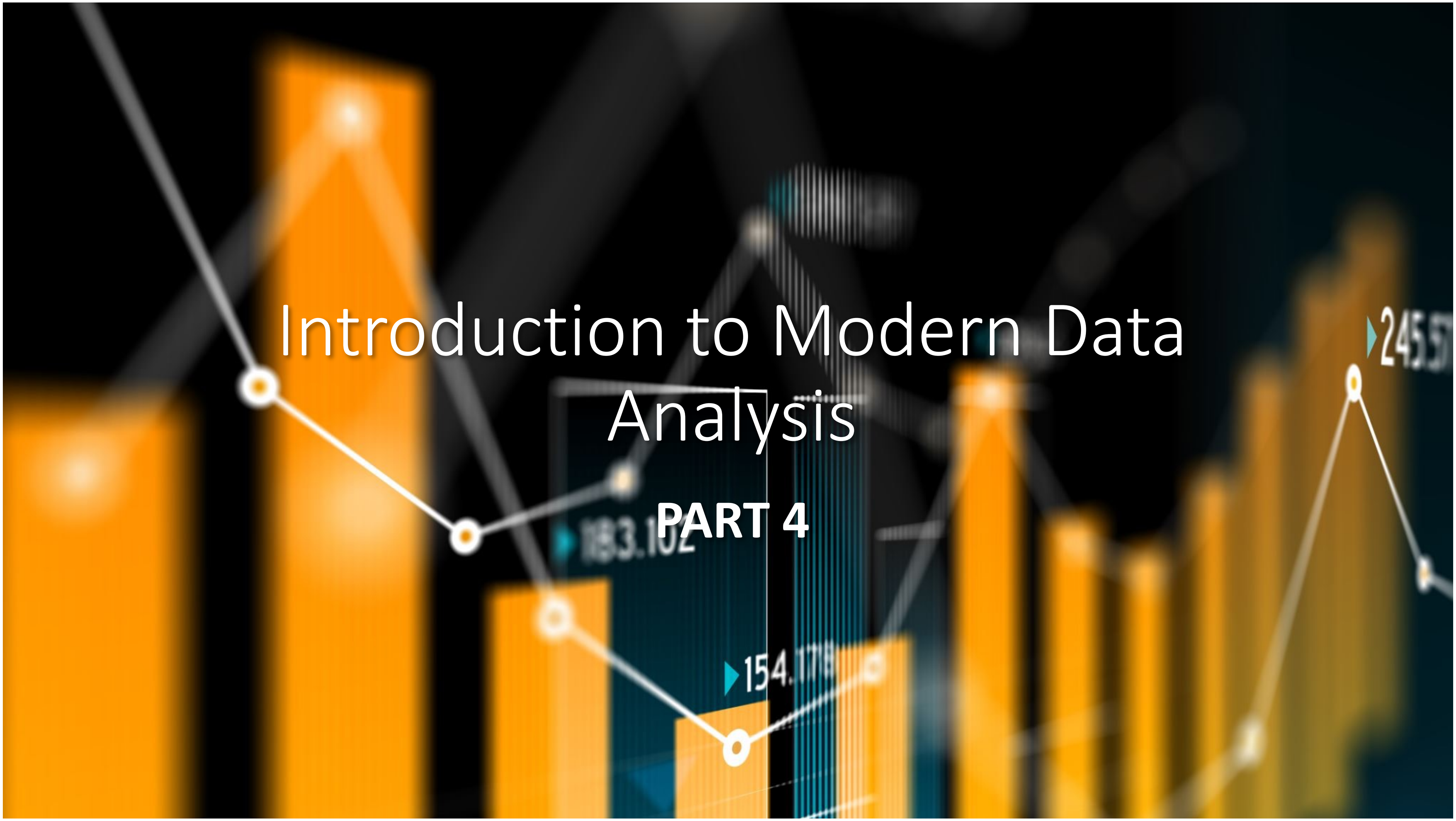
Introduction to Modern Data Analysis

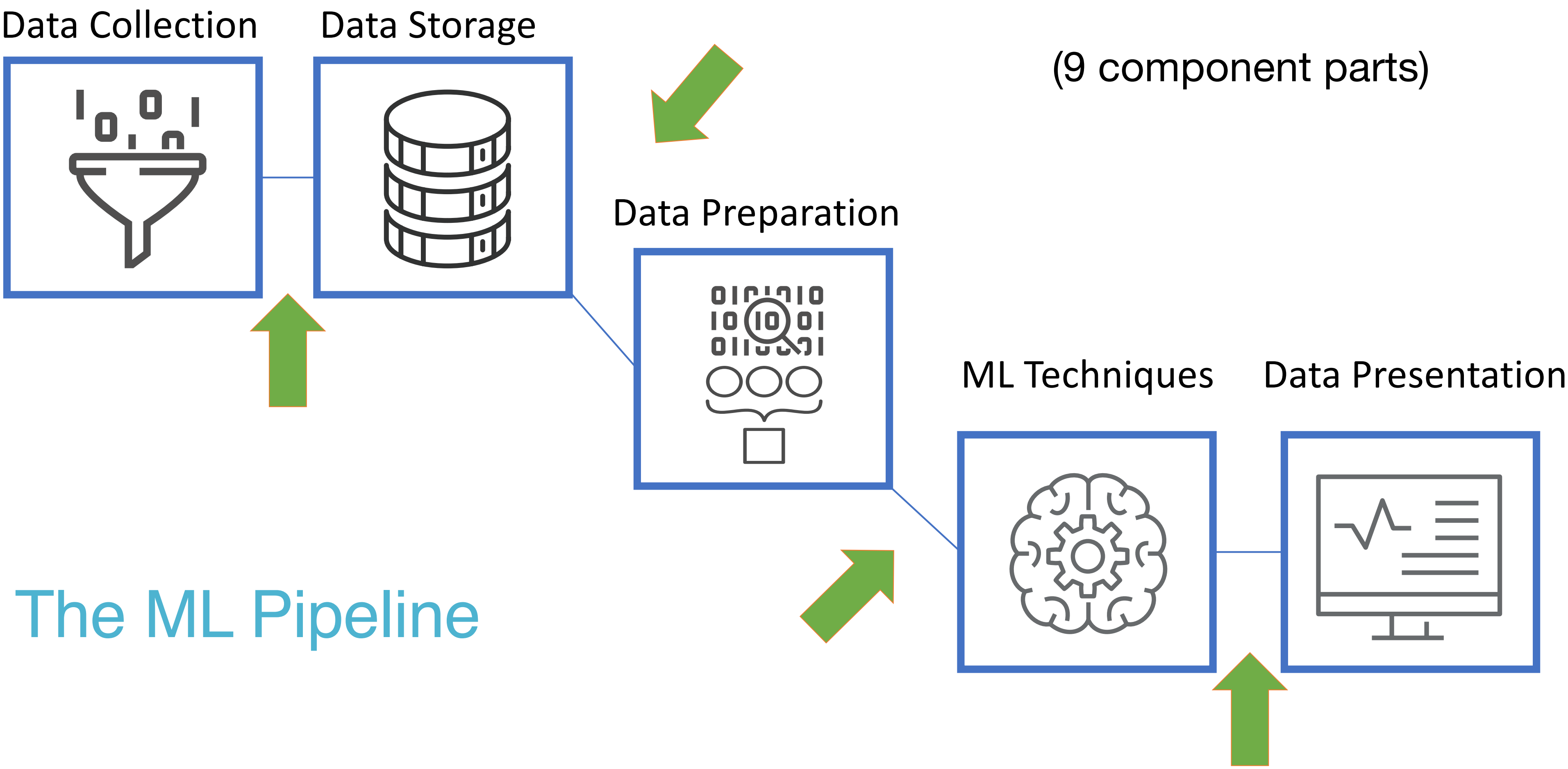
PART 4



Machine Learning/Artificial Intelligence

The ML Pipeline

(9 component parts)



Differences in this pipeline: Big Data

Machine learning can sometimes work on ‘small data’ and ‘business data’ BUT techniques are optimized for big data (plus sensor data)

This effects all aspects of the pipeline – collection, storage, cleaning and prep, presentation. For starters, manual is usually no longer an option!

Easy to stand up an ad-hoc pipeline in R BUT will this scale to a professional level?

To be able to champion these considerations in your organization, you must understand how ML analysis works

ML/AI A Very Quick Tool Discussion

Things to think about when you select analysis tools

A. Capability: What is their functionality + performance – do they have all the techniques, do they have the processing power
