DATA ANALYSIS UNIVERSALS

"Reports that say that something hasn't happened are always interesting to me, because as we know, there are **known knowns**; there are things we know that we know. There are **known unknowns**; that is to say, there are things that we now know we don't know. But there are also **unknown unknowns** – there are things we do not know we don't know."

Donald Rumsfeld, US Department of Defense News Briefing, 2002

OUTLINE

1. Data, M.L., and A.I. in the News
2. Data 101 – Basic Data Concepts
3. Some Practical Definitions
4. Workflows and Pipelines – the Process of Working with Data
5. Models and Systems Thinking
6. Ethical Considerations and Best Practices

DATA, MACHINE LEARNING, AND ARTIFICIAL INTELLIGENCE IN THE NEWS

DATA ANALYSIS UNIVERSALS

MODULE LEARNING OBJECTIVES

Increase awareness of the growing role of Data Science, Machine Learning, and A.I. in society, in different domains.

Increase awareness of the possible functionality/capabilities of these technologies.

Increase awareness of some of the social issues coming out of this growing role of these technologies.

News

Robots are better than doctors at diagnosing some cancers, major study finds

share

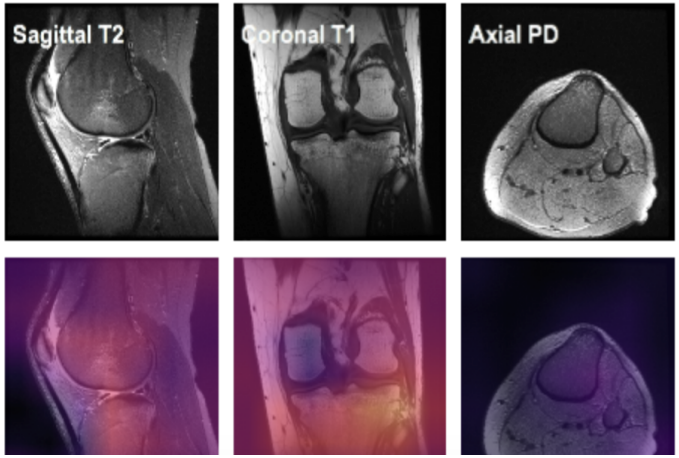
Save 7

Stanford ML Group

MRNet: Deep-learning-assisted diagnosis for knee magnetic resonance imaging

Nicholas Bien *, Pranav Rajpurkar *, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng†, Matthew P. Lungren†

We developed an algorithm to predict abnormalities in knee MRI exams, and measured the clinical utility of providing the algorithm's predictions to radiologists and surgeons during interpretation.



Magnetic resonance (MR) imaging of the knee is the standard of care imaging modality to evaluate knee

Google AI Claims 99 Percent Accuracy In Metastatic Breast Cancer Detection

34

Posted by **BeauHD** on Friday October 12, 2018 @08:00PM from the promising-solutions dept.

Researchers at the Naval Medical Center San Diego and Google AI, a division within Google dedicated to artificial intelligence research, are using cancer-detecting algorithms to detect metastatic tumors by autonomously evaluating lymph node biopsies. VentureBeat reports:

> Their AI system -- dubbed Lymph Node Assistant, or LYNA -- is described in a paper titled "Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection," published in The American Journal of Surgical Pathology. In tests, it achieved an area under the receiver operating characteristic (AUC) -- a measure of detection accuracy -- of 99 percent. That's superior to human pathologists, who according to one recent assessment miss small metastases on individual slides as much as 62 percent of the time when under time constraints. LYNA is based on Inception-v3, an open source image recognition deep learning model that's been shown to achieve greater than 78.1 percent accuracy on Stanford's ImageNet dataset. As the researchers explained, it takes as input a 299-pixel image (Inception-v3's default input size), outlines tumors at the pixel level, and, in the course of training, extracts labels -- i.e., predictions -- of the tissue patch ("benign" or "tumor") and adjusts the model's algorithmic weights to reduce error.
>
> In tests, LYNA achieved 99.3 percent slide-level accuracy. When the model's sensitivity threshold was adjusted to detect all tumors on every slide, it exhibited 69 percent sensitivity, accurately identifying all 40 metastases in the evaluation dataset without any false positives. Moreover, it was unaffected by artifacts in the slides such as air bubbles, poor processing, hemorrhage, and overstaining. LYNA wasn't perfect -- it occasionally misidentified giant cells, germinal cancers, and bone marrow-derived white blood cells known as histiocytes -- but managed to perform better than a practicing pathologist tasked with evaluating the same slides. And in a second paper published by Google AI and Verily, Google parent company Alphabet's life sciences subsidiary, the model halved the amount of time it took for a six-person team of board-certified pathologists to detect metastases in lymph nodes.

Data scientists find connections between birth month and health

Date: June 8, 2015

Source: Columbia University Medical Center

Summary: Scientists have developed a computational method to investigate the relationship between birth month and disease risk. The researchers used this algorithm to examine New York City medical databases and found 55 diseases that correlated with the season of birth. Overall, the study indicated people born in May had the lowest disease risk, and those born in October the highest.

Share:

9
Oct
2018

Scientists Using GPS Tracking on Endangered Dhole Wild Dogs

Researchers Successfully Tag a Dhole. Wildlife scientists around the globe are ecstatic to hear that researchers were able to successfully place a GPS tracking device onto a dhole. This marks the first time in history that conservationists have been able to place a collar on one of these very rare Indian wild dogs. It's estimated that less than 2,500 of these creatures still exist globally.

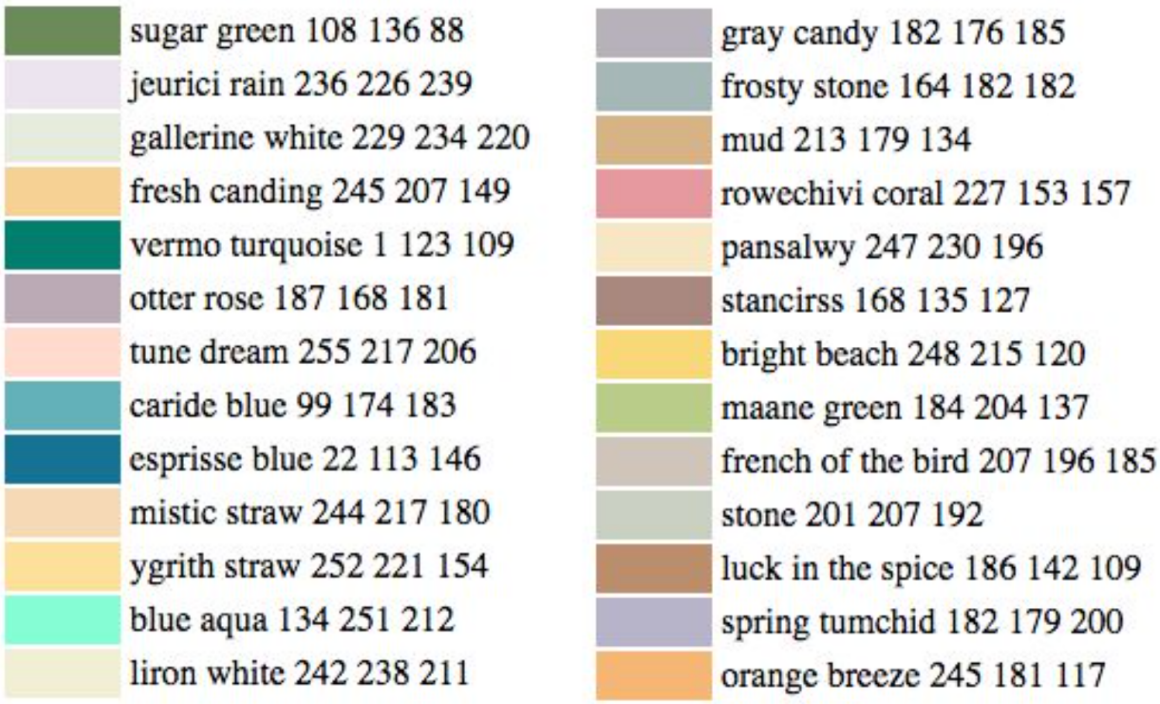
DESIGN

These AI-invented paint color names are so bad they're good

1

What's in a (paint) name?

By **Sam Reichman** | May 31, 2017, 5:12pm EDT



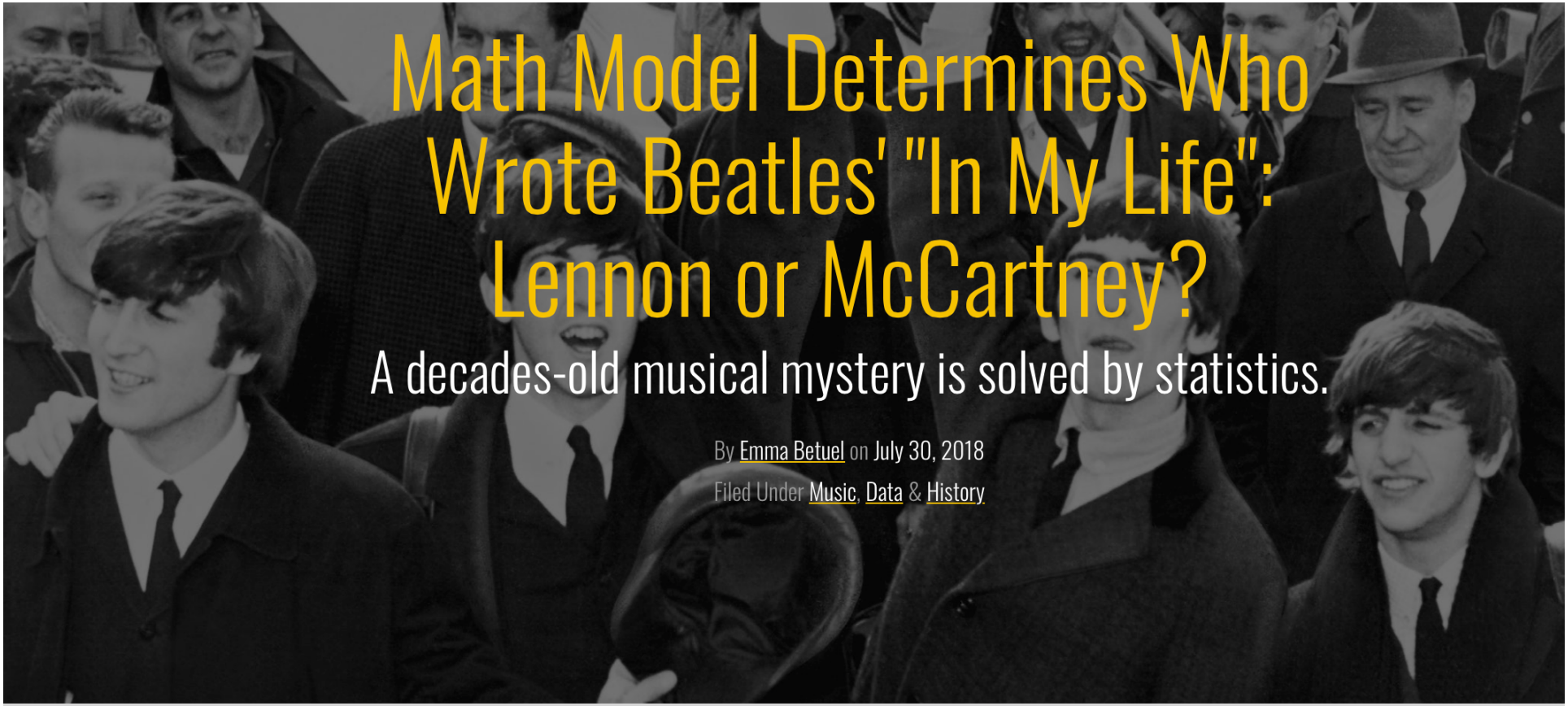
Intelligent Machines

We tried teaching an AI to write Christmas movie plots. Hilarity ensued. Eventually.

