Introduction to Data Science

From Business Problems to Machine Learning Tasks

Luis Rodrigues

luis-rodrigues-phd

AGENDA

01 › **02** › 03

Personal Presentation

Data Science Theory

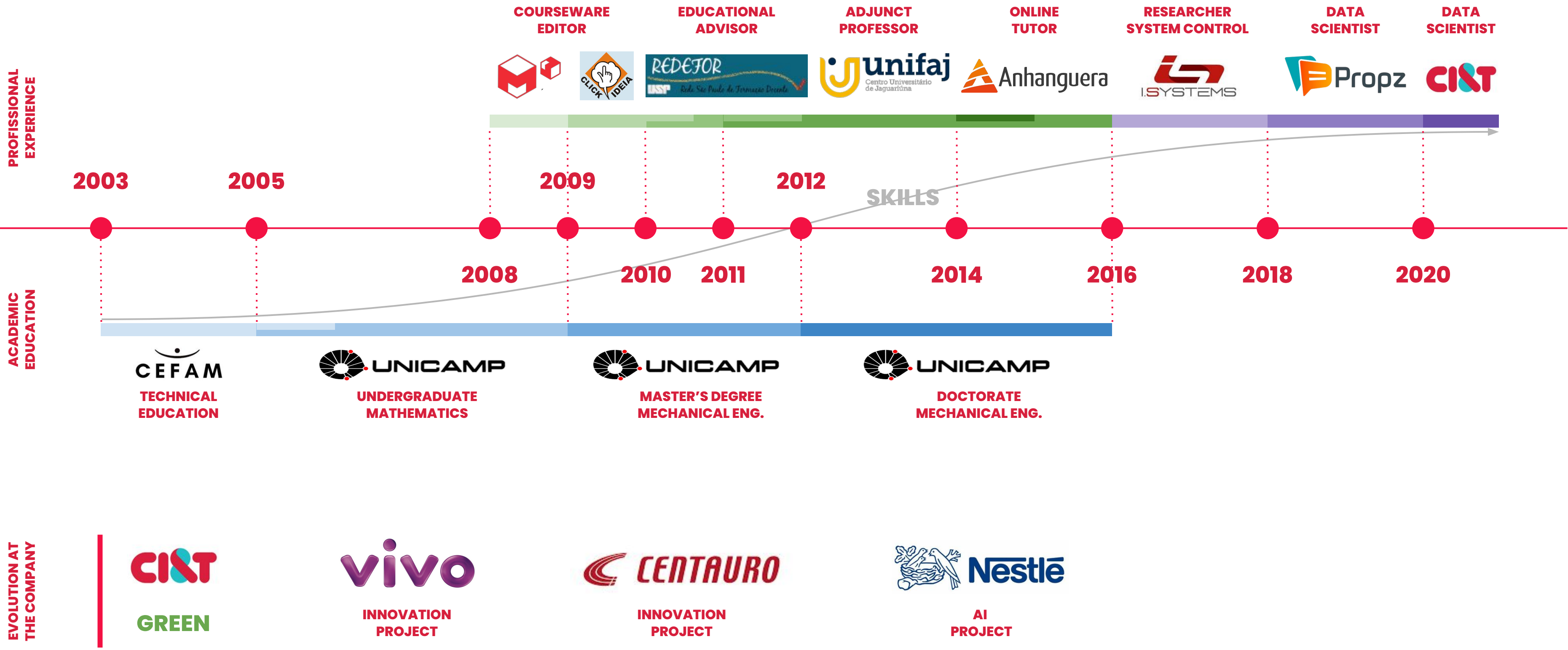
Hands-on Machine Learning

PERSONAL PRESENTATION

About who the presenter is

LUIS ANTONIO RODRIGUES

Professional Trajectory



PROFISSIONAL EXPERIENCE

COURSEWARE EDITOR — CLICK IDEIA

EDUCATIONAL ADVISOR — REDEFOR USP Rede São Paulo de Formação Docente

ADJUNCT PROFESSOR — unifaj Centro Universitário de Jaguariúna

ONLINE TUTOR — Anhanguera

RESEARCHER SYSTEM CONTROL — I.SYSTEMS

DATA SCIENTIST — Propz

DATA SCIENTIST — CI&T

2003 2005 2008 2009 2010 2011 2012 2014 2016 2018 2020

SKILLS

ACADEMIC EDUCATION

CEFAM — TECHNICAL EDUCATION

UNICAMP — UNDERGRADUATE MATHEMATICS

UNICAMP — MASTER’S DEGREE MECHANICAL ENG.

UNICAMP — DOCTORATE MECHANICAL ENG.

