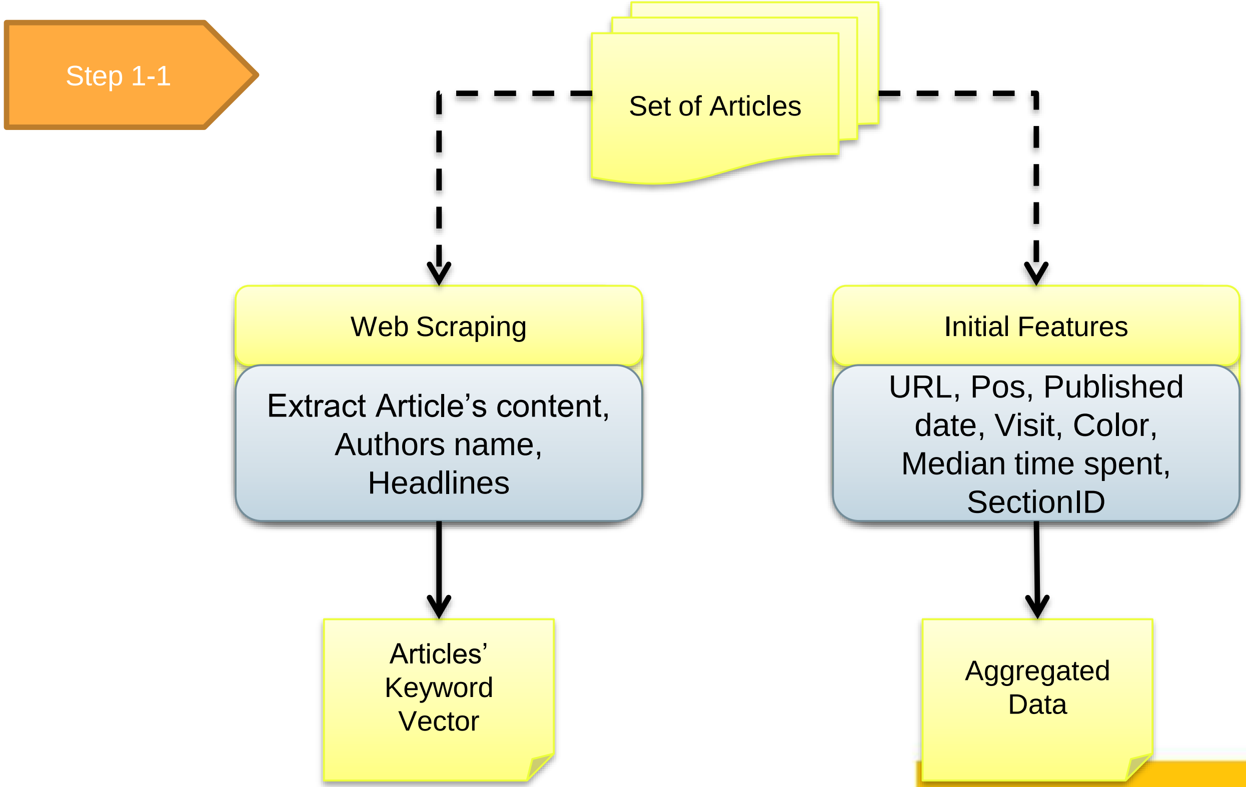
Learning and Prediction

Feature selection

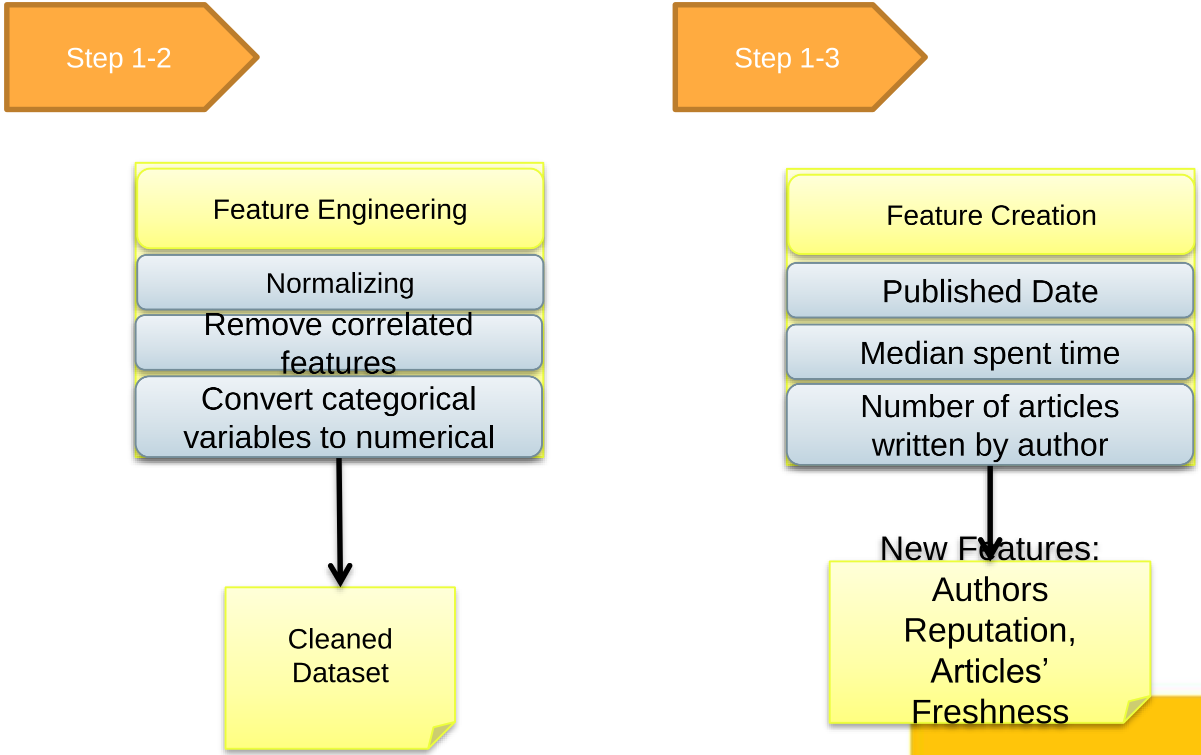
Classification algorithm

SVM, Logistic regression, Knn, Random Forest

Methodology: Data Preprocessing


Step 1-1
Set of Articles
Web Scraping
Extract Article’s content, Authors name, Headlines
Initial Features
URL, Pos, Published date, Visit, Color, Median time spent, SectionID
Articles’ Keyword Vector
Aggregated Data

Methodology: Data Preprocessing



Step 1-2

Feature Engineering

- Normalizing
- Remove correlated features
- Convert categorical variables to numerical

Cleaned Dataset

Step 1-3

Feature Creation

- Published Date
- Median spent time
- Number of articles written by author

New Features: Authors Reputation, Articles' Freshness

Methodology: TF-IDF Approach

- We tokenize each article and extract their keywords
- Calculate keywords' weights
    - TF: measures how often a term appears
        - Assuming that important terms appear more often
    - IDF: aims to reduce the weight of terms that appear in all articles

Methodology: Keyword Popularity Approach

- Keywords’ weights are measured based on
  - How many people visited a particular keyword e.g. *Canada*
  - Duration of keyword
    - [ArticleID=20, (Canada,2)] , [ArticleID=30, (Canada, 5)] a *Keyword_duration* for *Canada* is 7.
  - Combine these two factors and generate Keyword’s weights