Using a neural network to create ridiculous plot lines takes a lot of work—and reveals the challenges of generating human language.

by Karen Hao December 21, 2018

MR. TECH



Math Model Determines Who Wrote Beatles' "In My Life": Lennon or McCartney?

A decades-old musical mystery is solved by statistics.

By Emma Betuel on July 30, 2018

Filed Under Music, Data & History

Scientists use Instagram data to forecast top models at New York Fashion Week

Method is 80 percent accurate in identifying most popular models for the following season

Date: September 3, 2015

Source: Indiana University

Summary: Researchers have predicted the popularity of new faces to the world of fashion modeling with over 80 percent accuracy using advanced computational methods and data from Instagram.

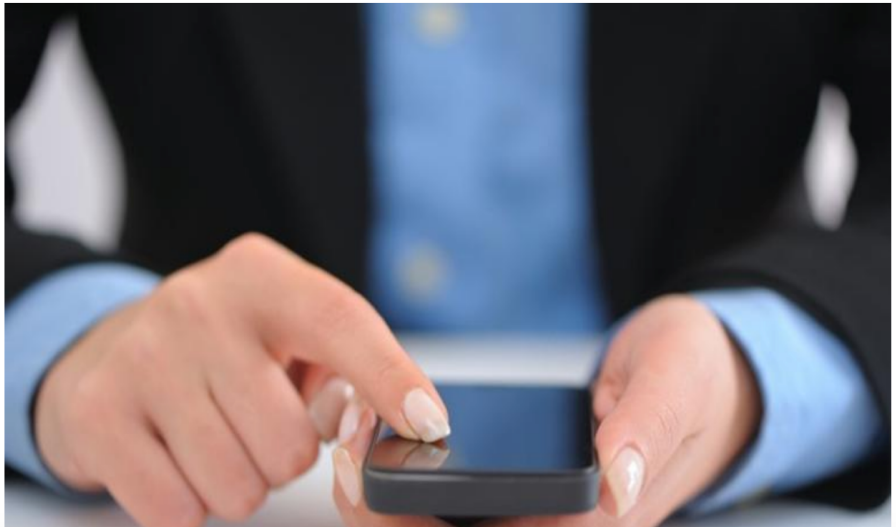
Share:

How big data will solve your email problem

That deluge in your inbox needs to be handled. A team of Israeli researchers thinks big data has some answers that can help.

By Jason Hiner | October 2, 2013 -- 16:05 GMT (09:05 PDT) | Topic: Going Deep on Big Data

6



NEWSLETTERS

ZDNet Big Data

Keep up with the latest developments in extracting maximum information value for today's business.

Your email address SUBSCRIBE

SEE ALL

MORE RESOURCES

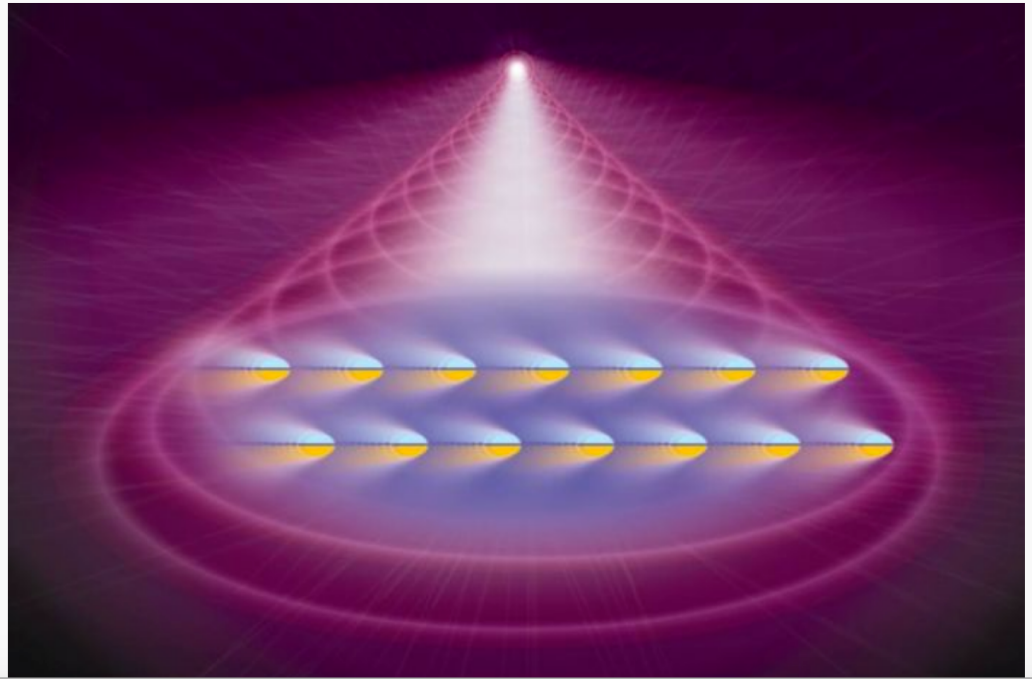
Special report: From cloud

Artificial intelligence better than physicists at designing quantum science experiments

Share on Facebook Share on Twitter

ABC Science By science reporter Belinda Smith

Posted 19 October 2018 at 3:36 pm



Wonkblog • Analysis

This researcher studied 400,000 knitters and discovered what turns a hobby into a business

Most Read Business

1 **Perspective**
I ordered a box of crickets from the Internet and it went about as well as you'd expect

2 As a grocery chain is dismantled, investors recover their money. Worker pensions are short millions.

3 Markets poised to finish year with worst performance in a decade — and the volatility seems certain to continue

SCIENCE

Wait, Have We Really Wiped Out 60 Percent of Animals?

The findings of a major new report have been widely mischaracterized —although the actual news is still grim.

ED YONG OCT 31, 2018

MORE STORIES

Animals Are Riding an Escalator to Extinction
ED YONG

It Will Take Millions of Years for Mammals to Recover From Us
ED YONG

In a Few Centuries, Cows Could Be the Largest Land Animals Left
ED YONG

An Ancient Tradition

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 2 DAYS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin
8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Facebook documents seized by MPs investigating privacy breach

25 November 2018 | Business

A cache of Facebook documents has been seized by MPs investigating the Cambridge Analytica data scandal.

The New York Times

Firm Led by Google Veterans Uses A.I. to 'Nudge' Workers Toward Happiness

At Netflix, Who Wins When It's Hollywood vs. the Algorithm?

As the company plunges deeper into originals, its L.A. wing is doing the once-unthinkable: overriding the metrics

The cast of Netflix original show 'GLOW.' NETFLIX

By Shalini Ramachandran and Joe Flint

100 COMMENTS

Nov. 10, 2018 12:00 a.m. ET

Netflix Inc.'s executives were torn. On the one hand they trusted the company's algorithm. On the other they were worried about ticking off Jane Fonda.

After the streaming-video giant released the second season of the comedy "Grace and Frankie" in 2016, its product team put up an image to promote the show to U.S. subscribers that only included Ms. Fonda's co-star, Lily Tomlin. Tests showed that more users clicked on the show when the photo didn't include Ms. Fonda.

DATA 101 – BASIC DATA CONCEPTS

DATA ANALYSIS UNIVERSALS

"You can have data without information, but you cannot have information without data."

Daniel Keys Moran (attributed)

MODULE LEARNING OBJECTIVES

Preliminary familiarity with the following concepts:

- data, attribute (property, factor, variable)
- predictive models, explanatory models
- classification, class probability estimation, clustering, association rules, time series analysis, anomaly detection, decision tree, supervised learning, unsupervised learning

Compare and contrast: data science vs analytics (Business Intelligence).

Awareness of appropriate levels of trust in models.

WHAT IS DATA? WHERE DOES IT COME FROM?

4,529 ‘red’ 25.782 ‘Y’

OBJECTS AND ATTRIBUTES

Object: apple

Shape: spherical

Colour: red

Function: food

Location: fridge

Owner: Jen

Remember: a person or an object is not simply the sum of its attributes!

FROM ATTRIBUTES TO DATASETS

Attributes are **fields** (or columns) in a database; objects are **instances** (or rows)

Objects are described by their **feature vector**, the collection of attributes associated with value(s) of interest

ID#	Shape	Colour	Function	Location	Owner
1	spherical	red	food	fridge	Jen
2	rectangle	brown	food	office	Pat
3	round	white	tell time	lounge	School
...	...	...	...	...	...

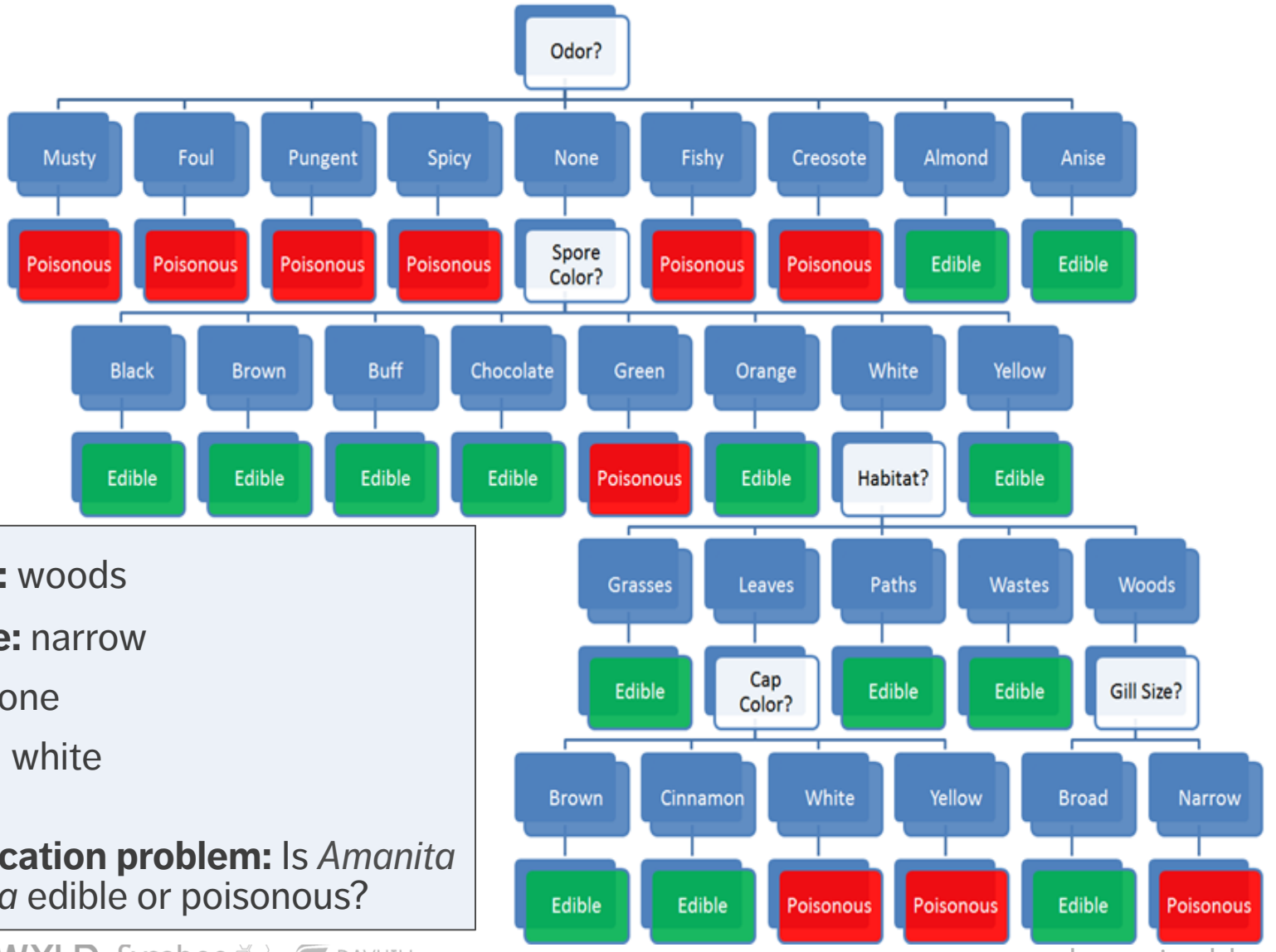
[https://archive.ics.uci.edu/ml/datasets/Mushroom]

POISONOUS MUSHROOMS DATASET

Amanita muscaria

Habitat: woods
Gill Size: narrow
Odor: none
Spores: white

Classification problem: Is *Amanita muscaria* edible, or poisonous?