EVOLUTION AT THE COMPANY

CI&T — GREEN

vivo — INNOVATION PROJECT

CENTAURO — INNOVATION PROJECT

Nestlé — AI PROJECT

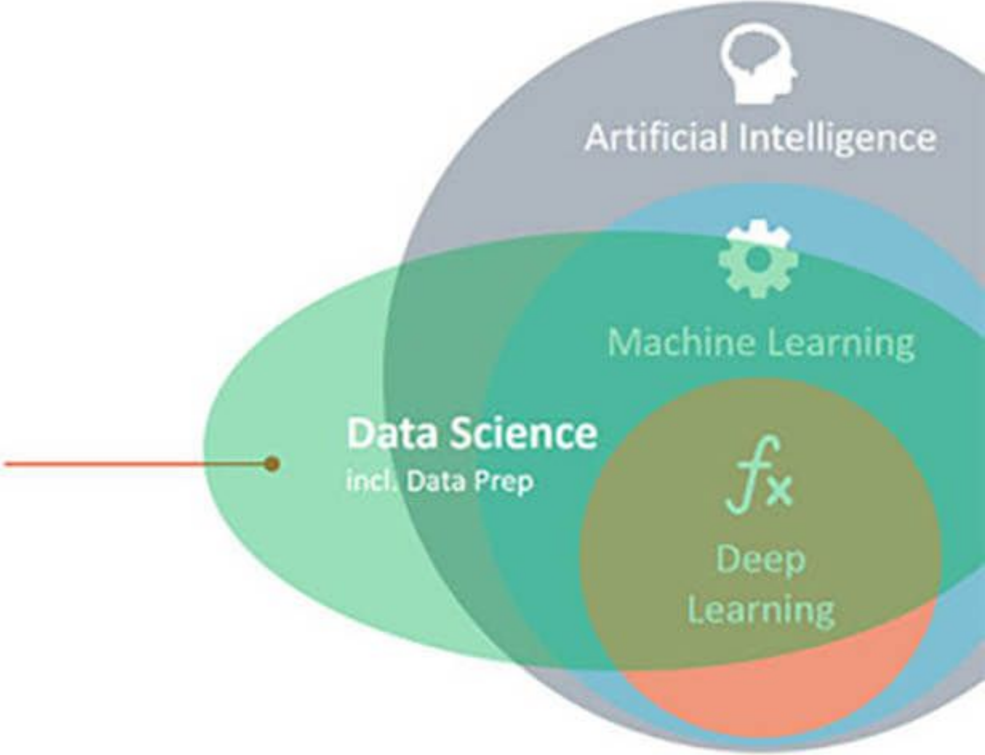
DATA SCIENCE THEORY

Introduction, ..., ML Tasks, ..., DM Process

DATA WHAT? **NOMENCLATURE** OVERVIEW

Data Science

Covers the practical application of advanced analytics, statistics, machine learning, and the necessary data preparation in a business context.



Source: https://rapidminer.com/wp-content/uploads/2019/01/data-science-ai_ml_dl.jpg

STAR WITH WHY: THE BUSINESS PROBLEM

How to solve the following business problem?

Improve the profit of a retail company

STAR WITH WHY: THE BUSINESS PROBLEM

How to solve the following business problem?

Improve the profit of a retail company

profit = revenue - cost

profit = (# clients × avg spend) - (marketing cost + …)

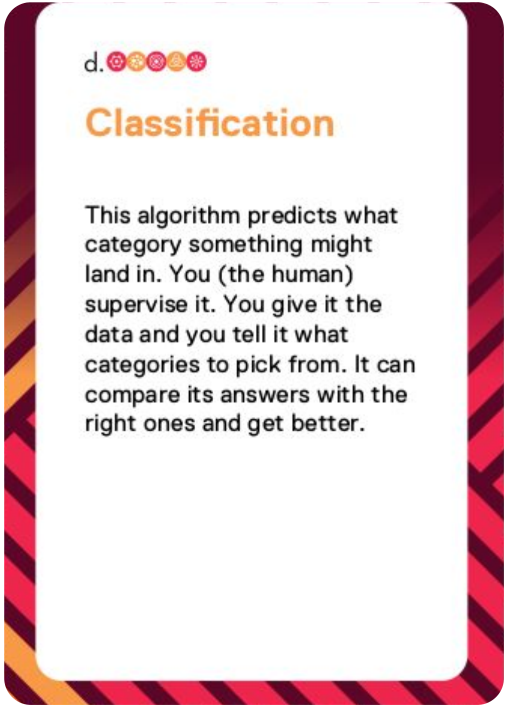
Ex: Increase the average ticket of the customers

1. Profiling customers to better understand the company niche
2. Making assertive offers through recommendation systems

Ex: Reduce the cost with marketing campaign

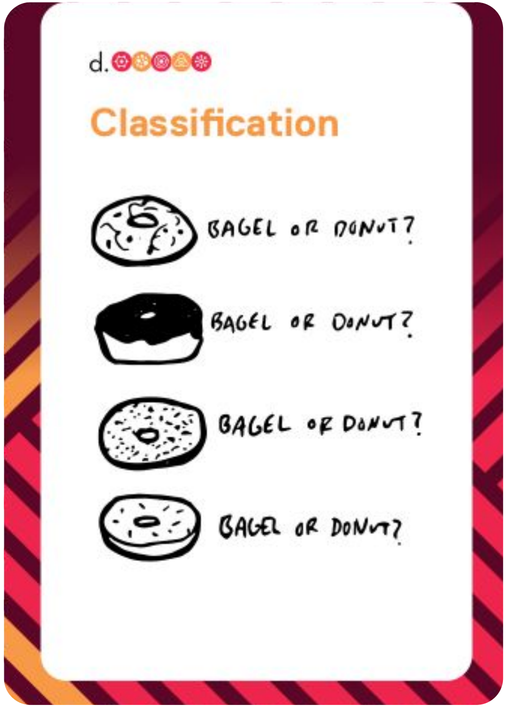
1. Contacting only customer with high propensity to click
2. Making special offers only for with high chances of churn

CLASSIFICATION TASK



Classification

This algorithm predicts what category something might land in. You (the human) supervise it. You give it the data and you tell it what categories to pick from. It can compare its answers with the right ones and get better.



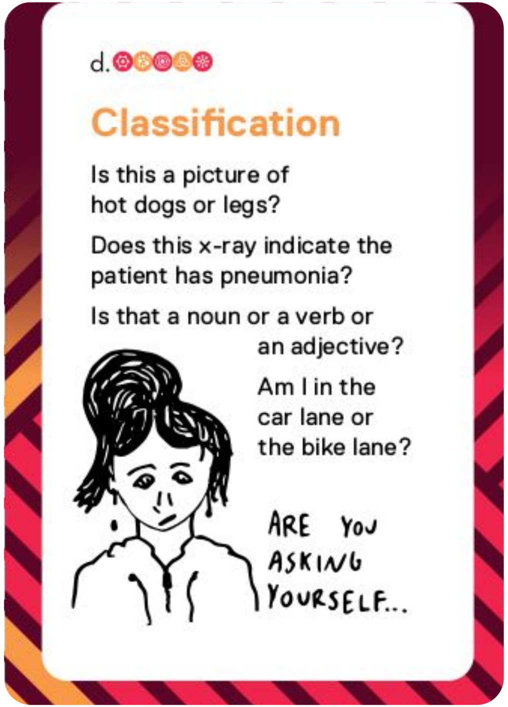
Classification

BAGEL OR DONUT?