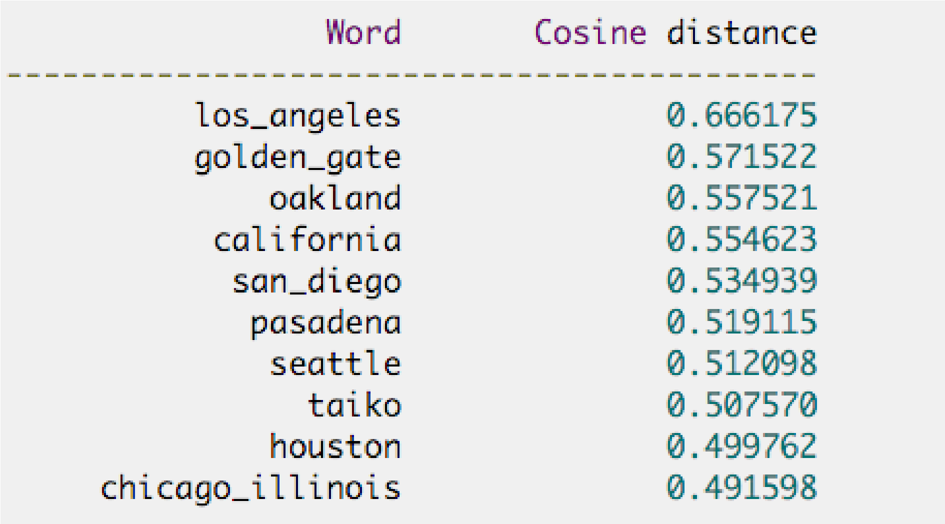
Methodology: Word2vec Approach

- Published by Google in 2013, is a two-layered neural net that processes text
    - Takes a text corpus as input and produces the word vectors as an output.

1. Constructs vocabulary from the training text
2. Creates numerical representation of words, feature vectors
3. Measures cosine similarity of words and group them together
    - e.g. word clustering, sentiment analysis, etc.

Example

- Given the corpus, the example output with the closest token to ‘San-francisco’ is:

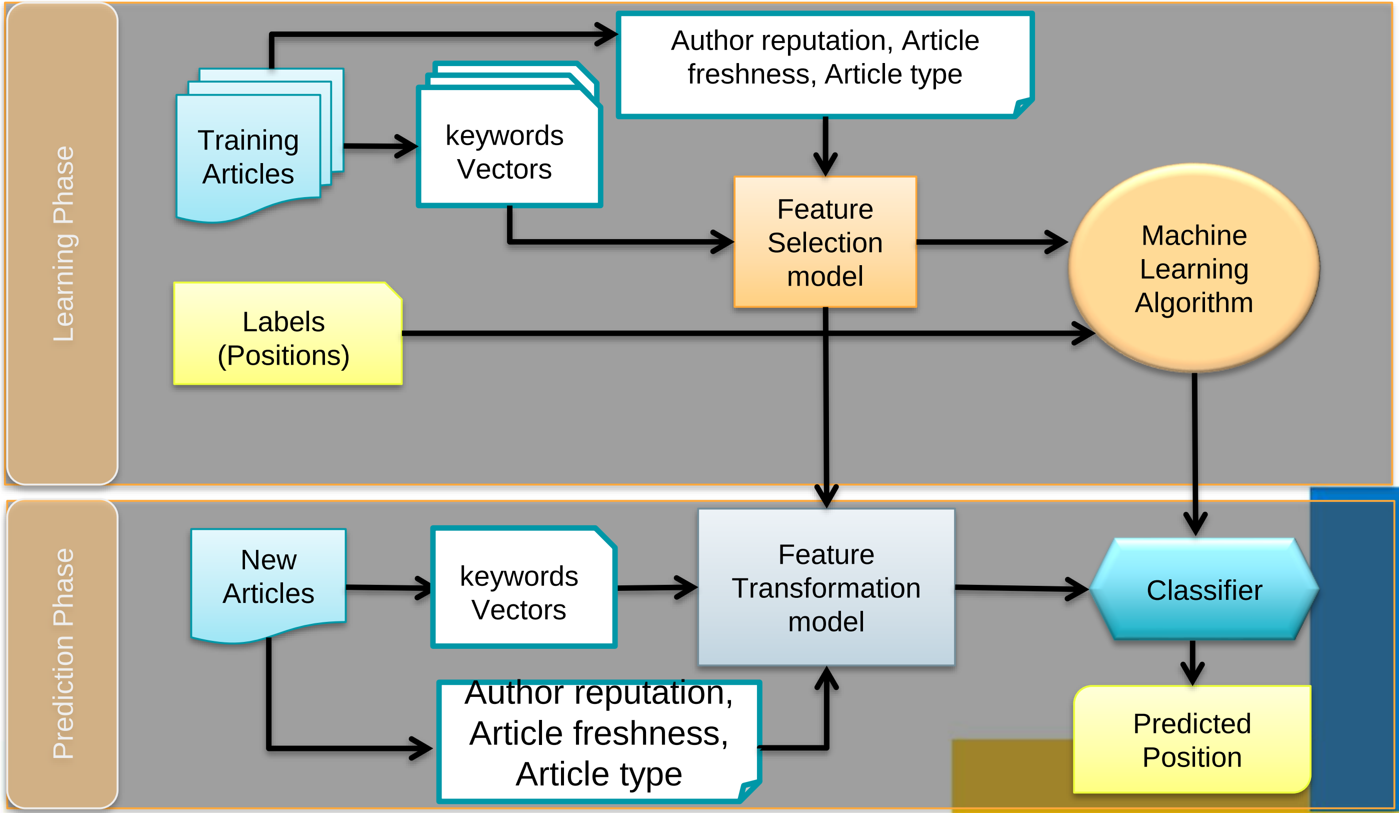


```
                Word       Cosine distance
-------------------------------------------
          los_angeles             0.666175
          golden_gate             0.571522
              oakland             0.557521
           california             0.554623
            san_diego             0.534939
             pasadena             0.519115
              seattle             0.512098
                taiko             0.507570
              houston             0.499762
     chicago_illinois             0.491598
```