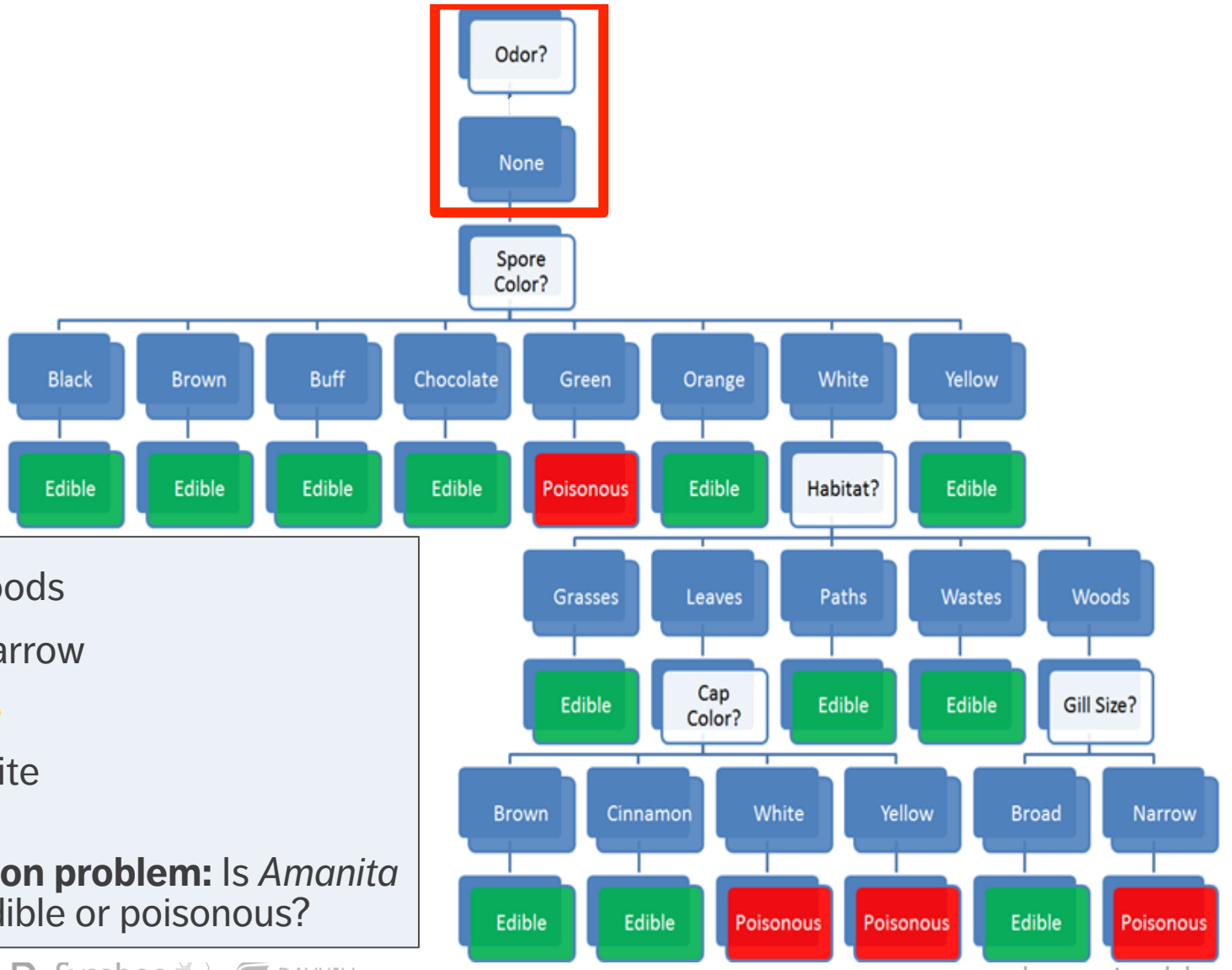
Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* edible or poisonous?



Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* edible or poisonous?

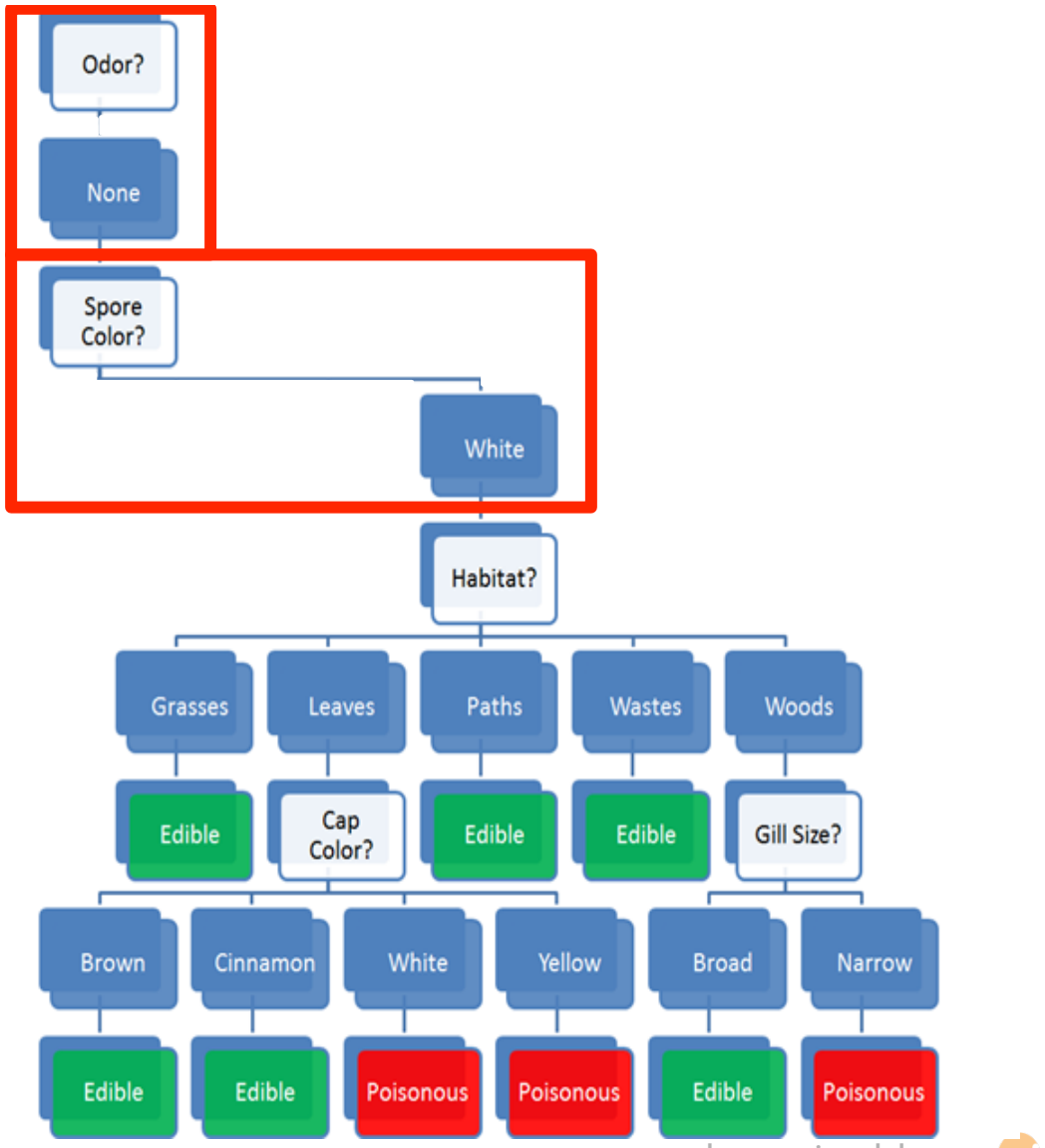
Habitat: woods

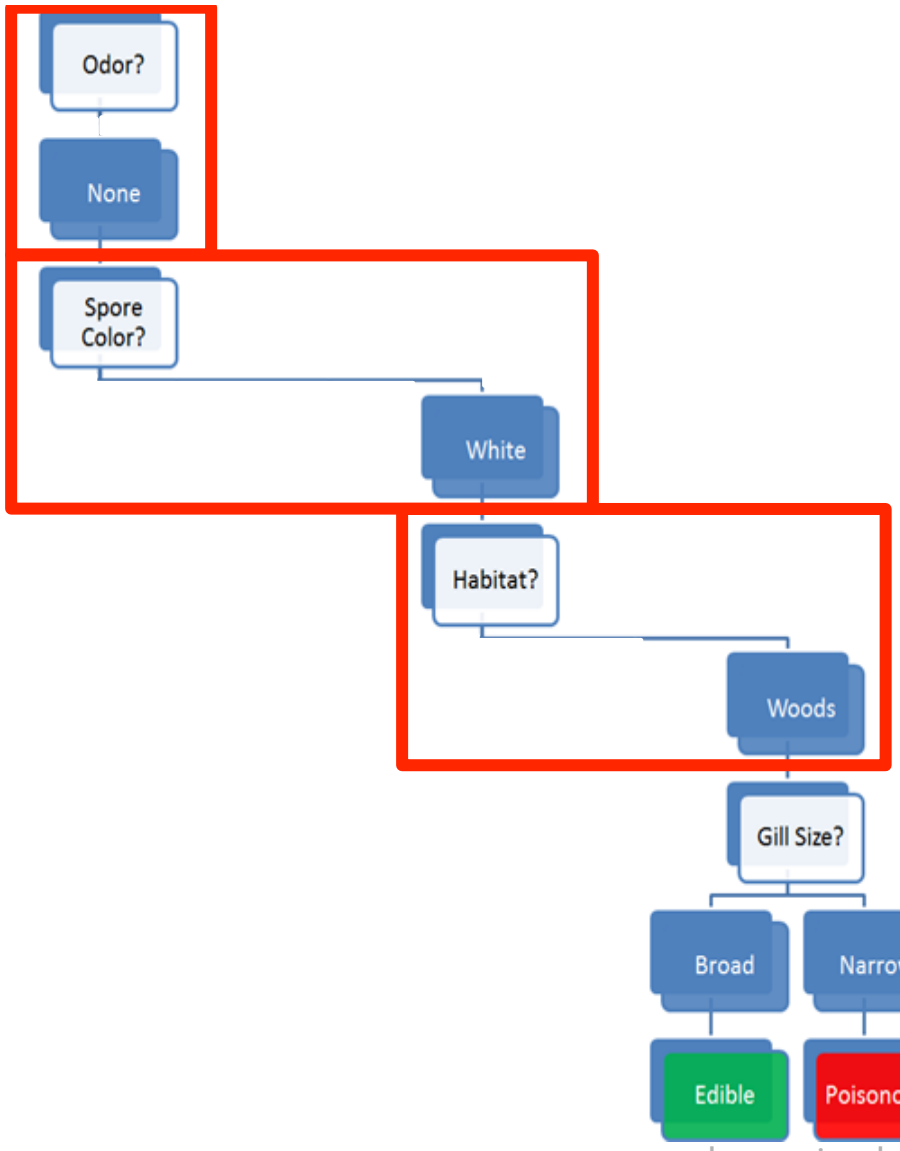
Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* edible or poisonous?