BAGEL OR DONUT?

BAGEL OR DONUT?

BAGEL OR DONUT?



Classification

Is this a picture of hot dogs or legs?

Does this x-ray indicate the patient has pneumonia?

Is that a noun or a verb or an adjective?

Am I in the car lane or the bike lane?

ARE YOU ASKING YOURSELF...

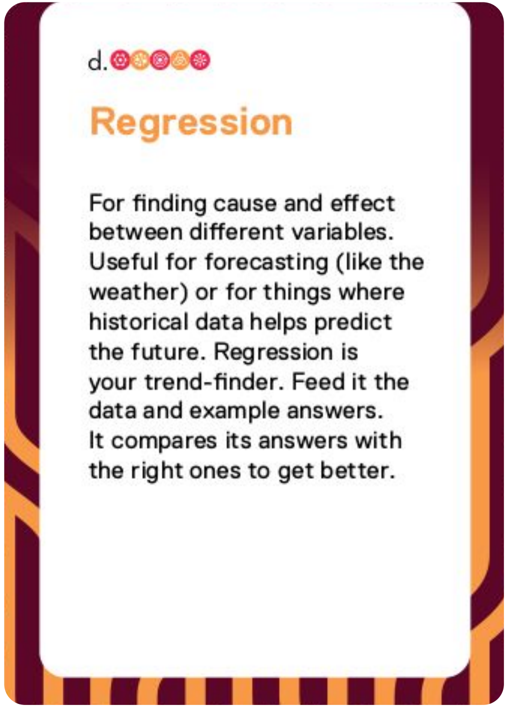
Remarks

It could be used to estimate an unknown variable and predict a future value of a categorical variable as well

Algorithms

Logistic Regression, SVM, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, Neural Networks

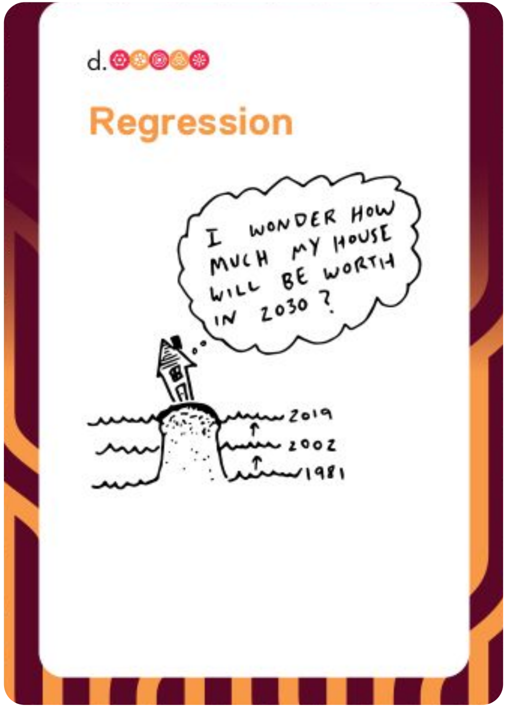
REGRESSION TASK



d.

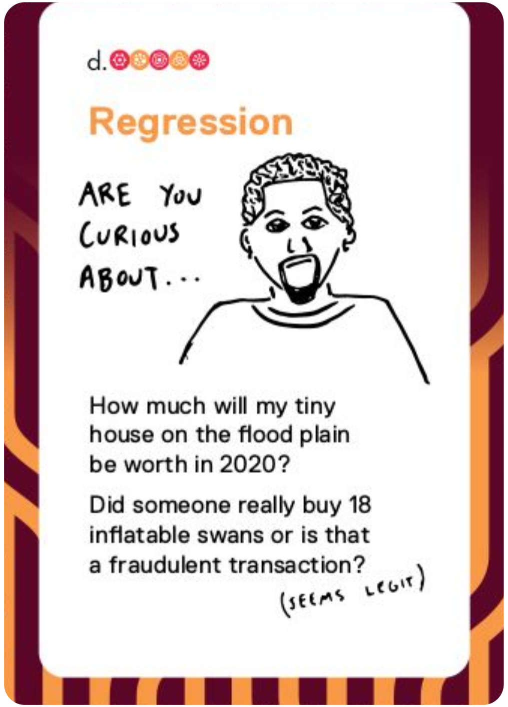
Regression

For finding cause and effect between different variables. Useful for forecasting (like the weather) or for things where historical data helps predict the future. Regression is your trend-finder. Feed it the data and example answers. It compares its answers with the right ones to get better.



d.

Regression



d.

Regression

How much will my tiny house on the flood plain be worth in 2020?

Did someone really buy 18 inflatable swans or is that a fraudulent transaction?

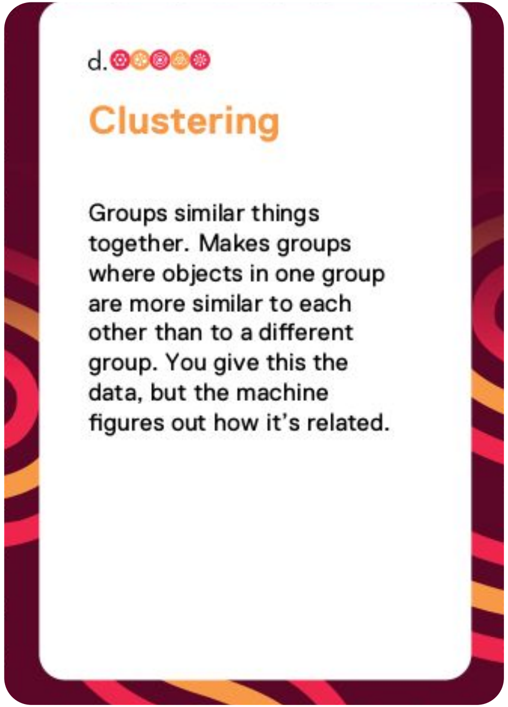
Remarks

It could be used for descriptive analytics and for predictive analytics as well.