B. Integration: How do they connect to other parts of your pipeline
C. User-Experience: What is the user experience like – what background/level of expertise do you need to operate this tool, how easy is it to use this tool?
D. Cost – short and long term

Tool for Machine Learning Analysis

R packages

Python modules

R will 99.99% guaranteed have any machine learning technique you want (no matter how esoteric!) for free. Python probably will have most as well.

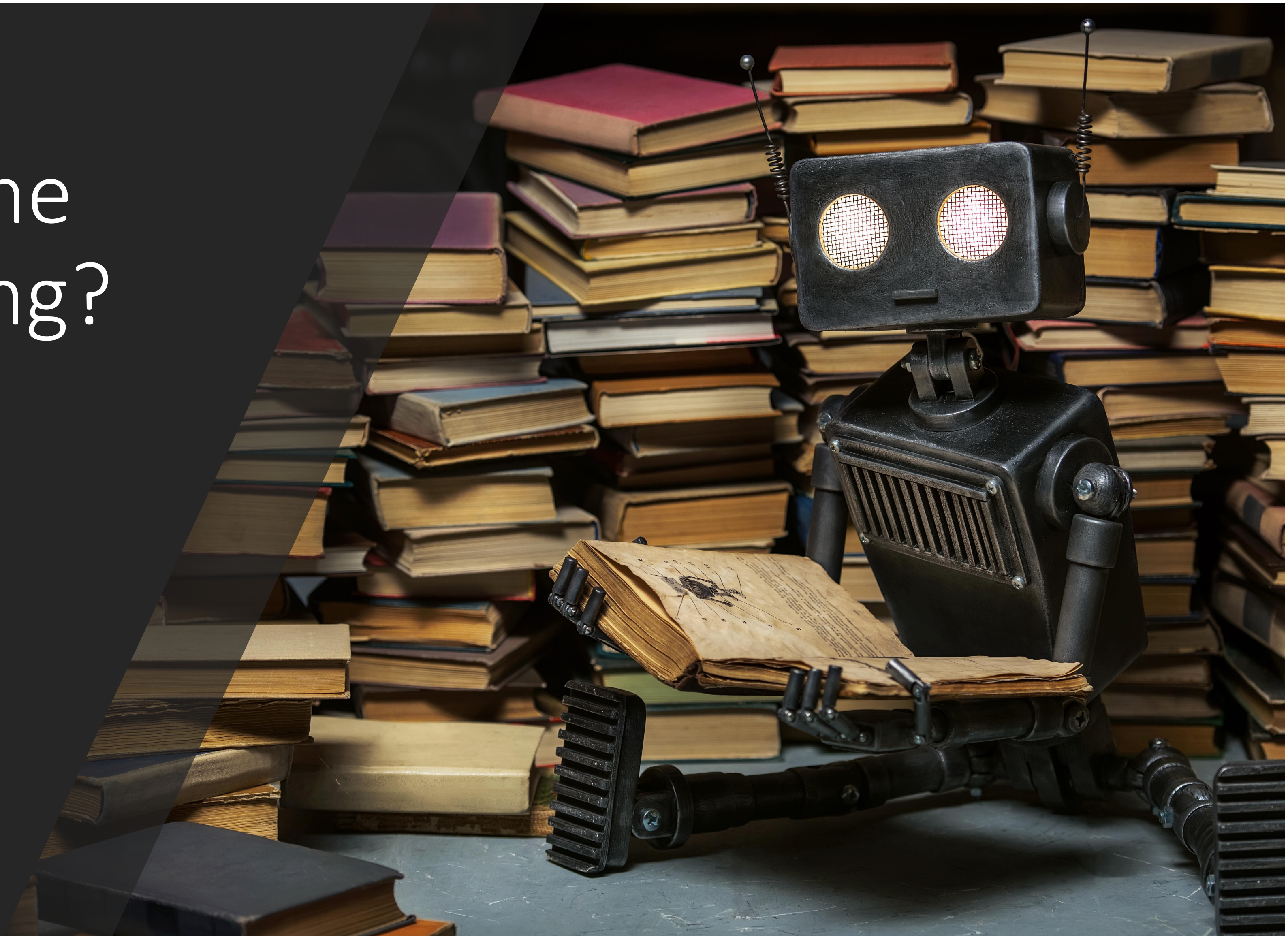
They will also have hooks into more niche/sophicated options (e.g. tensorflow pytorch)

Other tools will say they have machine learning implemented. This typically means one of a few things:

- You can embed R or Python into their tool
- They have implemented a small representative selection of available ML techniques (e.g. k-means clustering algorithm but not DB-scan, EM-clustering, etc.)

ML/AI Methods

Machine Learning?