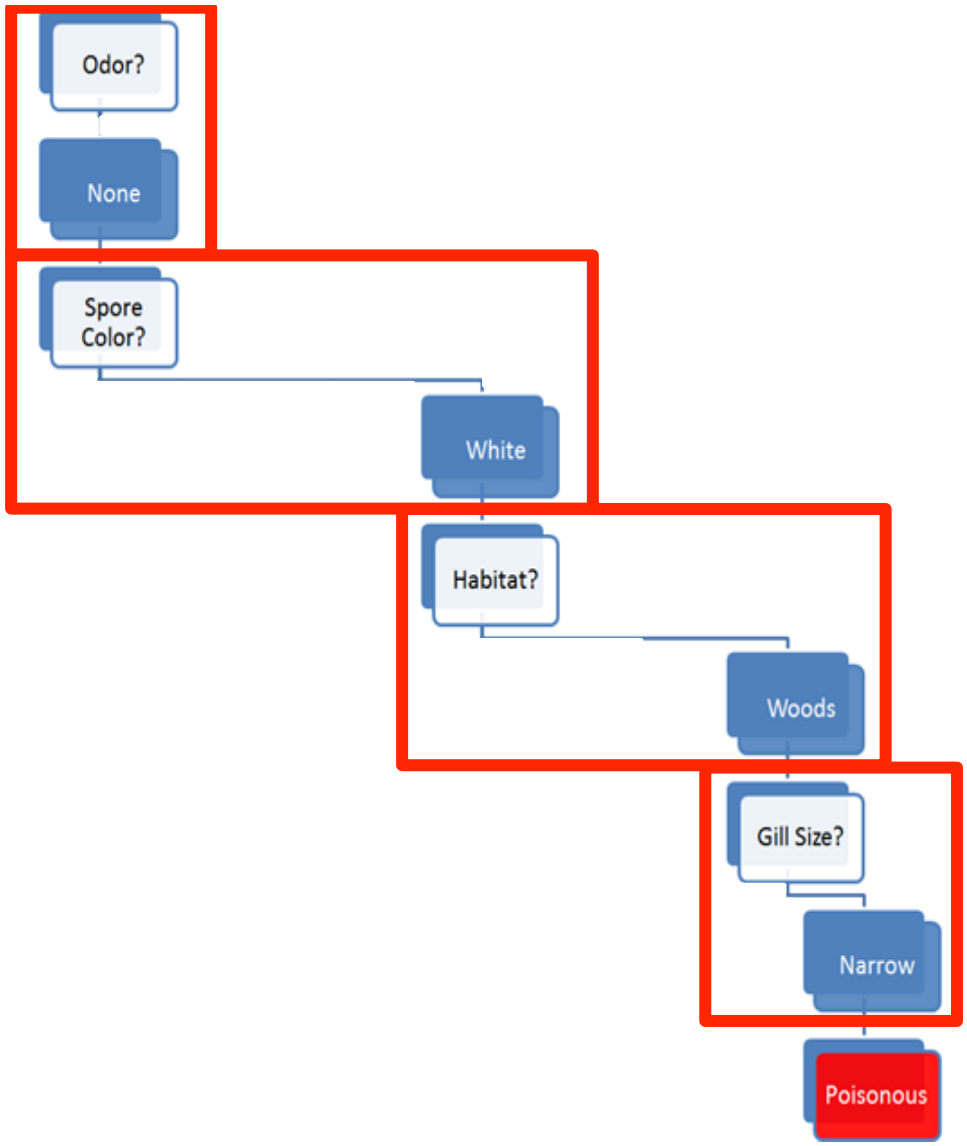
Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* edible or poisonous?



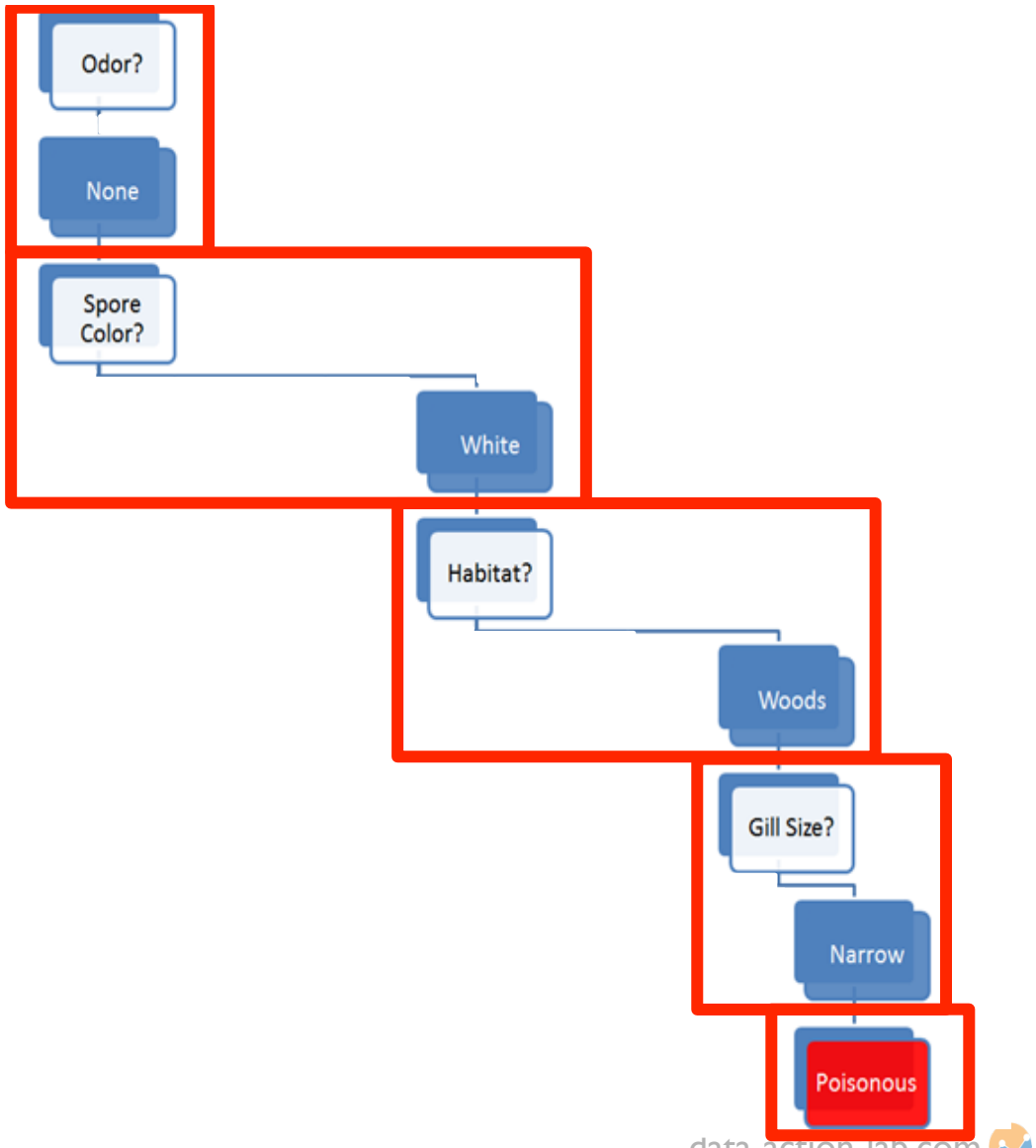
Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* edible or poisonous?



Habitat: woods

Gill Size: narrow

Odor: none

Spores: white

Classification problem: Is *Amanita muscaria* edible or **poisonous**?

DISCUSSION

Would you trust an "**edible**" prediction?

Where is the model coming from?

What would you need to know to trust the model?

What's the cost of making a classification mistake, in this case?

ASKING THE RIGHT QUESTIONS

Data science is really about asking and answering questions:

- **Analytics:** "How many clicks did this link get?"
- **Data Science:** "Based on this user's previous purchasing history, can I predict what links they will click on the next time they access the site?"

Data mining/science models are usually **predictive** (not **explanatory**): they show connections, but don't reveal why these exist.

Warning: not every situation calls for data science, artificial intelligence, machine learning, or analytics.

DATA SCIENCE/MACHINE LEARNING/A.I. TASKS

Classification and **class probability estimation:** which clients are likely to be repeat customers?

Clustering: do customers form natural groups?

Association rule discovery: what books are commonly purchased together?

Others:

profiling and behaviour description; **link prediction**; **value estimation** (how much is a client likely to spend in a restaurant); **similarity matching** (which prospective clients are similar to a company's best clients?); **data reduction**; **influence/causal modeling**, etc.

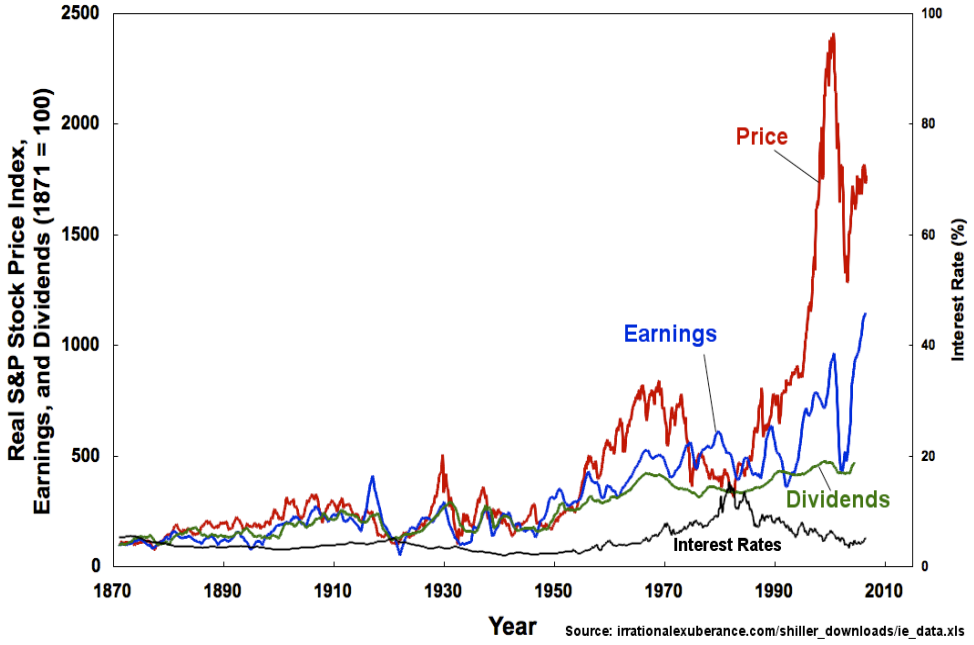
TIME SERIES ANALYSIS

A simple time series:

- Has two variables: time + 2nd variable
- The second variable is *sequential*

What is the pattern of behaviour of this second variable over time? Relative to other variables?

Can we use this information to forecast the behaviour of the variable in the future?