Algorithms

Linear Regression, elastic net (ridge + lasso), RANSAC, K-NN, Random Forest, Gradient Boosting, Neural Networks

CLUSTERING TASK



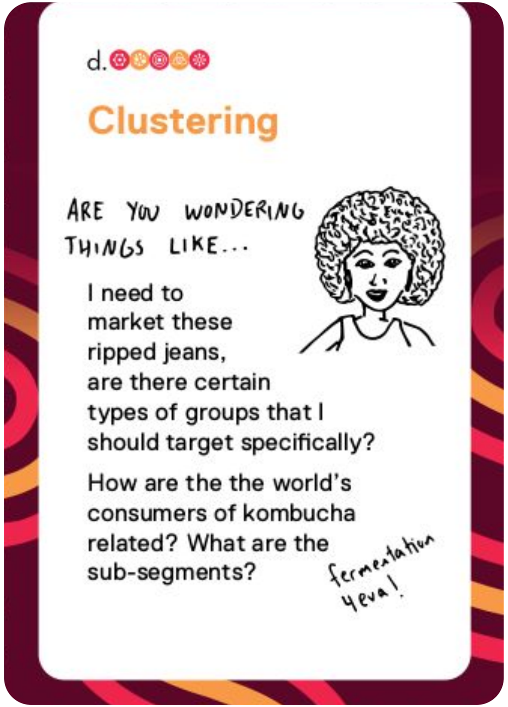
d.

Clustering

Groups similar things together. Makes groups where objects in one group are more similar to each other than to a different group. You give this the data, but the machine figures out how it's related.

d.

Clustering



d.

Clustering

ARE YOU WONDERING THINGS LIKE...

I need to market these ripped jeans, are there certain types of groups that I should target specifically?

How are the the world's consumers of kombucha related? What are the sub-segments?

fermentation 4eva!

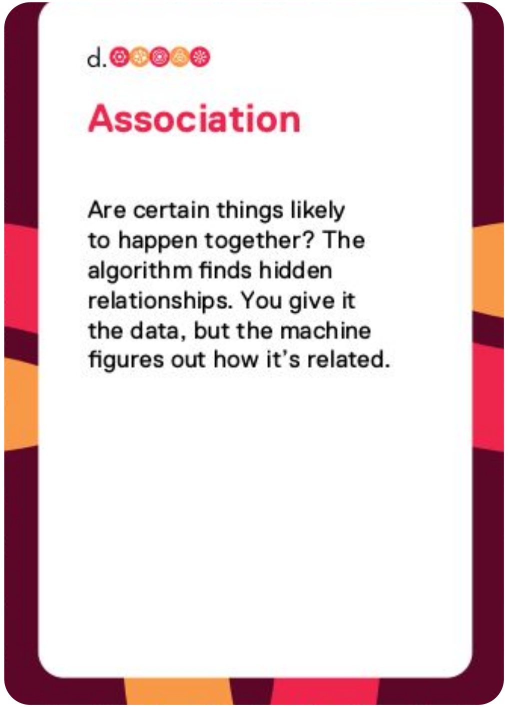
Remarks

It could be used for hierarchical (tree like structure) and non-hierarchical clustering (partitional, density based,...)

Algorithms

Hierarchical: HDBSCAN, Linkage. Non-hierarchical: K-Means, Gaussian Mixture Models, Mean Shift, DBSCAN

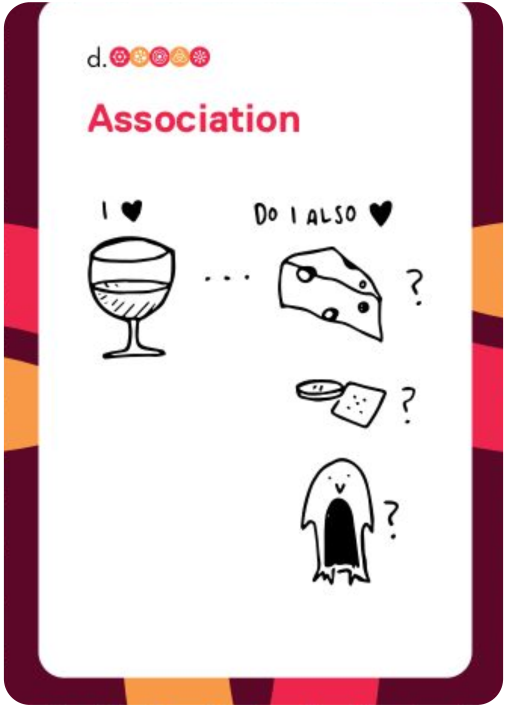
CO-OCCURRENCE GROUPING TASK



d.