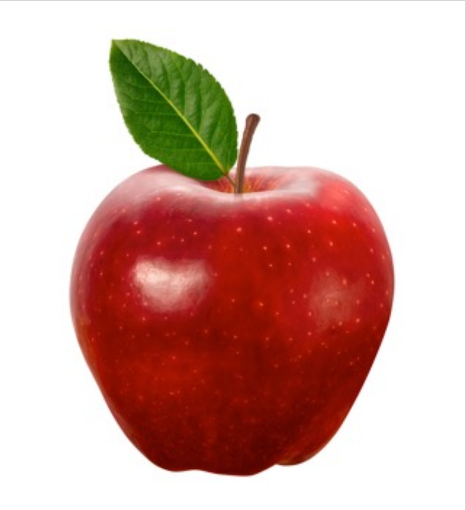
Human Learning

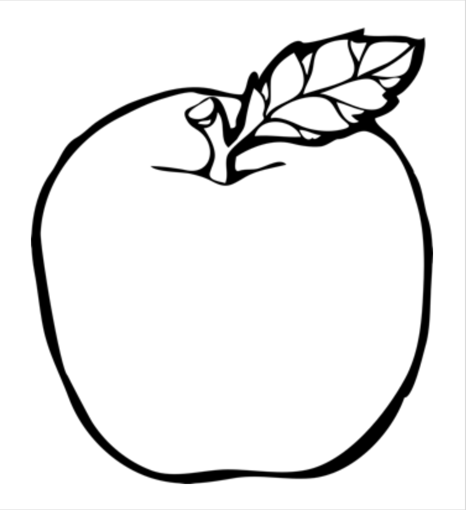
- **Supervised**: we give you some examples, you learn from them
- **Unsupervised**: you learn on your own, based on what you experience
- **Reinforcement**: The environment is your teacher

<u>All of these activities require a lot of data!</u>

Conceptual Model

The result of our learning is a concept or set of concepts that we can use when we encounter new situations





Machine Learning

- **Supervised techniques**:
  - Classification
  - Value Prediction (Regression)
- **Unsupervised techniques**:
  - Association rules
  - Clustering (Novel Categories + Concepts)
- **Reinforcement Learning**

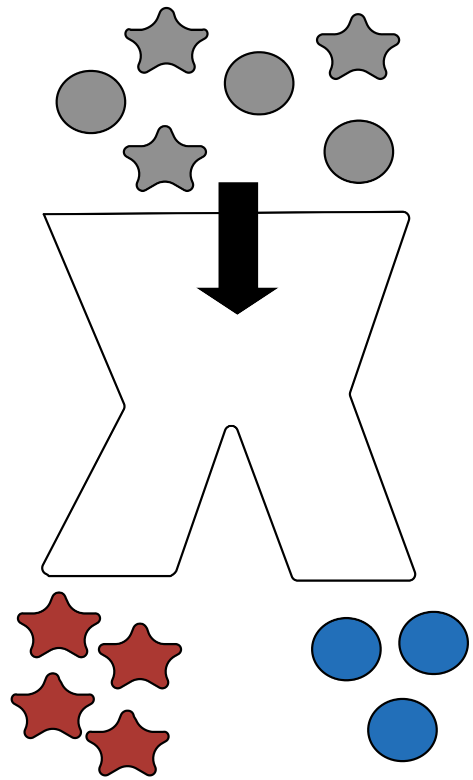
These output a <u>model</u> (black box?) that can be used to process new data

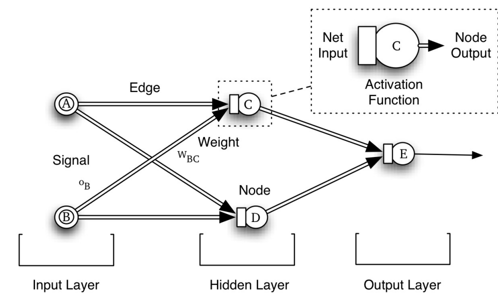
Machine Learning Practicalities

Just as is the case with humans, data science / machine learning techniques often need a large amount of the right kind of data to be successful

Classification

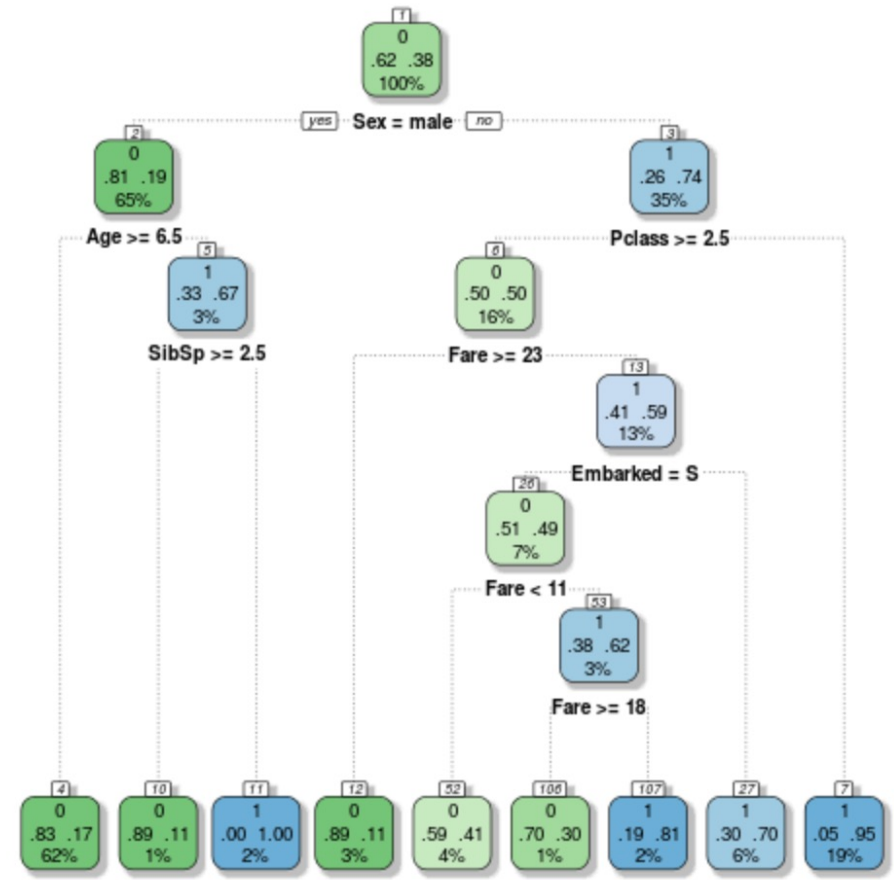
- **Classifier**: If I'm presented with an object, can I classify it into one of several predefined categories?
- Many different techniques to carry this out, but the steps are the same:
    - Use a *training set* to teach the classifier how to classify.
    - Test/validate the classifier using *new data*
    - Use the classifier to classify *novel instances*
- Some classifiers (e.g. neural nets) are very 'black box'. They might be good at classifying, but you don't know why!