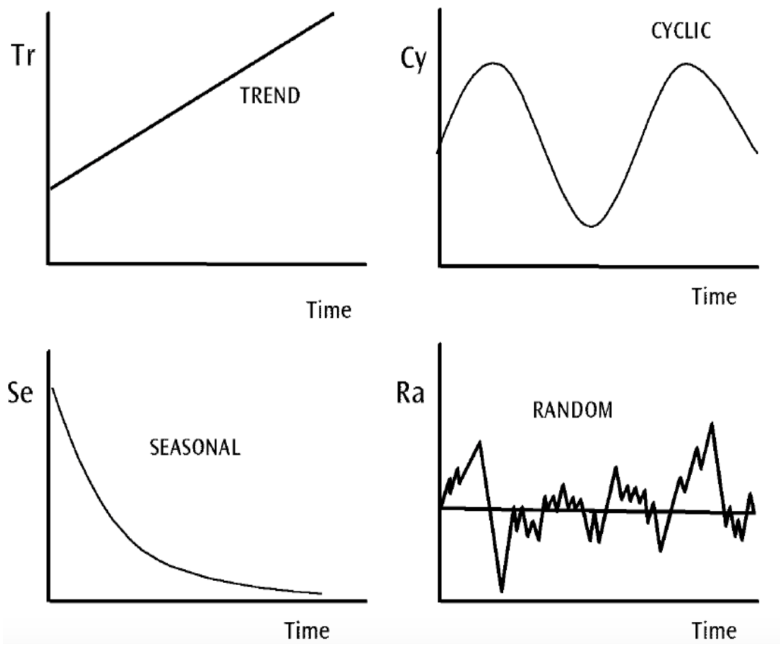
TEMPORAL PATTERNS

The goal here is our familiar analysis goals:

- find patterns in the data
- create a (mathematical) model that captures the essence of these patterns

The patterns can be quite complex – some fancy analysis typically required!

In particular – the overall series can often be broken down into being made up of multiple **component models**. There are software libraries that can help!



TIME SERIES CASE STUDIES

A Time-Series Analysis of International Public Relations Expenditure and Economic Outcome

Communication Research
2018, Vol. 45(7) 1012–1030

DOI: 10.1177/0093650215581370
journals.sagepub.com/home/crx
SAGE

Suman Lee[1] and Byungwook Kim[2]

Abstract

This study tested a causal relationship between international public relations (PR) expenditure and its economic outcome at the country level by using a time-series analysis. International PR expenditures of four client countries (Japan, Colombia, Belgium, and the Philippines) were collected from the semi-annual reports of the Foreign Agency Registration Act (FARA) from 1996 to 2009. Economic outcome was measured by U.S. imports from the client countries and U.S. foreign direct investment (FDI) toward them. This study found that the past PR expenditure holds power in forecasting future economic outcomes for Japan, Belgium, and the Philippines except Colombia.

Keywords

international public relations, PR return on investment, bottom-line effect, time-series analysis, Granger causality test

RESEARCH ARTICLE

Seiya MAKI, Shuichi ASHINA, Minoru FUJII, Tsuyoshi FUJITA, Norio YABE, Kenji UCHIDA, Gito GINTING, Rizaldi BOER, Remi CHANDRAN

Employing electricity-consumption monitoring systems and integrative time-series analysis models: A case study in Bogor, Indonesia

Abstract The Paris Agreement calls for maintaining a global temperature less than 2°C above the pre-industrial level and pursuing efforts to limit the temperature increase even further to 1.5°C. To realize this objective and promote a low-carbon society, and because energy production and use is the largest source of global greenhouse-gas (GHG) emissions, it is important to efficiently manage energy demand and supply systems. This, in turn, requires theoretical and practical research and innovation in smart energy monitoring technologies, the identification of appropriate methods for detailed time-series analysis, and the application of these technologies at urban and national scales. Further, because developing countries contribute increasing shares of domestic energy consumption, it is important to consider the application of such innovations in these areas. Motivated by the mandates set out in global agreements on climate change and low-carbon societies, this paper focuses on the development of a smart energy monitoring system (SEMS) and its deployment in households and public and commercial sectors in Bogor, Indonesia. An electricity demand prediction model is developed for each device using the Auto-Regression eXogenous model. The real-time SEMS data and time-series clustering to explore similarities in electricity consumption patterns between monitored units, such as residential, public, and commercial buildings, in Bogor is, then, used. These clusters are evaluated using peak demand and Ramadan term characteristics. The resulting energy-prediction models can be used for low-carbon planning.

Keywords electricity monitoring, electricity demand prediction, multiple-variable time-series modeling, time-series cluster analysis, Indonesia

Received Dec. 30, 2017; accepted Mar. 28, 2018; online May 30, 2018

1 Introduction

1.1 Background and objectives

To attain a low-carbon society, it is necessary to transform the centralized energy system into distributed systems at city and regional scales. Because energy demand patterns vary spatially, more detailed data on energy demand provided by innovative Information Communication Technologies (ICTs) is expected to enable local energy demand and supply system optimization in which distributed renewable energy resources can be integrated with large-scale grid energy supply systems.

Energy information and data at local scales, particularly in developing countries, is persistently unavailable. However, there is enormous potential to reduce energy use in various sectors through the use of rapidly developing ICT systems in energy management. The

MAKI ET AL: DATA ANALYSIS SYSTEM

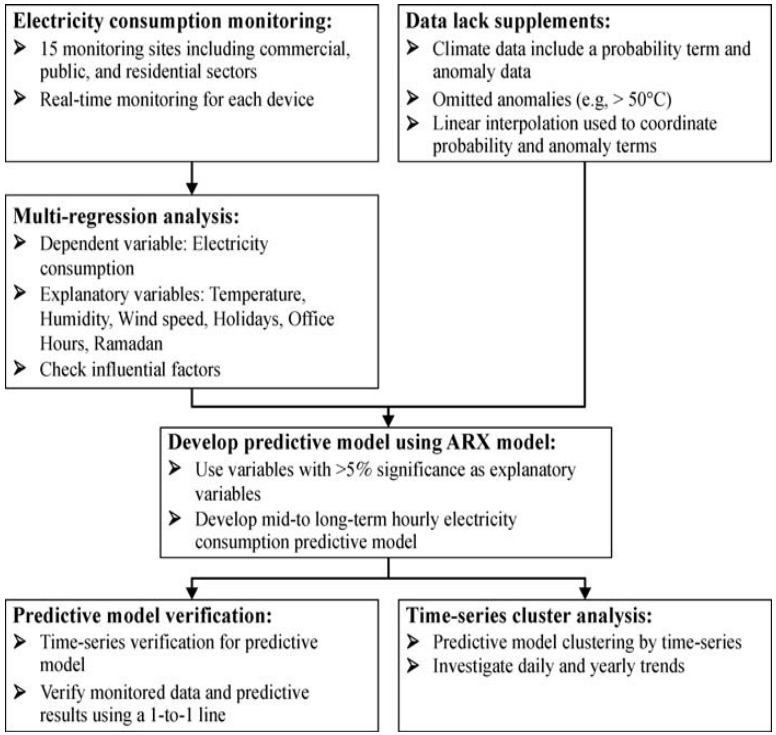


Fig. 1 Analytical procedure used in this paper

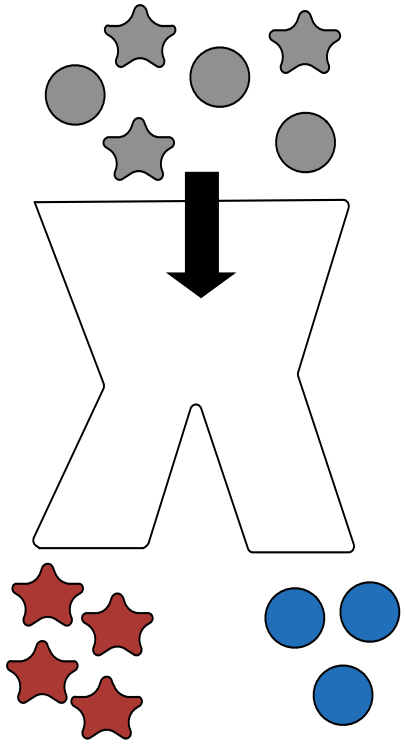
CLASSIFICATION

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!



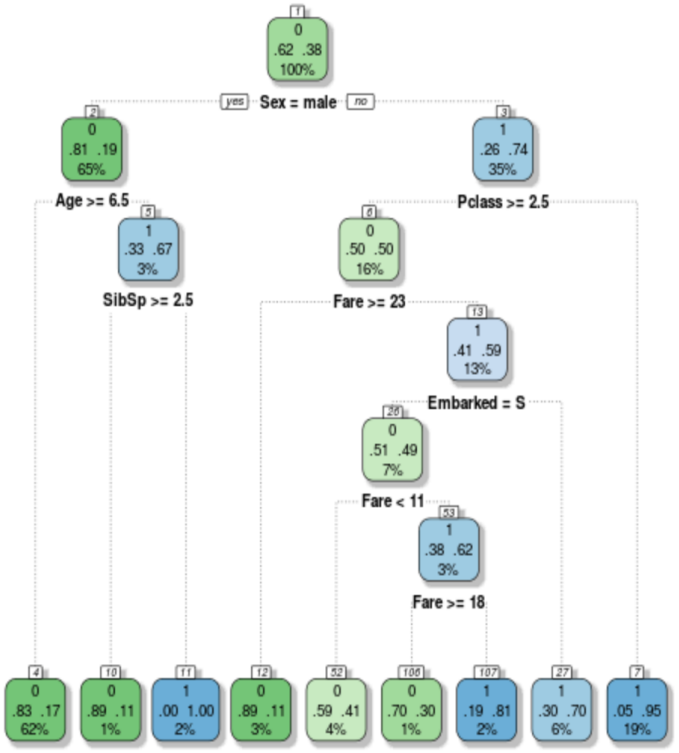
DECISION TREE CLASSIFIERS

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



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.

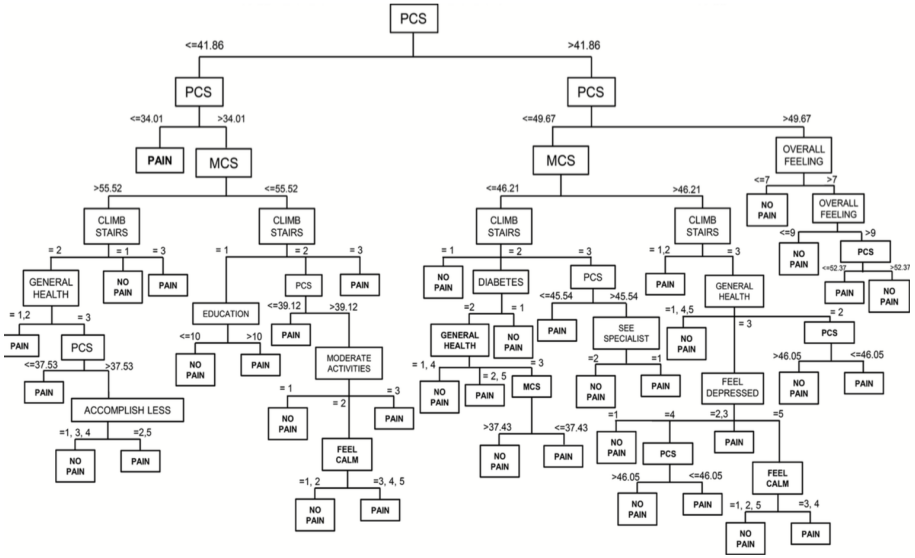
■ **Abstract**

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.



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

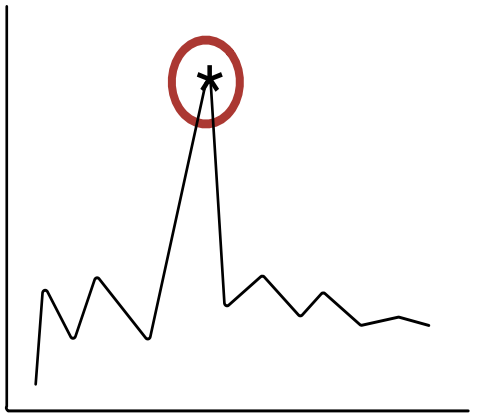
ANOMALY DETECTION

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

ARTICLE INFO

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

ABSTRACT

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.