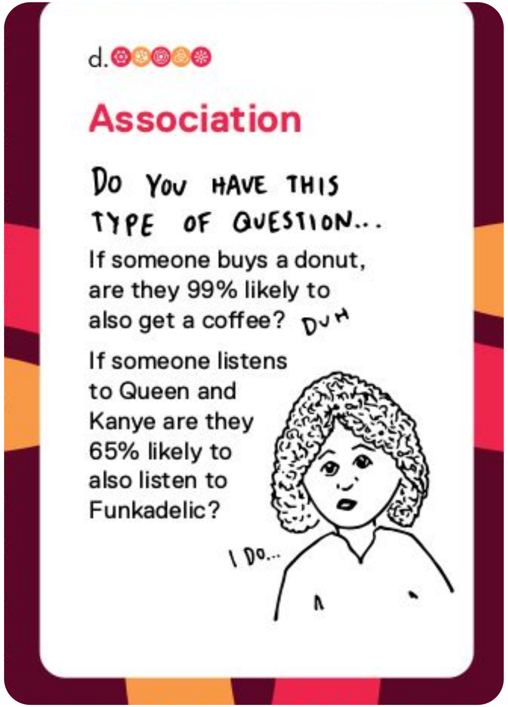
Association

Are certain things likely to happen together? The algorithm finds hidden relationships. You give it the data, but the machine figures out how it's related.



d.

Association



d.

Association

DO YOU HAVE THIS TYPE OF QUESTION...

If someone buys a donut, are they 99% likely to also get a coffee? DUH

If someone listens to Queen and Kanye are they 65% likely to also listen to Funkadelic?

I DO...

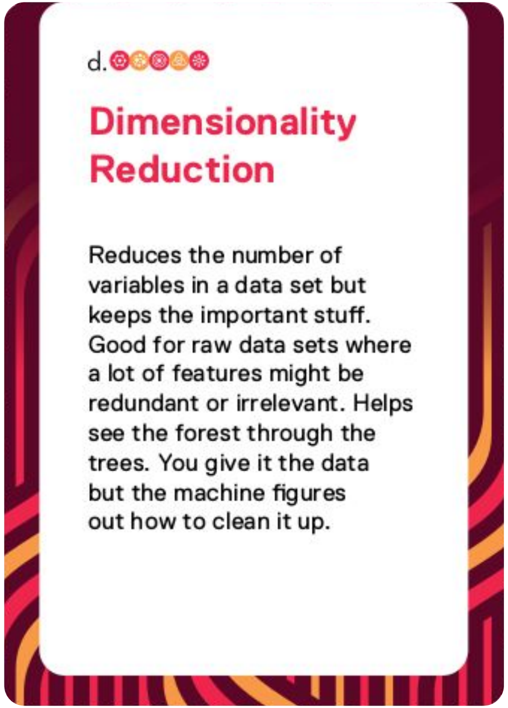
Remarks

Also known as market-basket analysis,
it could be used to find associations between entities

Algorithms

Apriori, Eclat, SETM, FP-growth

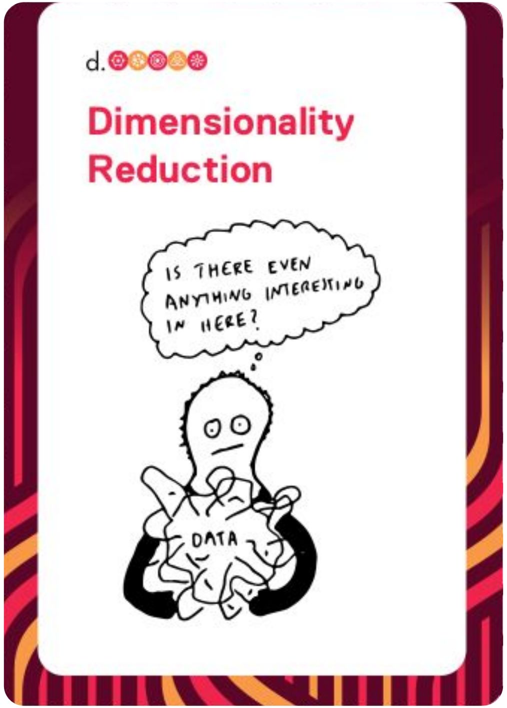
DATA REDUCTION TASK



d.

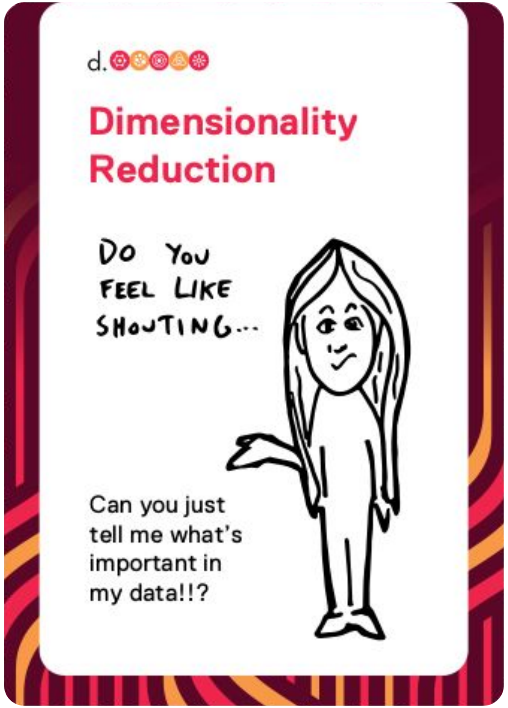
Dimensionality Reduction

Reduces the number of variables in a data set but keeps the important stuff. Good for raw data sets where a lot of features might be redundant or irrelevant. Helps see the forest through the trees. You give it the data but the machine figures out how to clean it up.



d.

Dimensionality Reduction



d.

Dimensionality Reduction

Can you just tell me what's important in my data!!?