Decision Tree Classifiers

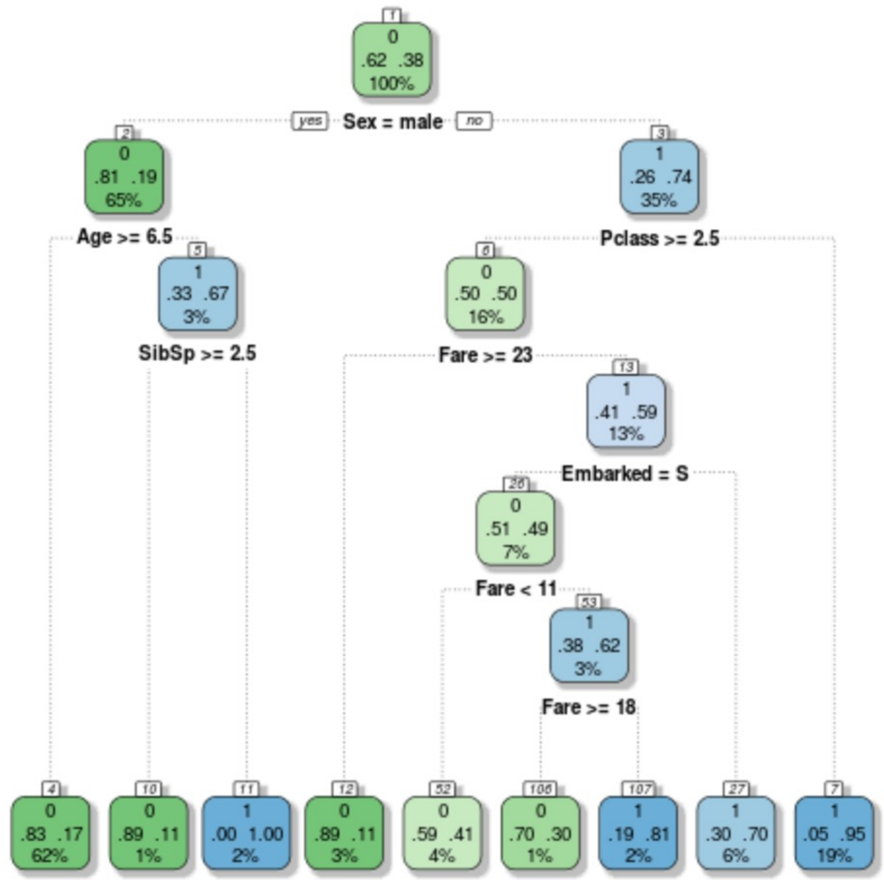
- Decision tree: what properties do you have? I'll (methodically) use this information to help me classify you.
- There are techniques we can use to *automatically* build these decision trees.
- Once the tree is built, we can see how the decision is made.
- These are also useful for expert systems



62% are male aged >=6.5. Of those, 83% are 'green category'.

Decision Tree Classifiers

- Decision tree: what properties do you have? I'll (methodically) use this information to help me classify you.
- There are techniques we can use to *automatically* build these decision trees.
- Once the tree is built, we can see how the decision is made.
- These are also useful for expert systems



Examples: Classification

See How Artificial Intelligence Can Improve Medical Diagnosis And Healthcare

Jennifer Kite-Powell Contributor

A digital health company from the UK wants to change the way a patient interacts with a doctor through the creation of an artificial intelligence (AI) doctor in the form of an AI chatbot.

Doctors using Infervision's AI powered CT diagnosis at Shanghai Changzheng Hospital in China. IMAGE COURTESY OF INFERVISION

Babylon Health raised close to $60 million in April 2017 to diagnose illnesses with an AI chatbot on your smartphone. Around the same time, Berlin and London based start up Ada announced its push into the

Predict Loan Default Using Seahorse and SparkR

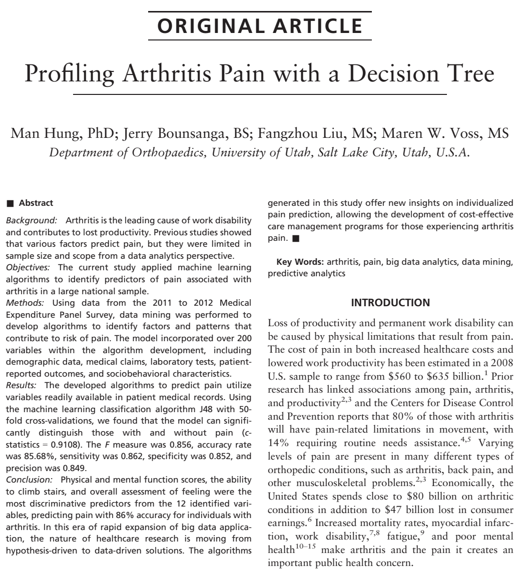
Loans can be risky, to say the least — so predicting loan defaults is an awesome possibility. Learn how to predict loan default of Lending Club, the largest online marketplace to connect borrowers and investors.

by Rathnadevi Manivannan · Nov. 02, 17 · AI Zone · Tutorial

Like (1) Comment (0) Save Tweet 2,960 Views

Data scientists are using Python and R to solve data problems due to the ready availability of these packages. These languages are often limited, as the data is processed on a single machine where the

Decision Tree Case Study



ORIGINAL ARTICLE

Profiling Arthritis Pain with a Decision Tree

Man Hung, PhD; Jerry Bounsanga, BS; Fangzhou Liu, MS; Maren W. Voss, MS

Department of Orthopaedics, University of Utah, Salt Lake City, Utah, U.S.A.

■ **Abstract**

Background: Arthritis is the leading cause of work disability and contributes to lost productivity. Previous studies showed that various factors predict pain, but they were limited in sample size and scope from a data analytics perspective.
Objectives: The current study applied machine learning algorithms to identify predictors of pain associated with arthritis in a large national sample.
Methods: Using data from the 2011 to 2012 Medical Expenditure Panel Survey, data mining was performed to develop algorithms to identify factors and patterns that contribute to risk of pain. The model incorporated over 200 variables within the algorithm development, including demographic data, medical claims, laboratory tests, patient-reported outcomes, and sociobehavioral characteristics.
Results: The developed algorithms to predict pain utilize variables readily available in patient medical records. Using the machine learning classification algorithm J48 with 50-fold cross-validations, we found that the model can significantly distinguish those with and without pain (*c*-statistics = 0.9108). The *F* measure was 0.856, accuracy rate was 85.68%, sensitivity was 0.862, specificity was 0.852, and precision was 0.849.
Conclusion: Physical and mental function scores, the ability to climb stairs, and overall assessment of feeling were the most discriminative predictors from the 12 identified variables, predicting pain with 86% accuracy for individuals with arthritis. In this era of rapid expansion of big data application, the nature of healthcare research is moving from hypothesis-driven to data-driven solutions. The algorithms generated in this study offer new insights on individualized pain prediction, allowing the development of cost-effective care management programs for those experiencing arthritis pain. ■