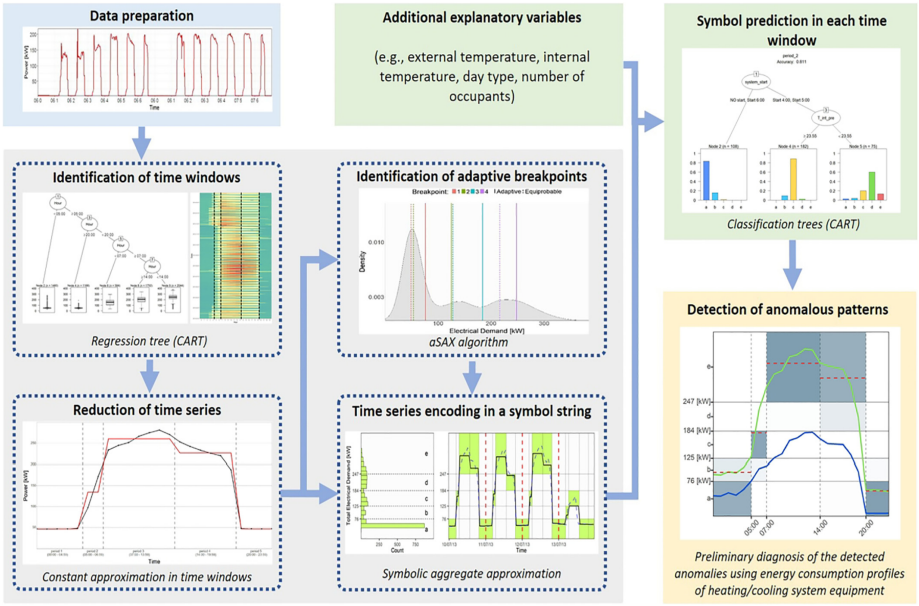


Fig. 2. – Framework for advanced energy consumption characterisation in buildings and anomalous pattern detection.

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)

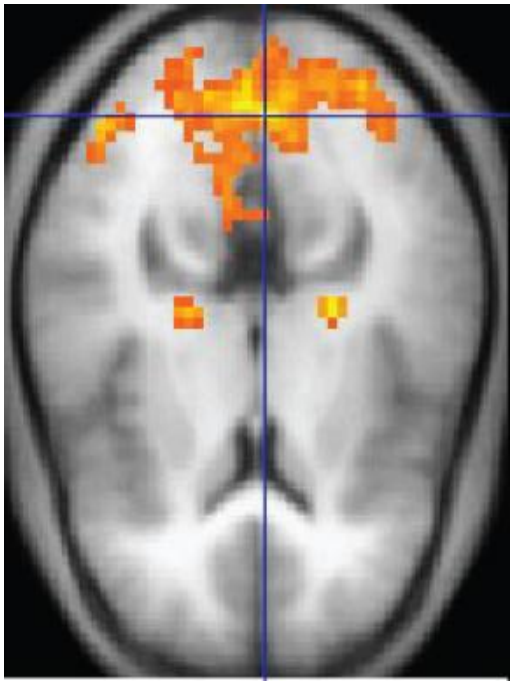
CLUSTERING CASE STUDY

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.



FMRI highlighting some areas of the pre-frontal cortex.

SOME PRACTICAL DEFINITIONS

DATA ANALYSIS UNIVERSALS

"What's in a name? That which we call a rose
By any other name would smell as sweet."

W. Shakespeare, Romeo and Juliet, Act II, Scene 2

MODULE LEARNING OBJECTIVES

Preliminary familiarity with the following concepts:

- data analysis
- data science
- machine learning
- patterns
- system
- artificial intelligence
- augmented intelligence

WHAT IS DATA ANALYSIS?

Finding **patterns** in data

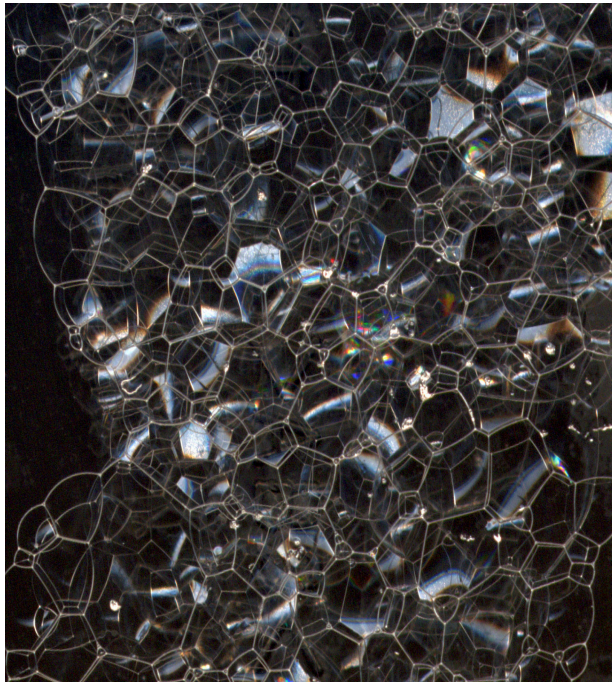
Using data to do something (answer a question, help decision-making, predict the future, draw a conclusion)

Creating models of your data

Describing or explaining your situation (your **system**)

(Testing (scientific) hypotheses?)

(Carrying out calculations on data?)



The more complicated the pattern, the more complicated the analysis

WHAT IS DATA SCIENCE?

Data science is the collection of processes by which we extract useful and **actionable insights** from data.

T. Kwartler (paraphrased)

Data science is the **working intersection** of statistics, engineering, computer science, domain expertise, and "hacking." It involves two main thrusts: **analytics** (counting things) and **inventing new techniques** to draw insights from data.

H. Mason (paraphrased)

WHAT IS MACHINE LEARNING?

Starting around the 1940s researchers began in earnest to teach machines how to learn

The goal of **machine learning** was to create machines that could learn and adapt and respond to novel situations

A wide variety of techniques, accompanied by a great deal of theoretical underpinning, was created in an effort to achieve this goal

WHAT IS ARTIFICIAL/AUGMENTED INTELLIGENCE?

Artificial Intelligence (A.I.) is non-human intelligence that has been engineered rather than one that has evolved naturally.

Artificial intelligence research is research carried out in pursuit of this goal.

Pragmatically speaking, A.I. is “computers carrying out tasks that only humans can usually do”.

Augmented Intelligence is human intelligence that is supported or enhanced by machine intelligence.

WORKFLOWS AND PIPELINES

DATA ANALYSIS UNIVERSALS

"All models are wrong. Some models are useful."

George Box

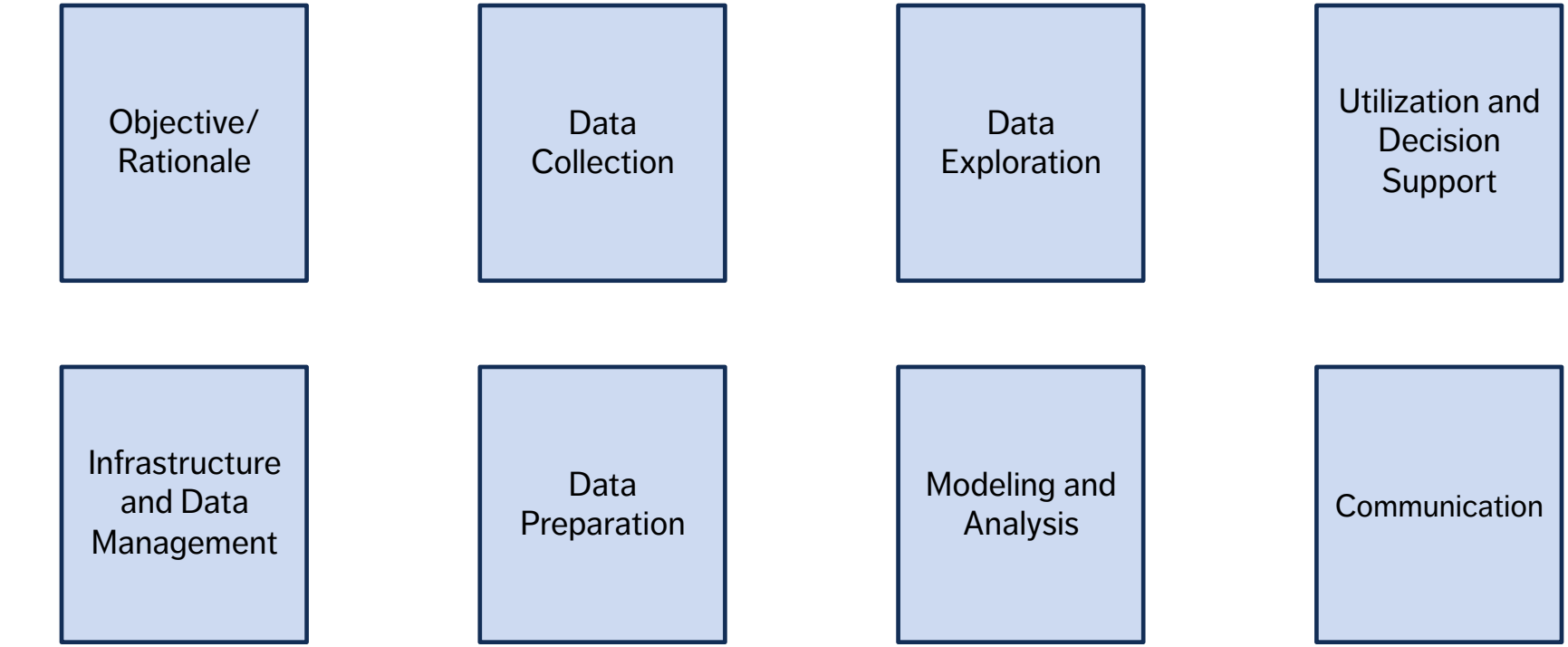
MODULE LEARNING OBJECTIVES

Preliminary familiarity with the following concepts:

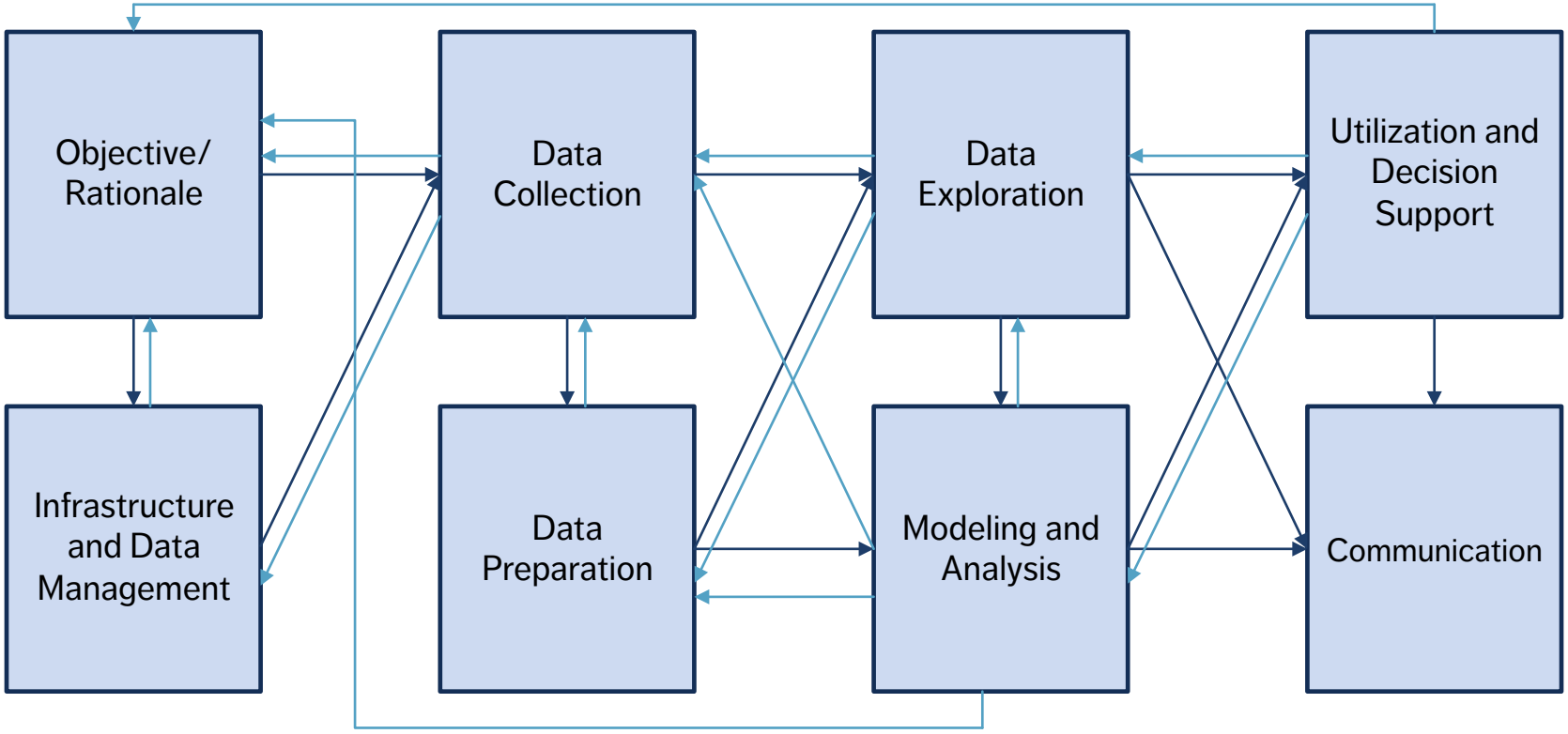
- workflow and components (data collection, data exploration, etc.)
- analytical model
- data mining
- analytic decay
- data science ecosystem
- data science teams

Awareness of the non-linearity of the data analytical process.

THE DATA "WORKFLOW"



THE DATA SCIENCE "WORKFLOW"



THE DATA ANALYSIS PROCESS

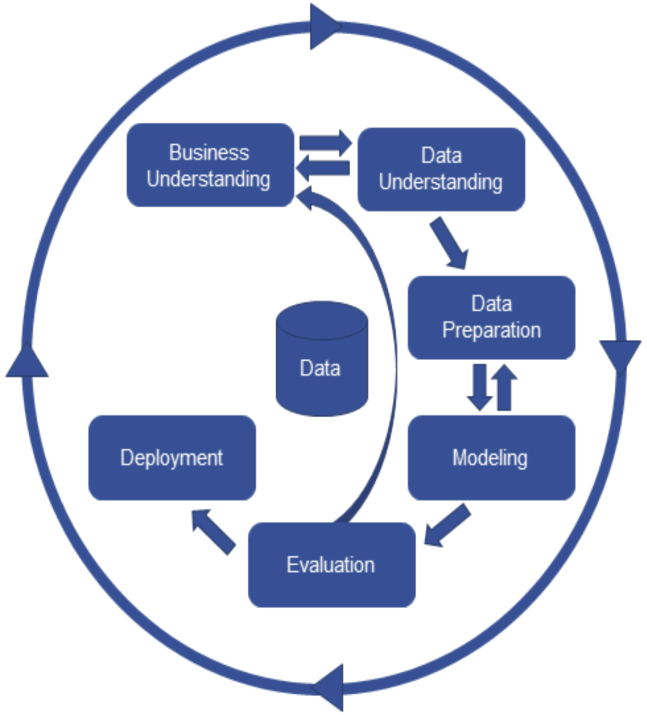
A **large number of analytical models** have to be generated before a final selection can be made.

Iterative process: feature selection and data reduction may require numerous visits to domain experts before models start yielding promising results.

Domain-specific knowledge has to be integrated in the models in order to beat random classifiers and clustering schemes, **on average**.

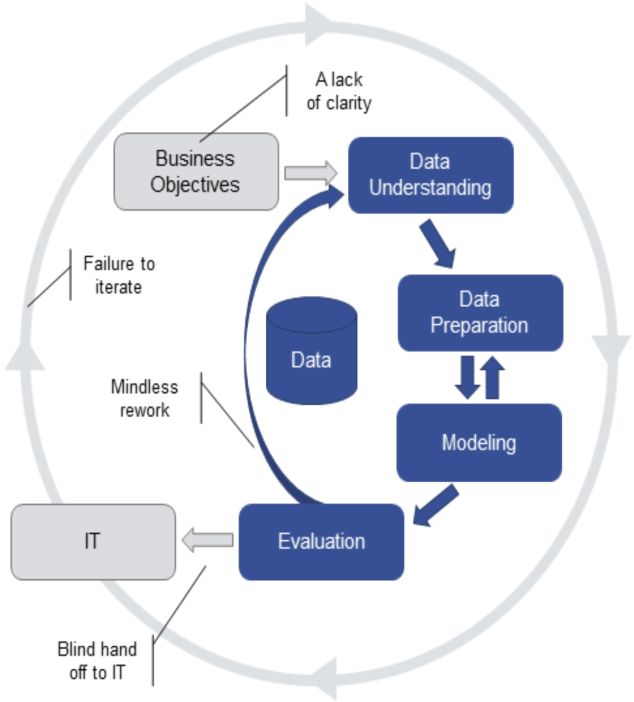
[James Taylor]

CROSS INDUSTRY STANDARD PROCESS, DATA MINING



[James Taylor]

CROSS INDUSTRY STANDARD PROCESS, DATA MINING



LIFE AFTER ANALYSIS

When an analysis or model is 'released into the wild', it can take on a life of its own.

Analysts may eventually have to relinquish control over dissemination. Results may be misappropriated, misunderstood, or shelved. What can the analyst do to prevent this?

Finally, because of **analytic decay**, it's important to view the last analytical step NOT as a static dead end, but rather as an invitation to return to the beginning of the process.

DATA SCIENCE ECOSYSTEM

Data analysis is a **team sport**, with team members needing a good understanding of both **data** and **context**

- data management
- data preparation
- analysis
- communications

Even slight improvements over a current approach can find a useful place in an organization – **data science is not solely about Big Data and disruption!**

MODELS AND SYSTEMS THINKING

DATA ANALYSIS UNIVERSALS

"What if the only valid model of the Universe is the Universe itself?"

Unknown

MODULE LEARNING OBJECTIVES

Preliminary familiarity with the following concepts:

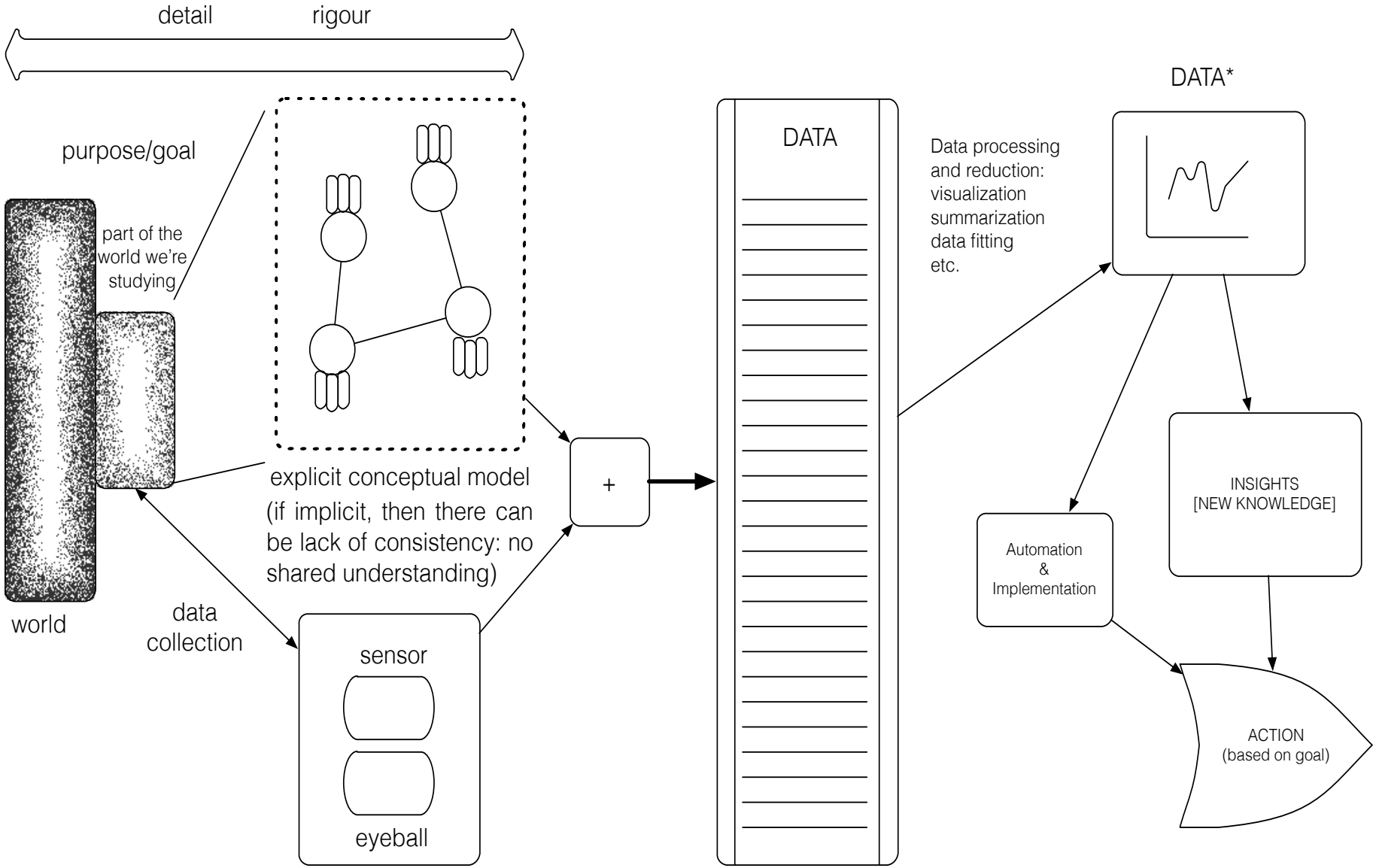
- representation
- systems
- models
- properties
- knowledge gap
- conceptual model

REPRESENTATION

A representation is an object that stands in for another object.

A representation may or may not physically resemble the object it represents.

Representations of the world help us to understand, navigate and manipulate the world.


detail
rigour
purpose/goal
part of the world we're studying
world
data collection
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
sensor
eyeball
+
DATA
Data processing and reduction: visualization summarization data fitting etc.
DATA*
Automation & Implementation
INSIGHTS [NEW KNOWLEDGE]
ACTION (based on goal)

THINKING IN SYSTEMS TERMS

In order to understand how various aspects of the World interact with one another, we need to **carve out chunks** corresponding to the aspects and define their **boundaries**.

Working with other intelligences requires **shared understanding** of what is being studied.

A **system** is made up of **objects** with **properties** that potentially change over time. Within the system we perceive **actions** and **evolving** properties leading us to think in terms of **processes**.

THINKING IN SYSTEMS TERMS

Objects themselves have various properties. Natural processes generate (or destroy) objects, and may change the properties of these objects over time.