Learning with Word2vec Features

- Build feature vectors of n-dimensions for each keyword
- Given set of keywords of each article
  - We calculate the *centroids* of the keywords; assign it to its article
- Train the model

Methodology: Learning and Prediction



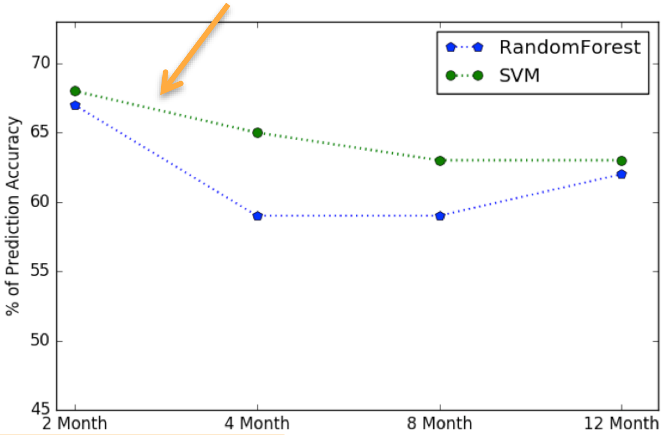
Methodology: Classification Algorithms, Evaluation, and Dataset

- Learning with articles from different time intervals
  - 2, 4, 8, 12 month prior to the test article’s date
- Accuracy measure

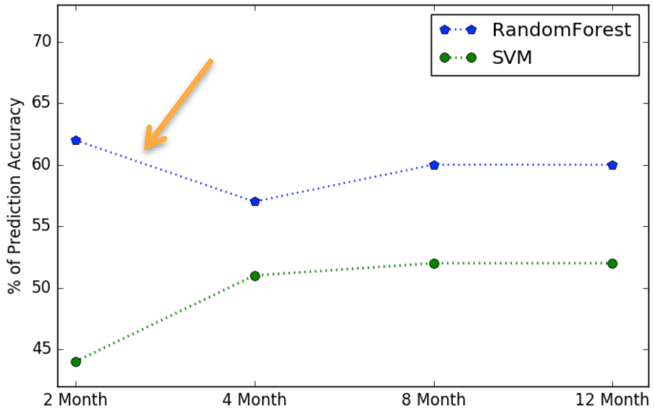
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