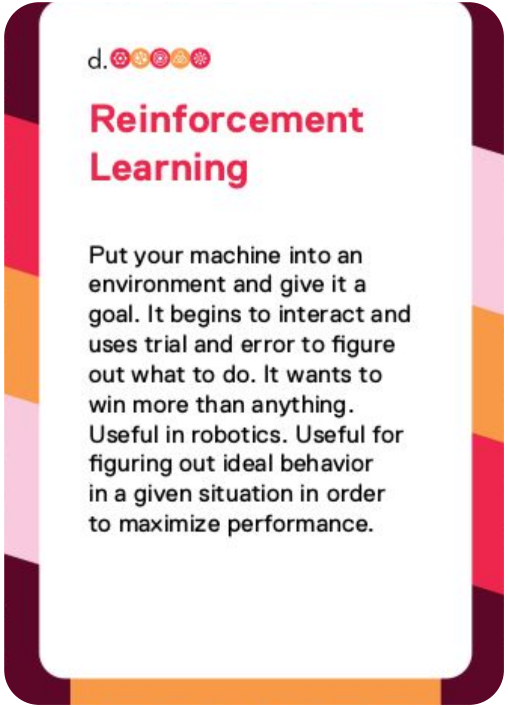
Remarks

It could be used to transform data from a high-dimensional space into a low-dimensional space

Algorithms

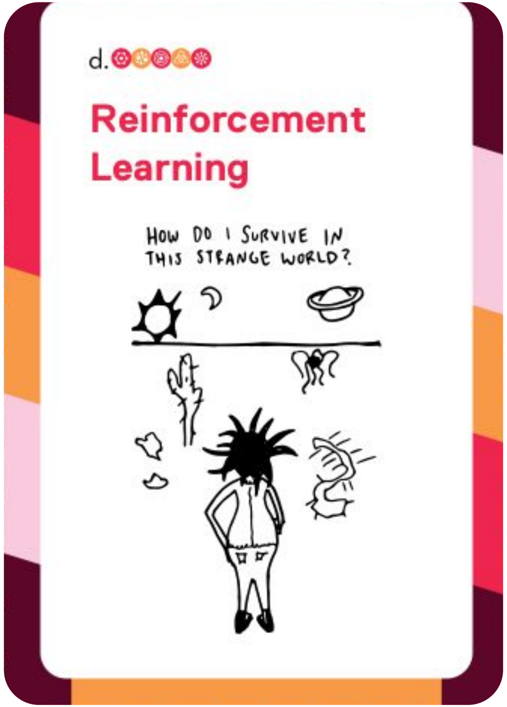
PCA, TruncatedSVD (also LSA for DT matrices), t-SNE, UMAP, Auto-Encoders, LDA (for topic modeling)

REINFORCEMENT LEARNING



Reinforcement Learning

Put your machine into an environment and give it a goal. It begins to interact and uses trial and error to figure out what to do. It wants to win more than anything. Useful in robotics. Useful for figuring out ideal behavior in a given situation in order to maximize performance.



Reinforcement Learning

Reinforcement Learning

How do I win this game?

How might this car drive itself?

How to optimize marketing so someone will click click click?

Remarks

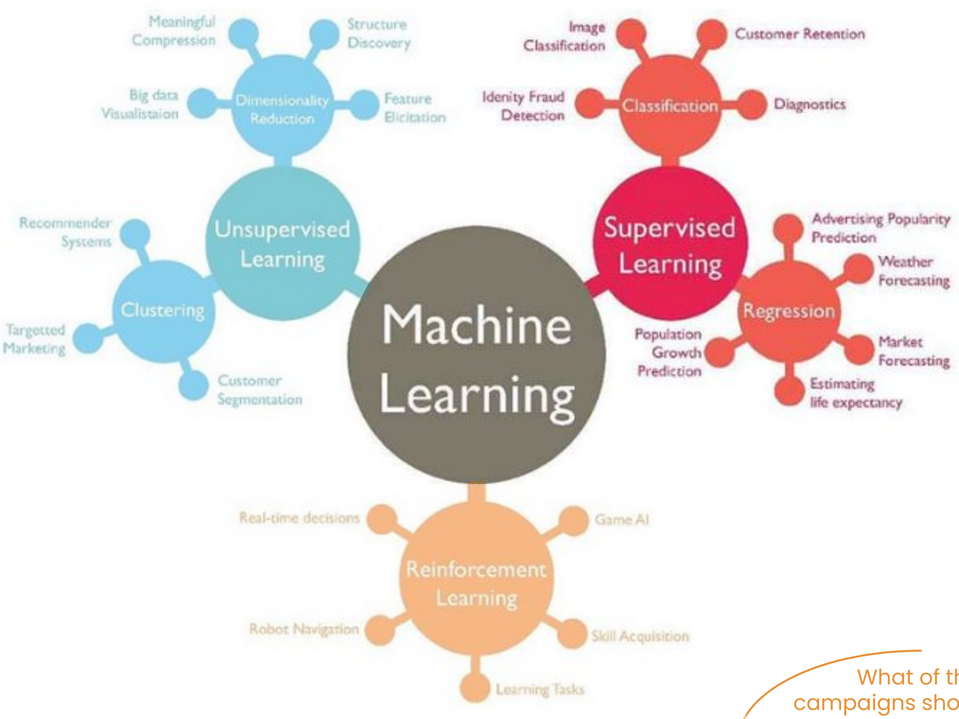
It could be used to define the best sequence of decisions to solve a problem while maximizing a reward

Algorithms

Epsilon-greedy, UCB1, Thompson Sampling (Bayesian bandit), SARSA, Q-learning

TYPES OF LEARNING AND TASKS COVERED

Do our customers naturally fall into different groups? No specific purpose or target has been specified for the grouping.