Key Words: arthritis, pain, big data analytics, data mining, predictive analytics

INTRODUCTION

Loss of productivity and permanent work disability can be caused by physical limitations that result from pain. The cost of pain in both increased healthcare costs and lowered work productivity has been estimated in a 2008 U.S. sample to range from \$560 to \$635 billion.[1] Prior research has linked associations among pain, arthritis, and productivity[2,3] and the Centers for Disease Control and Prevention reports that 80% of those with arthritis will have pain-related limitations in movement, with 14% requiring routine needs assistance.[4,5] Varying levels of pain are present in many different types of orthopedic conditions, such as arthritis, back pain, and other musculoskeletal problems.[2,3] Economically, the United States spends close to \$80 billion on arthritic conditions in addition to \$47 billion lost in consumer earnings.[6] Increased mortality rates, myocardial infarction, work disability,[7,8] fatigue,[9] and poor mental health[10–15] make arthritis and the pain it creates an important public health concern.

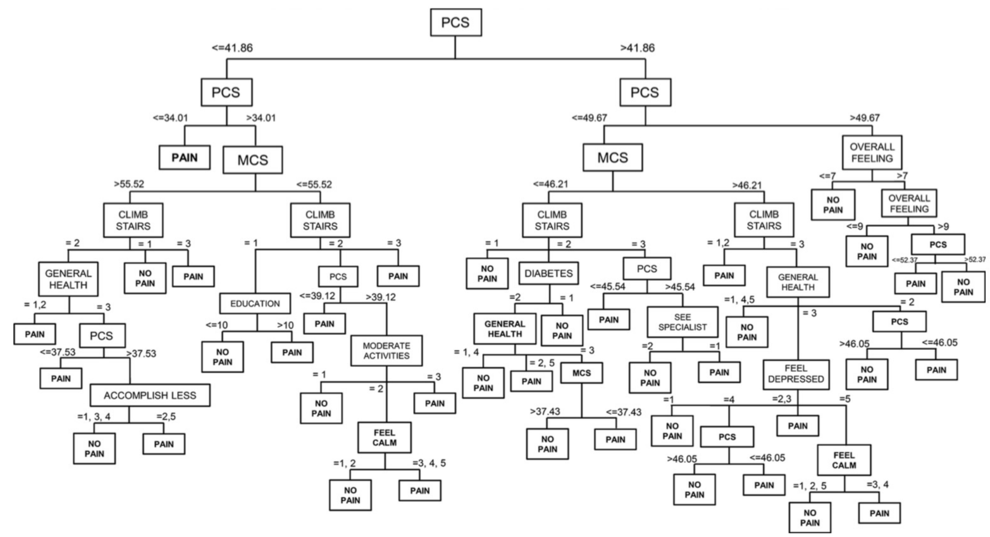
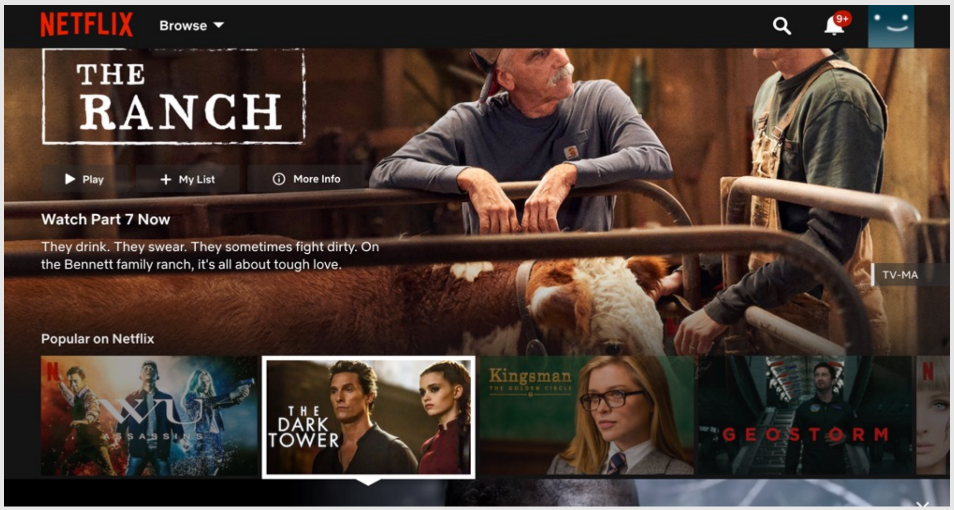


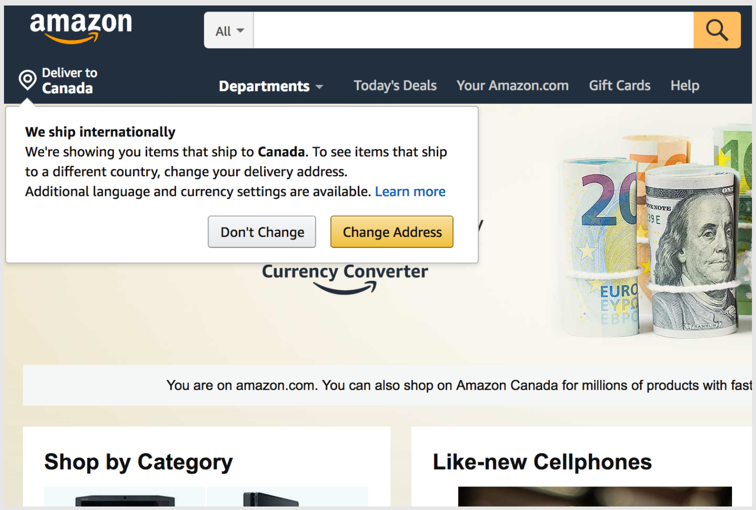
Figure 3. Predictors of pain tree diagram. PCS, Physical Component Summary; MCS, Mental Component Summary.