Number of observation for Different Time Intervals			
2 Month	**4 Month**	**8 Month**	**12 Month**
356	1114	2676	4532

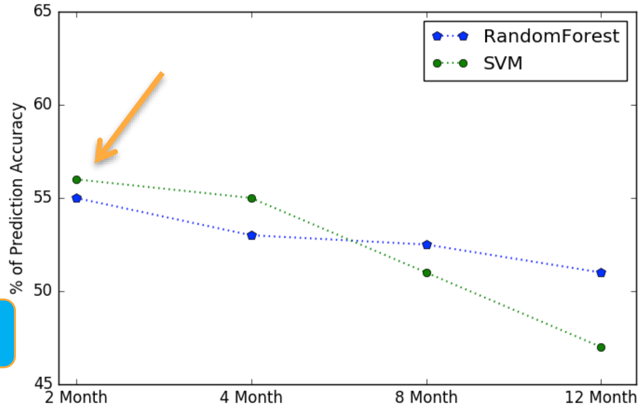
Results



TF-IDF

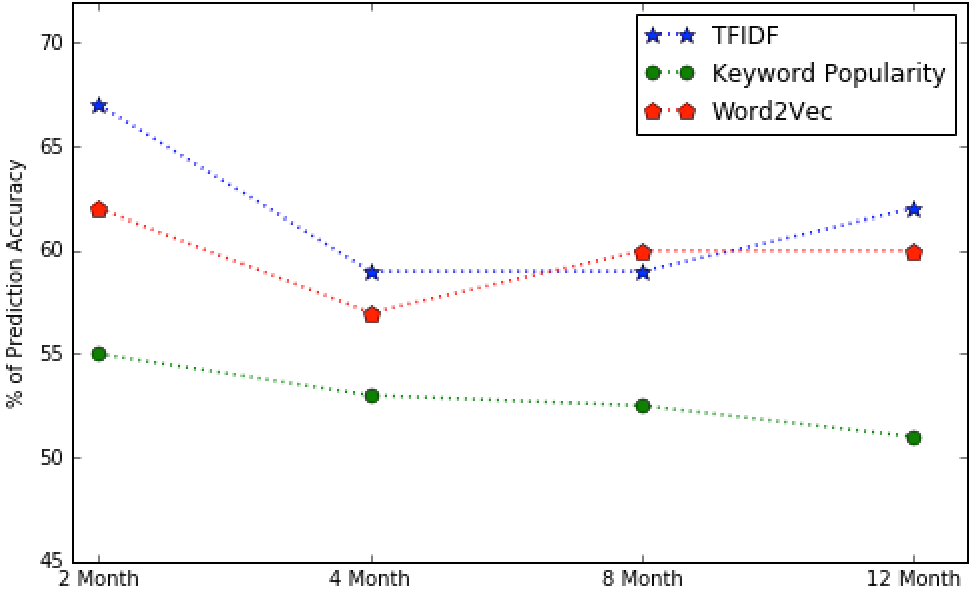


Word2Vec



Keyword popularity

Results



70
65
60
55
50
45
% of Prediction Accuracy
2 Month
4 Month
8 Month
12 Month
TFIDF
Keyword Popularity
Word2Vec

Threats to Validity

- The external threats to validity of this work is minimal
    - Extracted unique dataset, pre-processed the data to match our described analysis
- We minimized the internal threat to validity of this work
    - Reporting the results in terms of accuracy, a commonly used measure to communicate the result of prediction
- Threats to construct validity of this work is minimal
    - Compared the result of multiple text analytics techniques as well as classification algorithms
- The threats to conclusion validity of this work exists
    - Used predefined packages in R and Python to perform analysis and also the analysis is performed on a singe dataset

Conclusion and Future Work

- We may model expert’s article ranking with the suggested algorithms with more than 65% accuracy
  - For this dataset the best combination is TF-IDF and SVM respectively
- Experimental analysis shows that the classification performance is best when more recent articles are used for training
- Future Work:
  - Experiment with larger and more recent datasets
    - Large scale application of all algorithms
  - Use other keyword evaluation techniques
    - LDA or a hybrid version of former techniques