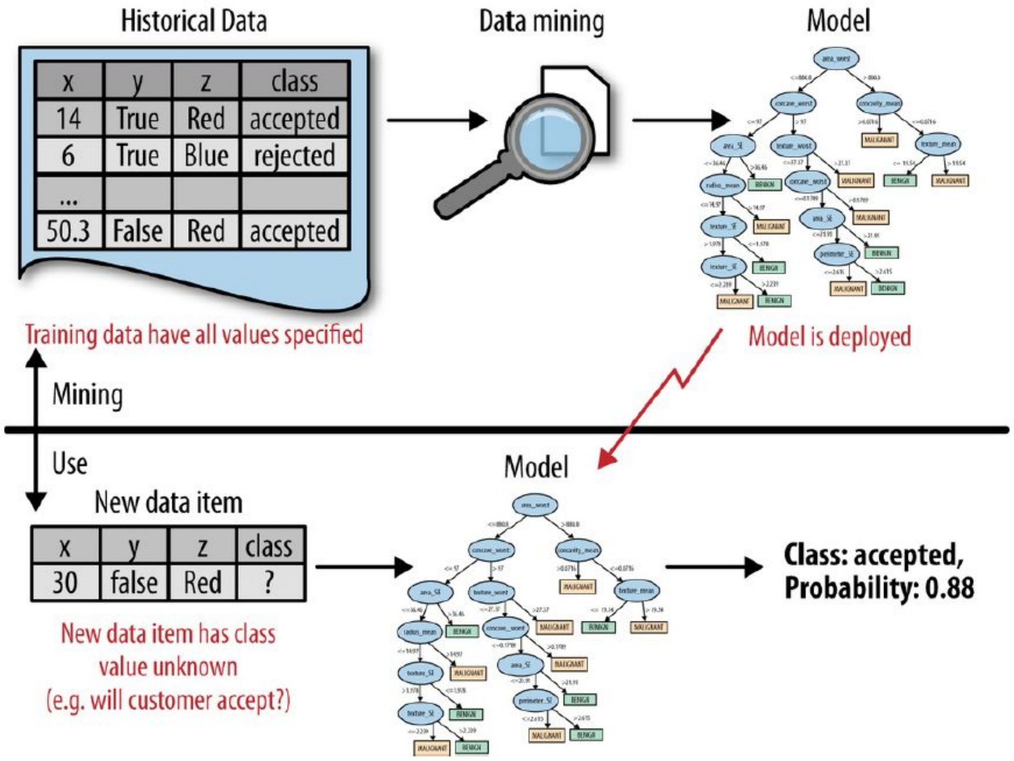
Can we find groups of customers who have particularly high likelihoods of canceling their service soon?

What of these new marketing campaigns should be shown to our customers? There is no previous data, actions are learned through the interaction with environment and observation of reward

Source: https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/

Modeling: **Training vs Prediction**

The output of the modeling process is an algorithm (the trained model)



x	y	z	class
14	True	Red	accepted
6	True	Blue	rejected
...			
50.3	False	Red	accepted

Training data have all values specified

Model is deployed

x	y	z	class
30	false	Red	?

New data item has class value unknown (e.g. will customer accept?)

Class: accepted, Probability: 0.88

Training

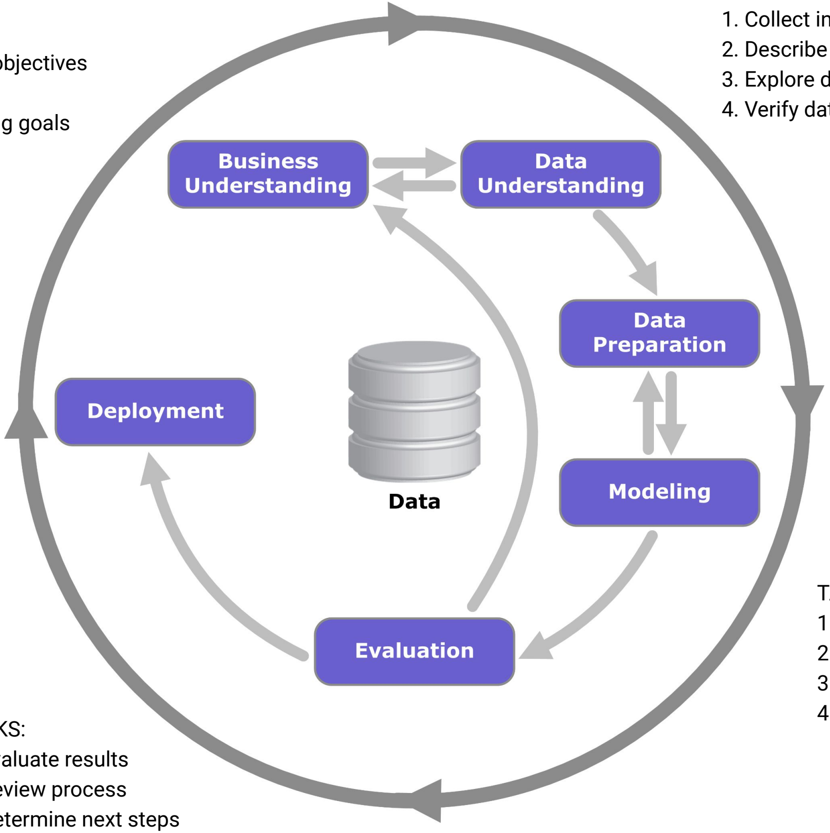
Prediction

The trained model is used to get the answer for new cases (data that was not used in training)

Source: Data Science for Business book.

CRISP-DM | CRoss Industry Standard Process for Data Mining

Process Diagram



TASKS:
1. Determine business objectives
2. Assess situation
3. Determine data mining goals
4. Produce project plan

TASKS:
1. Collect initial data
2. Describe data
3. Explore data
4. Verify data quality

TASKS:
1. Select data
2. Clean data
3. Construct data
4. Integrate data
5. Format data

TASKS:
1. Select modeling techniques
2. Generate test design
3. Build model
4. Assess model

TASKS:
1. Evaluate results
2. Review process
3. Determine next steps

TASKS:
1. Plan deployment
2. Plan monitoring and maintenance
3. Produce final report
4. Review project

ANSWERING BUSINESS QUESTIONS WITH THESE ML TASKS

1. Who are the most profitable customers?

   A **database querying** could be used to retrieve a list of customer order by total spend

2. Is there a really difference between the profitable customers and the average customer?

   A **statistical hypothesis test** (the famous A/B test) could be used to confirm or reject.