We **observe**, **quantify**, and **record** particular values of these properties at particular points in time.

This generates data points, capturing the **underlying reality** to some degree of **accuracy** and **error** (biased or unbiased).

IDENTIFYING GAPS IN KNOWLEDGE

A **gap in knowledge** is identified when we realize that what we thought we knew about a system proves incomplete (or false).

This might happen repeatedly, at any moment in the process:

- data cleaning
- data consolidation
- data analysis

The solution is to be flexible. When faced with such a gap, **go back**, **ask questions**, and **modify the system representation**.

CONCEPTUAL MODELS

Exercise:

- assume that an acquaintance has just set foot in your living space for the first time.
- you are on the phone with them but not currently at home.
- explain to them how to go about preparing a cup of sugar.

Conceptual models are built using methodical investigation tools

- diagrams
- structured interviews
- structured descriptions
- etc.

RELATING THE DATA TO THE SYSTEM

Is the data which has been collected and analyzed going to be of any use when it comes to understanding the system?

This question can only be answered if we understand:

- how the data is **collected**
- the **approximate nature** of both data and system
- what the data **represents** (observations and features)

Is the combination of system and data **sufficient** to understand the aspects of the world under consideration?

TAKE-AWAYS

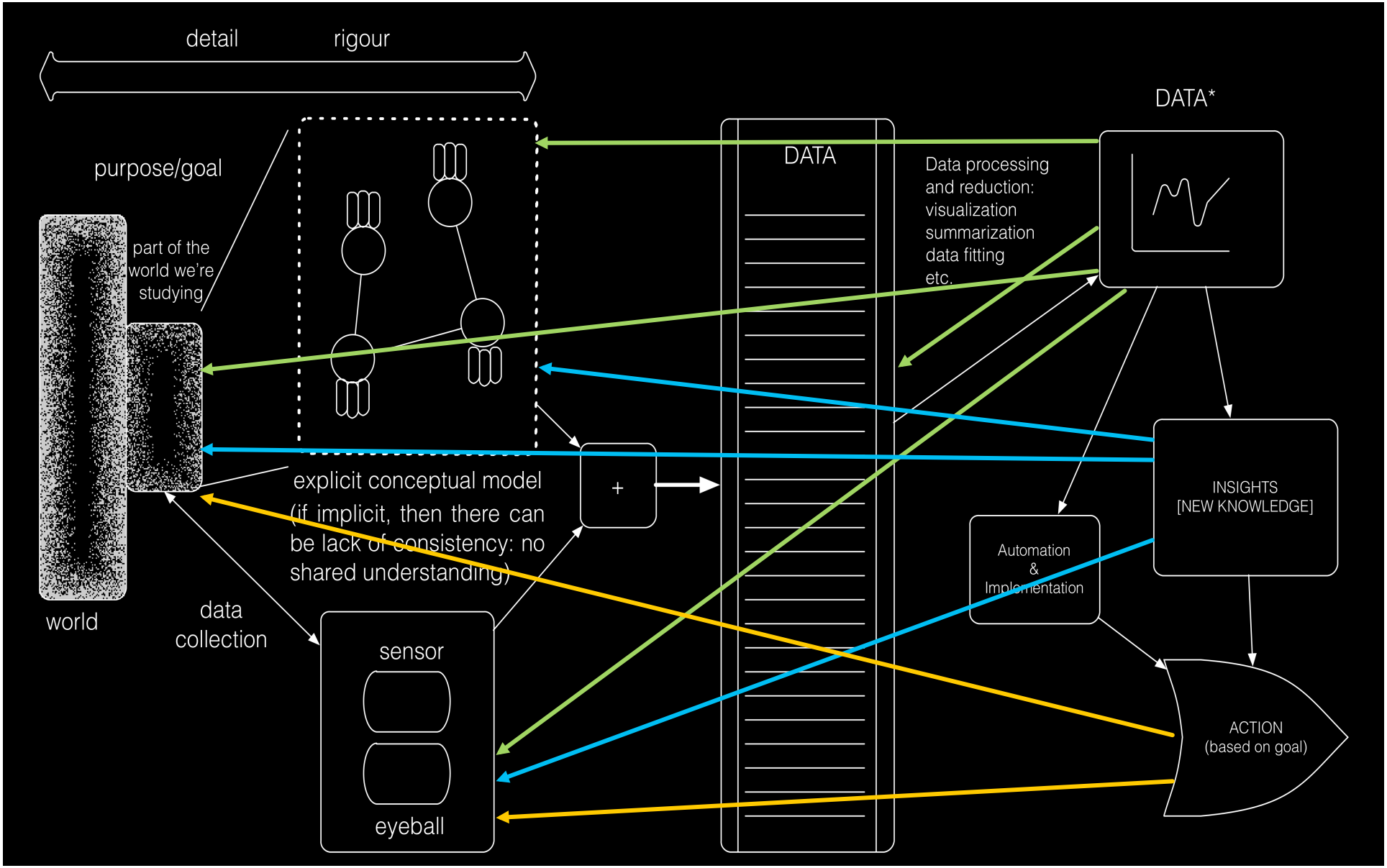
Certain aspects of the Universe can be approximated with the help of systems.

System models provide the basis under which data is identified and collected, but data itself is approximate and selective.

Knowledge gaps happen. Be prepared and ready to re-visit your set-up regularly.

We often only rely on implicit conceptual modeling, but there's danger that way.

If the data, the system, and the world are out of alignment, insights might prove useless.


detail
rigour
purpose/goal
part of the world we're studying
world
data collection
explicit conceptual model (if implicit, then there can be lack of consistency: no shared understanding)
+
sensor
eyeball
DATA
DATA*
Data processing and reduction: visualization summarization data fitting etc.
INSIGHTS [NEW KNOWLEDGE]
Automation & Implementation
ACTION (based on goal)

ETHICAL CONSIDERATIONS AND BEST PRACTICES

DATA ANALYSIS UNIVERSALS

"We have flown the air like birds and swum the sea like fishes, but have yet to learn the simple act of walking the Earth like brothers."

Martin Luther King, Jr.

MODULE LEARNING OBJECTIVES

Preliminary familiarity with the following concepts:

- ethics and best practices
- First Nations principles (OCAP)
- “do no harm“
- informed consent
- privacy
- model validity

DISCUSSION

What harm can come from data?

THE NEED FOR ETHICS

Formerly: "**Wild West**" mentality to data collection (and use). Whatever wasn't technologically forbidden was allowed.

Now: professional codes of conduct are being devised for data scientists (outline responsible ways to practice data science).

Additional responsibility for data scientists; but also **protection** against being hired to carry out questionable analyses.

Does your organization have a code of ethics for its data scientists? For its employees?

WHAT ARE ETHICS?

Broadly speaking, ethics refers to the **study** and **definition** of **right and wrong conducts:**

- "not [...] social convention, religious beliefs, or laws". (R.W. Paul, L. Elder)

Influential *Western* ethical theories:

- Kant's **golden rule** (do onto others...), **consequentialism** (the ends justify the means), **utilitarianism** (act in order to maximize positive effect), etc.

Influential *Eastern* ethical theories:

- **Confucianism**, **Taoism**, **Buddhism** (?), etc.

[OCAP® is a registered trademark of the *First Nations Information Governance Centre* (FNIGC), fnigc.ca/OCAP]

WHAT ARE ETHICS?

First Nations Principles of **OCAP®**:

- **Ownership**
  cultural knowledge, data, and information is owned by First Nations communities
- **Control**
  First Nations communities have the right to control all aspects of research and information management that impact them
- **Access**
  First Nations communities must have access to information and data about themselves no matter where it is held
- **Possession**
  First Nations communities must have physical control of relevant data

[University of Virginia, Centre for Big Data Ethics, Law, and Policy]

ETHICS IN THE DATA CONTEXT

Data ethics questions:

- **Who**, if anyone, owns data?
- Are there **limits** to how data can be used?
- Are there **value-biases** built into certain analytics?
- Are there categories that should **not** be used in analyzing personal data?
- Should some data be **publicly available** to **all** researchers?

Analytically, the **general** is preferred to the **anecdotal** – decisions made on the basis of machine learning and A.I. (security, financial, marketing, etc.) may affect real beings in **unpredictable ways**.

BEST PRACTICES

"Do No Harm": data collected from an individual **should not be used to harm** the individual.

Informed Consent:

- Individuals must **agree to the collection and use** of their data
- Individuals must have a **real understanding of what they are consenting to**, and of **possible consequences** for them and others

Respect "Privacy": excessively hard to maintain in the age of constant trawling of the Internet for personal data.

BEST PRACTICES

Keep Data Public: data should be kept **public** (all? most? any?).

Opt-In/Opt-Out: Informed consent requires the ability to **opt out**.

Anonymize Data: removal of id fields from data prior to analysis.

"Let the Data Speak":

- no cherry picking
- importance of validation (more on this later)
- correlation and causation (more on this later, too)
- repeatability

MODEL ASSESSMENT AND VALIDITY

Models should be **current**, **useful**, and **valid**.

Data can be used in conjunction with existing models to come to some conclusions, or can be used to update the model itself.

At what point does one determine that the current data model is **out-of-date** or is **not useful anymore**?

Past successes can lead to **reluctance** to re-assess and re-evaluate a model.

READINGS AND REFERENCES

DATA ANALYSIS UNIVERSALS

REFERENCES

First Nations – OCAP

Wikipedia article on Semi-Supervised Learning

Wikipedia article on Supervised Learning

Wikipedia article on Reinforcement Learning

Wikipedia article on Unsupervised Learning

J. Blitzen [2017], *What is it like to design a data science class?*, answer on *Quora*

J. Taylor [2017], *4 Problems with CRISP-DM*, KDNuggets.

Brin, D. [1998], *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?*, Perseus.

REFERENCES

Mayer-Schönberger, V. and Cukier, K. [2013], *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt.

Mayer-Schönberger, V. [2009], *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press.

Data Science Association, *Data Science Code of Professional Conduct*.

Chen, M. [2013], *Is 'Big Data' Actually Reinforcing Social Inequalities?*, The Nation.

Shin, L. [2013], *How the New Field of Data Science is Grappling With Ethics*, SmartPlanet.

Schutt, R. and O'Neill, C. [2013], *Doing Data Science: Straight Talk From the Front Line*, O'Reilly.

O'Neill, C. [2016], *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown.

REFERENCES

Chang, R.M., Kauffman, R.J., Kwon, Y. [2014], *Understanding the paradigm shift to computational social science in the presence of big data*, Decision Support Systems, 63:67–80, Elsevier.

Hurlburt, G.F., Voas, J. [2014], *Big Data, Networked Worlds*, IEEE Computer Society.

Introna, L.D. [2007], *Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible*, Ethics and Information Technology, 9:11–25, Springer.

Floridi, L. [2011], *The philosophy of information*, Oxford University Press.

Floridi, L. (ed) [2006], *The Cambridge handbook of information and computer ethics*, Cambridge University Press, 2006.

Big Data & Ethics

Mason, H. [2012], What is a Data Scientist?, Forbes.

REFERENCES

Schlimmer, J.S. [1987], *Concept Acquisition Through Representational Adjustment (Technical Report 87-19)*. Department of Information and Computer Science, UCalifornia, Irvine.

Iba, W., Wogulis, J., Langley, P. [1988], *Trading off Simplicity and Coverage in Incremental Concept Learning*, in Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.

Gorelik, B. [2017], Don't study data science as a career move; you'll waste your time!, gorelik.net

J. Leskovec, A. Rajaraman, J. Ullman [2015] *Mining of Massive Datasets*, Cambridge University Press.

Hastie, T., Tibshirani, R., and J. Friedman [2008], *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer.

Provost, F., Fawcett, T. [2013], *Data Science for Business*, O'Reilly.