Data Science Lab (DSL) - Ryerson

- A research lab dedicated for Machine Learning applications
- 10+ industry partners including
  - Toronto Stock Exchange
  - IBM Canada
  - St Michael's Hospital
  - Communications Research Centre Canada
  - Globe and Mail Canada
  - Blackberry Canada
  - Manulife Canada
  - Toronto Police Services
- Numerous grants: federal, provincial, industry, government
- Data Science and Analytics MSc / Certificate Programs

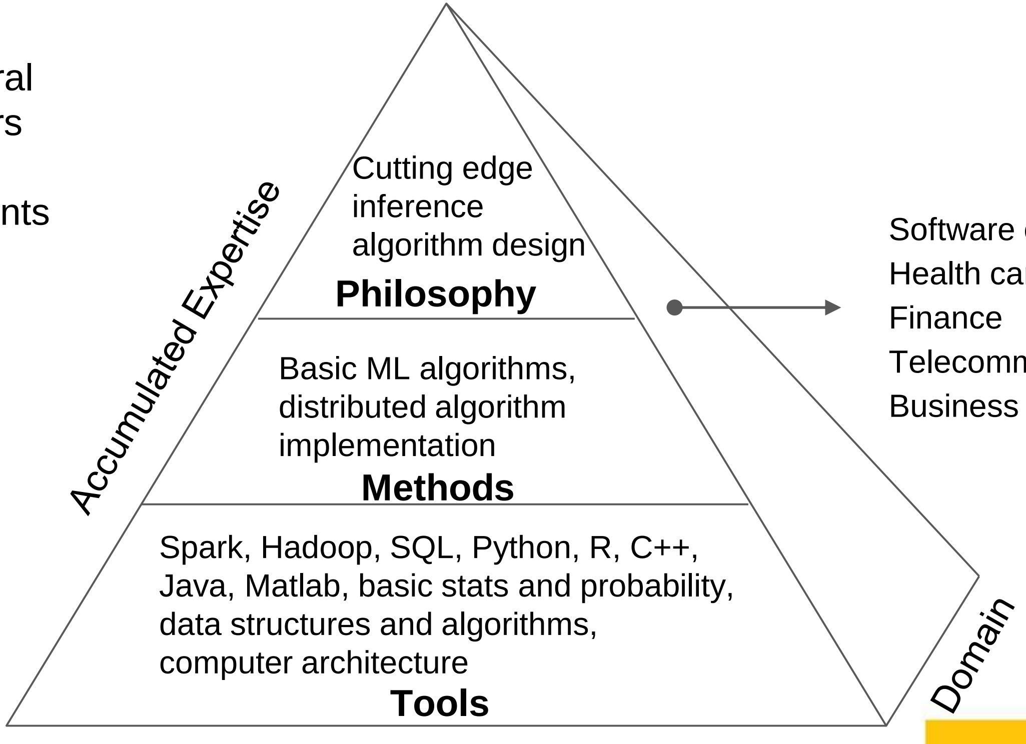
DSL expertise in Theoretical and Applied ML

Postdoctoral researchers

PhD students

Masters students

Certificate students



Cutting edge inference algorithm design

Philosophy

Basic ML algorithms, distributed algorithm implementation

Methods

Spark, Hadoop, SQL, Python, R, C++, Java, Matlab, basic stats and probability, data structures and algorithms, computer architecture