Unsupervised Learning Techniques

- Automated behaviours vs intelligent behaviours
- **Supervised**: we give you some examples, you learn from them
- **Unsupervised**: you learn on your own, based on what you experience
- Unsupervised techniques:
  - Association rules
  - Recommender engines
  - Novel categories (clustering)

Example: Clustering

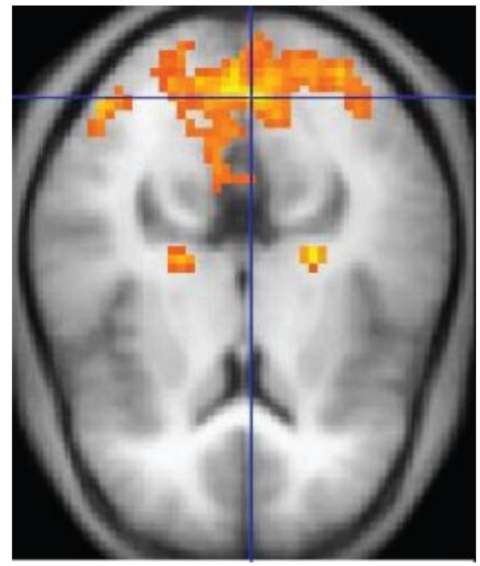




Clustering Case Study

Detecting Alzheimer's Disease

- Mild cognitive impairments (MCI) are a known to be a risk for factor for development of Alzheimer's Disease
- MCI are accompanied by changes in brain structure
- But which changes indicate that people will go on to develop Alzheimer's?
- A number of different data science techniques applied to MRI data: Support Vector Machines, Bayesian Statistics, Voting Feature Intervals, Feature Extraction and (last but not least) DBSCAN
- DBSCAN is used once voxels that provide high information about the classification of the image are identified using entropy based measures
- DBSCAN then groups pixels with similar spatial and information levels to determine which parts of the brain are the most important for the diagnosis



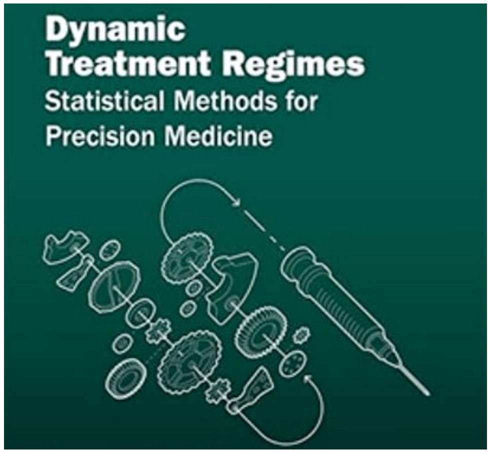
FMRI highlighting some areas of the pre-frontal cortex.

Reinforcement Learning Techniques

- Work in progress
- Requires environmental feedback
- Increasingly possible as things become more digital
- Still in the research lab stage
- As we increasingly see digitial transformation this will become more prevelant

Example: Reinforcement Learning

- We are only now starting to see real world applications of reinforcement learning.
- This is being enabled by an increasingly digital environment.
- Stay tuned!

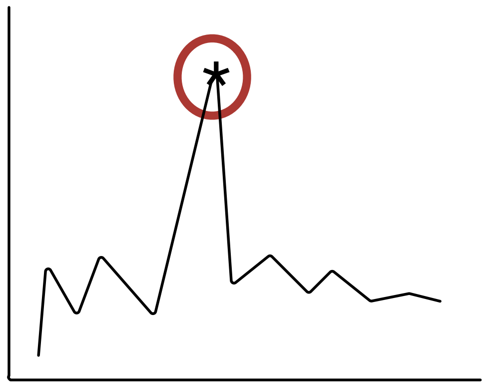


Mixed Methods

- There are some areas of ML that are focused more on the outcome
  - Anomaly Detection
  - Recommender Engines
- These may mix and match strategies from all of the categories

Anomaly Detection

- Anomaly: An unexpected, unusual, atypical or statistically unlikely event
- Wouldn't it be nice to have a data analysis pipeline that alerted you when things were out of the ordinary?
- Many different analytic approaches to take!
  - Clustering
  - Naïve Bayes
  - Association rules deviation
  - Ensemble techniques



Anomaly Detection Case Study

Energy 157 (2018) 336–352

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings

Alfonso Capozzoli*, Marco Savino Piscitelli, Silvio Brandi, Daniele Grassi, Gianfranco Chicco

Dipartimento Energia "Galileo Ferraris", Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

A R T I C L E I N F O

Article history:
Received 5 February 2018
Accepted 19 May 2018
Available online 21 May 2018

Keywords:
Energy consumption
Building energy management
Adaptive symbolic aggregate approximation
Anomaly detection
Data mining
Smart buildings

A B S T R A C T

The energy management of buildings currently offers a powerful opportunity to enhance energy efficiency and reduce the mismatch between the actual and expected energy demand, which is often due to an anomalous operation of the equipment and control systems. In this context, the characterisation of energy consumption patterns over time is of fundamental importance. This paper proposes a novel methodology for the characterisation of energy time series in buildings and the identification of infrequent and unexpected energy patterns. The process is based on an enhanced Symbolic Aggregate approXimation (SAX) process, and it includes an optimised tuning of the time window width and of the symbol intervals according to the building energy behaviour. The methodology has been tested on the whole electrical load of buildings for two case studies, and its flexibility and robustness have been confirmed. In order to demonstrate the implications for a preliminary diagnosis, some unexpected trends of the total electrical load have also been discussed in a post-mining phase, using additional datasets related to heating and cooling electrical energy needs.

The process can be used to support stakeholders in characterising building behaviour, to define appropriate energy management strategies, and to send timely alerts based on anomaly detection outcomes.