3. But who really are these customers? Can I characterize them?

   It could be used **database querying** to extract individual characteristics and summary statistics or **machine learning** techniques to automatically find patterns as well

4. Will some particular new customer be profitable? How much revenue should I expect this customer to generate?

   Machine learning could be used to produce predictive models of profitability from historical data that can be applied to new customer to generate predictions.

REFERENCES

1. Book: Data Science for Business

2. Resource: I Love Algorithms

3. Website: Machine Learning Algorithm - Backbone of emerging technologies

4. Website: What is CRISP-DM?

HANDS-ON
MACHINE LEARNING

Classification Task

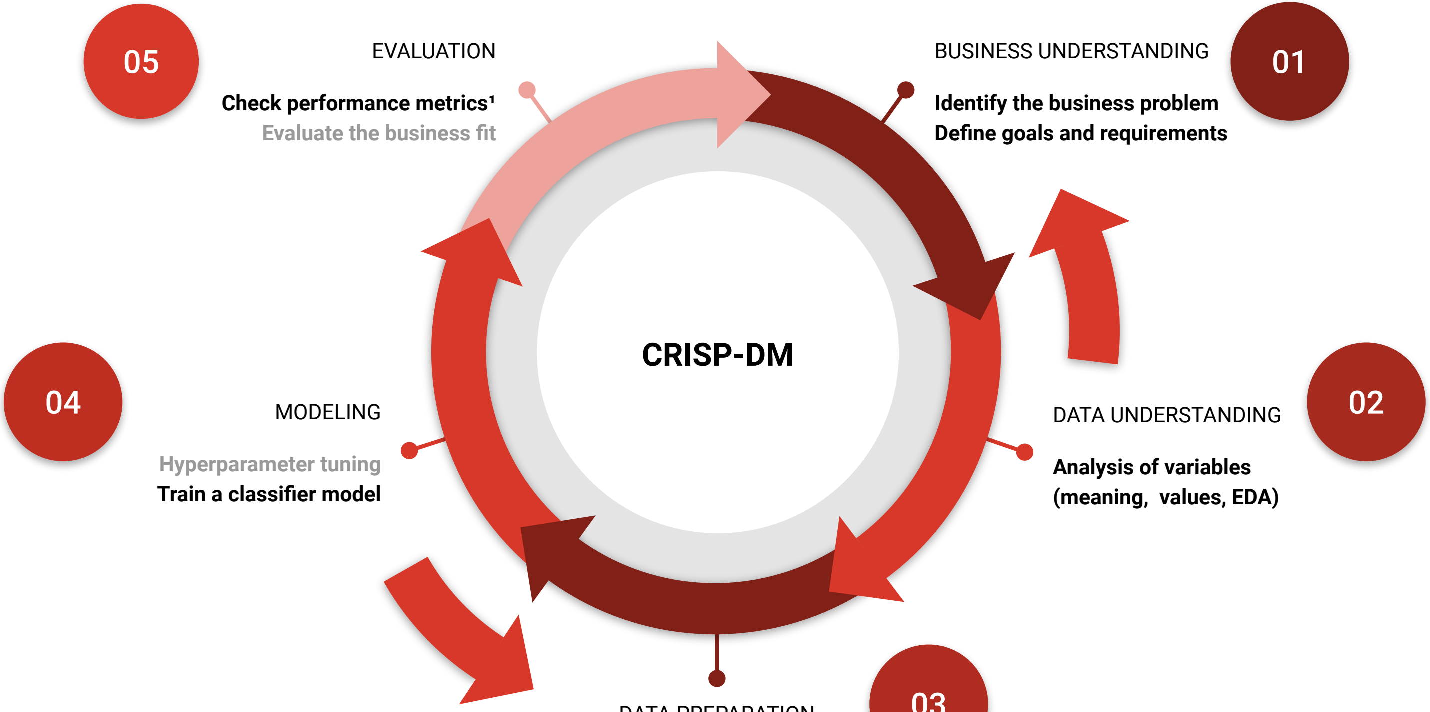
HANDS-ON ML: **BUILDING A ROCK CLASSIFIER**

Binary classification

- Database: Kaggle Database: Music Genre Classification
  - Training dataset: 17,996 rows with 17 columns
  - Column details: artist name; track name; popularity; ‘danceability’; energy; key; loudness; mode; ‘speechiness’; ‘acousticness’; ‘instrumentalness’; liveness; valence; tempo; duration in milliseconds and time_signature.
  - Target Variable: 'Genre’ such as Rock, Indie, Alt, Pop, Metal, HipHop, Alt_Music, Blues, Acoustic/Folk, Instrumental, Country, Bollywood.
- Study Case:
  - [Binary Classification] Build a genre classifier to identify if a song is Rock or other.
- Source code:
  - Hands-on ML: (Imbalanced) Binary Classification.ipynb
- Challenge:
  - 1: [Binary Classification] Build a genre classifier to identify if a song is Blues or other.
  - 2: [Multi-class classification] Build a genre classifier to identify if a song is Rock, Pop or other.

APPLICATION OF CRISP-DM FOR A **CLASSIFICATION TASK**

Incremental and Continuous Value Delivery



CRISP-DM

01 BUSINESS UNDERSTANDING

Identify the business problem
Define goals and requirements

02 DATA UNDERSTANDING

Analysis of variables (meaning, values, EDA)

03 DATA PREPARATION

Treat null values and outliers
Feature engineering

04 MODELING

Hyperparameter tuning
Train a classifier model

05 EVALUATION

Check performance metrics[1]
Evaluate the business fit

[1] Tasks usually included in the modeling step.
DISCLAIMER: In this project the deployment step will not be done, consequently, being removed.

ciandt.com

THANK YOU

CI&T Make Their **Tomorrow**