Tools

Accumulated Experience

Software engineering

Health care

Finance

Telecommunications

Business applications

Research Challenges

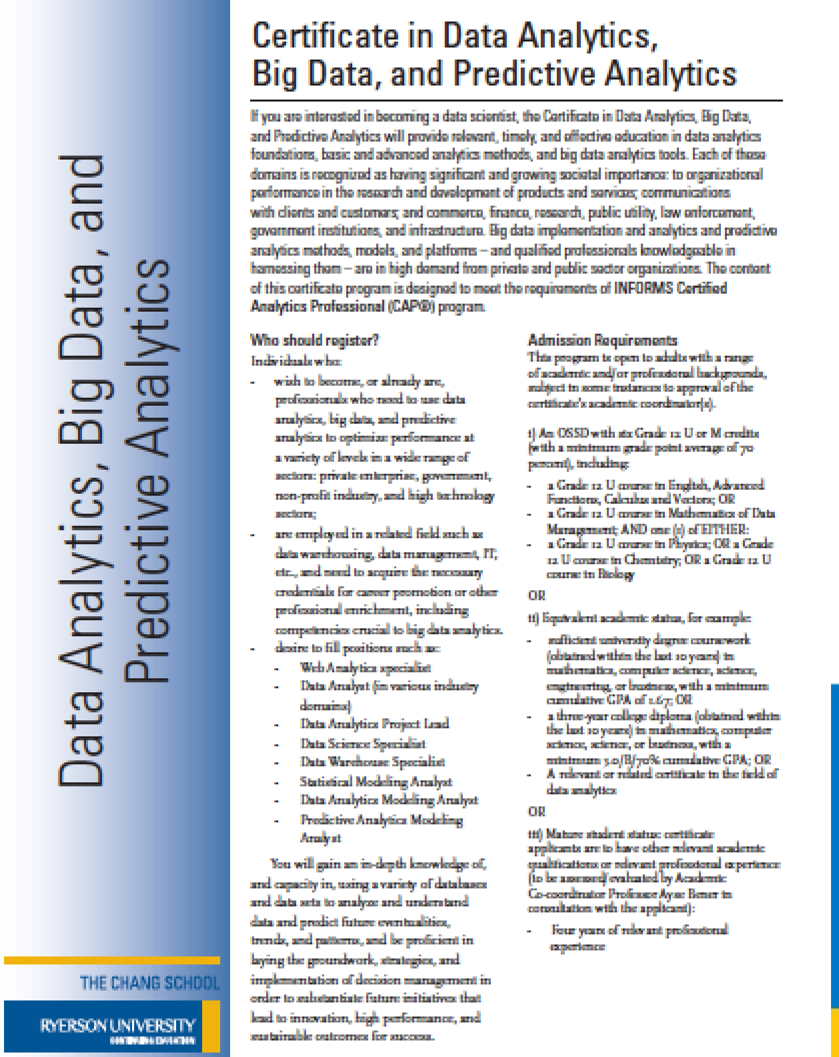
- Recommender systems and prediction models
  - Cold start
  - Temporal date
  - User biases, negative choices
  - Unstructured data
  - Change/ predict user behaviour
  - Recommendation update frequency
  - Randomness

Novelty in our Research

- Bayesian Machine Learning
  - Algorithms that learn, adopt, reason
    - Reinforcement learning, neural networks, Bayesian networks, model comparison
  - Embeddings
    - Tensor factorization, Gradient Boosted Decision Trees, Ensembles

Programs in Data Analytics

- Certificate in Data Analytics, Big Data, and Predictive Analytics
- Aligned with CAPS INFORMS
- Launched in September 2014



Data Analytics, Big Data, and Predictive Analytics

THE CHANG SCHOOL

RYERSON UNIVERSITY

Certificate in Data Analytics, Big Data, and Predictive Analytics

If you are interested in becoming a data scientist, the Certificate in Data Analytics, Big Data, and Predictive Analytics will provide relevant, timely, and effective education in data analytics foundations, basic and advanced analytics methods, and big data analytics tools. Each of these domains is recognized as having significant and growing societal importance: to organizational performance in the research and development of products and services; communications with clients and customers; and commerce, finance, research, public utility, law enforcement, government institutions, and infrastructure. Big data implementation and analytics and predictive analytics methods, models, and platforms – and qualified professionals knowledgeable in harnessing them – are in high demand from private and public sector organizations. The content of this certificate program is designed to meet the requirements of INFORMS Certified Analytics Professional (CAP®) program.

Who should register?