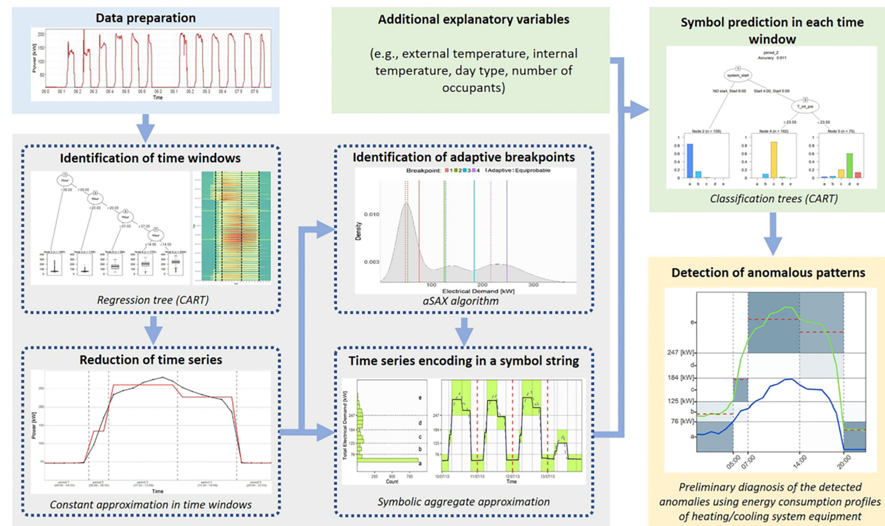


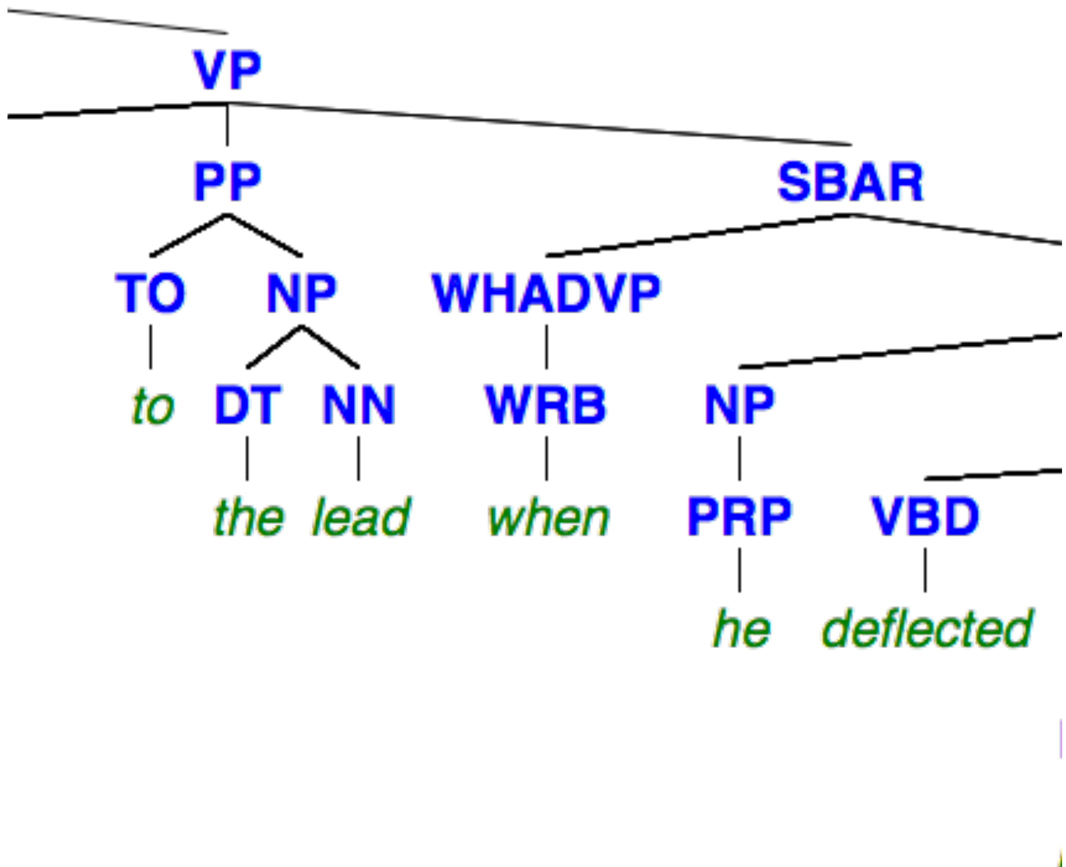
Fig. 2. – Framework for advanced energy consumption characterisation in buildings and anomalous pattern detection.


Spotlight on Text Mining

Semantic Parsing

The process of converting a sentence in a natural language to a **formal meaning representation**.

Word **order** and word **type**/role provide the word's **attributes**.

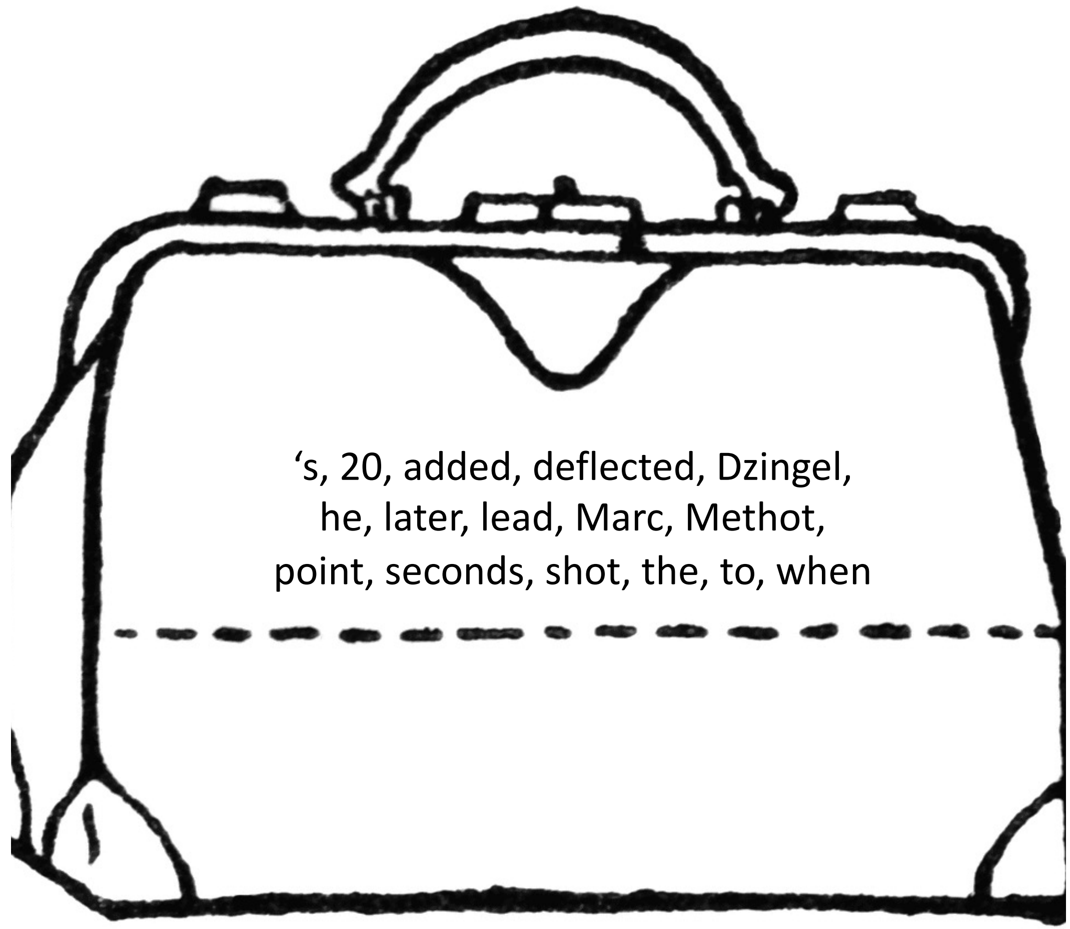


Bag of “Words” (BoW)

Only the **presence** (or **absence**) of “words” (stems, n-grams, sentences, etc.) is important.

Relative **frequencies** provide information (intent, theme, feeling, etc.) about the corpus.

The words **themselves** are attributes of the document.



Text Processing

Text data requires extensive cleaning and processing.

There are a number of challenges due to the nature of the data:

- what is an anomaly in the text?
- what is an outlier?
- are these concepts even definable?
- how do we deal with encoding errors?

Spelling mistakes and typographical errors are difficult to catch in large documents, even with spell-checkers.

Text Processing

The process can be simplified to some extent with the help of **regular expressions** and **text pre-processing functions**.

Specific pre-processing steps vary depending on the problem:

- *tweetish* uses a different vocabulary than *legalese*
- ditto for a child who's learning to speak and a Ph.D. candidate

As is almost everything else related to text mining, the cleaning process is **strongly context-dependent**.

Note that the order of pre-processing tasks can affect results.

Text Processing

- “Dzingel added lead deflected Marc Methot point shot twenty seconds later”

____

added, deflected, Dzingel, later, lead, Marc, Methot, point, seconds, shot, twenty

Text Processing – OPTIONS

Convert all letters to **lower case** (avoid when seeking names)

Remove all **punctuation** marks (avoid if seeking emojis)

Remove all **numerals** (avoid when mining for quantities)

Remove all extraneous **white space**

Remove characters within **brackets** (avoid if seeking tags)

Replace all **numerals with words**

Text Processing – OPTIONS

Replace **abbreviations**

Replace **contractions** (avoid if seeking non-formal speech)

Replace all **symbols with words**

Remove **stop words** and **uninformative words** (language-, era- and context-dependent)

Stem words and **complete stems** to remove empty variation

- “sleepiness”, “sleeping”, “sleeps”, “slept” convey the meaning of “sleep”
- in “operations research”, “operating systems” and “operative dentistry”, the stem “operati” needs to stand it for **different meanings**

Text Processing

- **Phonetic accent representation**
  ya new cah's wicked pissa!
- **Neologisms and portmanteaus**
  I'm planning prevenge?
- **Poor translations/foreign words**
- **Puns and play-on-words**
- **Mark-up, tags, and uninformative text**
  <b>; \includegraphics; ISBN blurb
- **Specialized vocabulary**
  clopen; poset; retro encabulator
- **Fictional names and places**
  Qo'noS; Kilgore Trout
- **Slang and curses**
  skengfire; #$&#!

EXERCISE

How would you process the following bit of text?

“<i>He<i> went to bed at   2 A.M. It\’s way too late! He was only 20% asleep at first, but sleep eventually came.”

Text Representation

Text must be stored to data structures with right properties:

- a **string** or vector of characters, with language-specific encoding
- a **corpus** (collection) of text documents (with meta information)
- a **document-term matrix** (DTM) where the rows are documents, the columns are terms, and the entries are an appropriate text statistic (or the transposed **term-document matrix** (TDM)
- a **tidy text dataset** with one **token** (single word, *n*-gram, sentence, paragraph) per row

No magic recipe: best format depends on the problem at hand. But this step is **crucial**, both for semantic analysis and BoW.

DTM/TDM Representation

	Document 1	Document 2	Document 3	…	Document N
Token 1	0	0	1	62	3
Token 2	0	1	0	61	2
Token 3	1	0	3	101	0
…	112	24	38	84	0
Token M	2	2	0	12	3
Sum	115	27	42	320	8

Sum
66
64
105
258
19

Text Statistics

Consider a corpus $\mathcal{C} = \{d_1, \dots, d_N\}$ consisting of N **documents** and M BoW **terms** $\mathcal{C} = \{t_1, \dots, t_M\}$.

For instance, if

$$\mathcal{C} = \left\{ \begin{array}{c} \text{“the dogs who have been let out”,} \\ \text{“who did that”,} \\ \text{“my dogs breath smells like dogs food”} \end{array} \right\},$$

then

$N = 3, d_1 =$ “the dogs who have been let out”,

$d_2 =$ “who did that”, $d_3 =$ “my dogs breath smells like dogs food”

Text Statistics

The **relative term frequency** of t in d is

$$tf^*_{t,d} = \frac{\text{\# of times } t \text{ occurs in } d}{M_d}$$

$tf^*_{t,d}$		t													
		1 been	2 breath	3 did	4 dogs	5 food	6 have	7 let	8 like	9 my	10 out	11 smells	12 that	13 the	14 who
d	1	1/7	0	0	1/7	0	1/7	1/7	0	0	1/7	0	0	1/7	1/7
	2	0	0	1/3	0	0	0	0	0	0	0	0	1/3	0	1/3
	3	0	1/7	0	2/7	1/7	0	0	1/7	1/7	0	1/7	0	0	0

Text Statistics

The **relative document frequency** of t is

$$df_t^* = \frac{\text{\# of documents in which } t \text{ occurs}}{N} = \frac{\sum_d \operatorname{sign}(tf_{t,d}^*)}{N}$$

df_t^*	t													
	1 been	2 breath	3 did	4 dogs	5 food	6 have	7 let	8 like	9 my	10 out	11 smells	12 that	13 the	14 who
	1/3	1/3	1/3	2/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	2/3

Text Statistics

The **term frequency – inverse document frequency** of t in d is

$$tf\text{–}idf^*_{t,d} = -tf^*_{t,d} \times \ln(df^*_t)$$

(coloured values are high scoring for a particular document (e.g. red = document 1)

$tf\text{–}idf^*_t$		t													
		1 been	2 breath	3 did	4 dogs	5 food	6 have	7 let	8 like	9 my	10 out	11 smells	12 that	13 the	14 who
d	1	0.16	0	0	0.06	0	0.16	0.16	0	0	0.16	0	0	0.16	0.06
	2	0	0	0.37	0	0	0	0	0	0	0	0	0.37	0	0.14
	3	0	0.16	0	0.12	0.16	0	0	0.16	0.16	0	0.16	0	0	0

Text Statistics

If **all the documents** contain the term t, then $\mathrm{df}_t^* = 1$ and

$$tf\text{–}idf_{t,d}^* = -tf_{t,d}^* \times \ln(1) = 0$$

(that terms does not provide information)

If a term t **rarely occurs** in a document d, then $tf_{t,d}^* \approx 0$ and

$$tf\text{–}idf_{t,d}^* \approx -0 \times \ln(df_t^*) \approx 0.$$

Terms that appear relatively often only in a small subset of documents are crucial to understanding those documents **in the general context** of the corpus.

DISCUSSION

- At the analysis stage, it is easy to forget where the data comes from and what it really applies to.
- Text comes unstructured and unorganized. After processing, text is clean, but still unstructured. Bag of Words provides a framework for a structured numerical representation of text.
- How does this affect the choice of text statistic in the DTM/TDM?

DEMO WITH TEXT DATA

Concluding Remarks on Machine Learning

- The number of available techniques in machine learning is growing all the time
- The key is figuring out how to apply these techniques to your particular problems or questions
- The key to this in turn is:
  - Having a very strong understanding of the business situation
  - Having a decent understanding of the functionality and appropriate (and inappropriate) applications of these techniques
  - Having the creativity to connect the one to the other