Individuals who:

- wish to become, or already are, professionals who need to use data analytics, big data, and predictive analytics to optimize performance at a variety of levels in a wide range of sectors: private enterprise, government, non-profit industry, and high technology sectors;
- are employed in a related field such as data warehousing, data management, IT, etc., and need to acquire the necessary credentials for career promotion or other professional enrichment, including competencies crucial to big data analytics.
- desire to fill positions such as:
  - Web Analytics specialist
  - Data Analyst (in various industry domains)
  - Data Analytics Project Lead
  - Data Science Specialist
  - Data Warehouse Specialist
  - Statistical Modeling Analyst
  - Data Analytics Modeling Analyst
  - Predictive Analytics Modeling Analyst

You will gain an in-depth knowledge of, and capacity in, using a variety of databases and data sets to analyze and understand data and predict future eventualities, trends, and patterns, and be proficient in laying the groundwork, strategies, and implementation of decision management in order to substantiate future initiatives that lead to innovation, high performance, and sustainable outcomes for success.

Admission Requirements

This program is open to adults with a range of academic and/or professional backgrounds, subject in some instances to approval of the certificate's academic coordinator(s).

i) An OSSD with six Grade 12 U or M credits (with a minimum grade point average of 70 percent), including:

- a Grade 12 U course in English, Advanced Functions, Calculus and Vectors; OR
- a Grade 12 U course in Mathematics of Data Management; AND one (1) of EITHER:
- a Grade 12 U course in Physics; OR a Grade 12 U course in Chemistry; OR a Grade 12 U course in Biology

OR

ii) Equivalent academic status, for example:

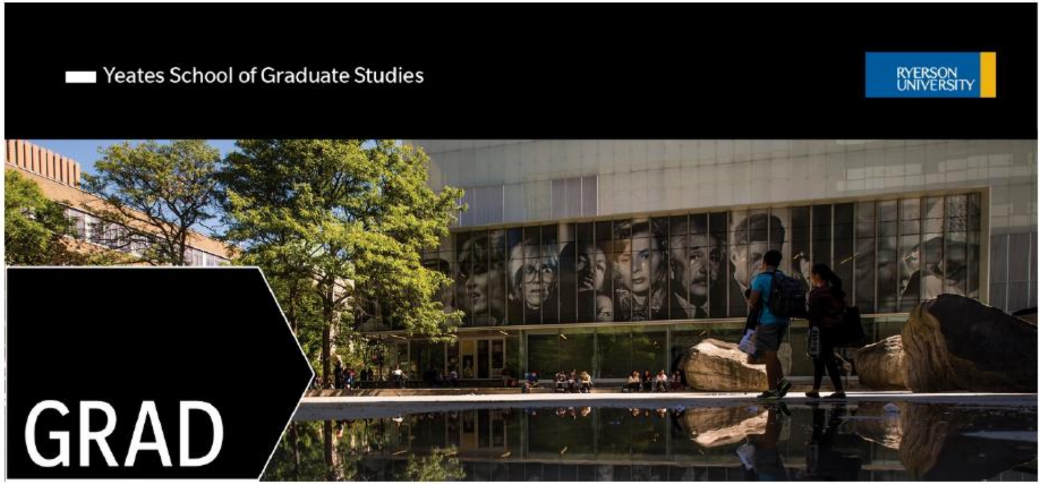
- sufficient university degree coursework (obtained within the last 10 years) in mathematics, computer science, science, engineering, or business, with a minimum cumulative GPA of 1.67; OR
- a three-year college diploma (obtained within the last 10 years) in mathematics, computer science, science, or business, with a minimum 3.0/B/70% cumulative GPA; OR
- A relevant or related certificate in the field of data analytics

OR

iii) Mature student status: certificate applicants are to have other relevant academic qualifications or relevant professional experience (to be assessed/evaluated by Academic Co-coordinator Professor Ayse Bener in consultation with the applicant):

- Four years of relevant professional experience

Programs in Data Analytics



Master of Science in Data Science and Analytics

https://www.ryerson.ca/graduate/datascience

Thanks for listening

http://www.datasciencelab.ca