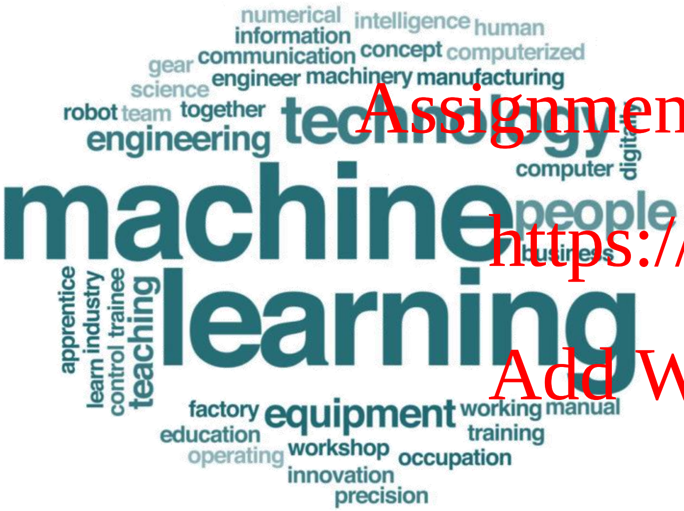
Understanding Machine Learning (Concepts)

December 2020



Contents

- What is Machine Learning
- Case Study: Using Machine Learning to Classify Emails
- Machine Learning Models
  - Regression
  - Classification
  - Clustering
  - Deep Learning
- The Machine Learning Process

What is Machine Learning (ML)

What does it mean to learn?

- How did you learn to read?
- Learning requires identifying patterns
  - Identify patterns
    - Identify letters and then the patterns of letters together to form words
  - Recognize those patterns when you see them again
- That is what machine learning (ML) does with data that we provide

Identifying patterns in some amount of data is easier but the predictive power of such patterns might be limited

Name	Amount	Fraudulent
Daniel	$2,600.45	No
Alex	$2,294.58	Yes
Adrian	$1,003.30	Yes
Vicky	$8,488.32	No

What is the pattern for fraudulent transactions?

It's obvious, isn't it?

If the name starts with “A”, they are a criminal

- The problem with having so little data is that it is easy to find patterns, but it is hard to find patterns that are correct
- Correct in the sense that they are predictive - they help us understand whether a new transaction is likely to be fraudulent

More data helps to identify more meaningful patterns but accuracy remains an issue

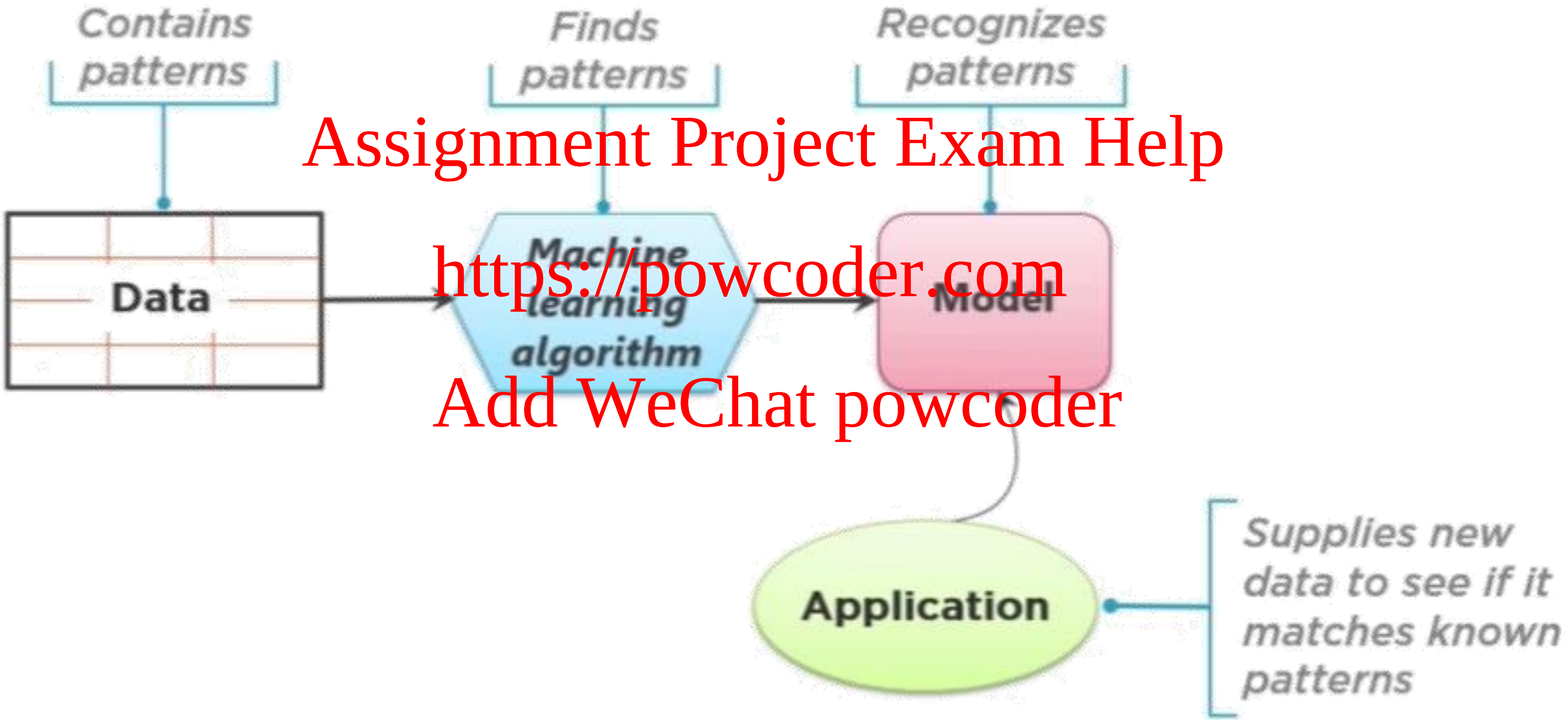
Name	Amount	Where Issued	Where Used	Age	Fraudulent
Daniel	$2,600.45	HK	HK	22	No
Alex	$2,294.58	HK	RUS	29	Yes
Adrian	$1,003.30	HK	RUS	25	Yes
Vicky	$8,488.32	JAP	HK	64	No
Adams	$200.12	AUS	JAP	53	No
Jones	$3,250.11	HK	RUS	43	No
Mary	$8,156.20	HK	RUS	27	Yes
Max	$7,475.11	UK	GER	32	No
Peter	$500.00	HK	RUS	27	No
Anson	$7,475.11	HK	RUS	20	Yes

A transaction is fraudulent

- if the card holder is in their 20's
- if the card is issued in Hong Kong (HK) and used in Russia (RUS)
- the amount is more than $1,000

But once again, do we know that that pattern is truly predictive?

Building AI systems is about finding patterns in data and developing AI models to recognise the patterns in new data



Why is machine learning so hot right now?

- Doing ML well requires
  - Lots of data
  - Lots of computer power
  - Effective ML algorithms
- All of those things are now more available than ever

Data and Technologies Have Been Democratised

Machine Learning Has Gone Mainstream

Who is interested in machine learning?

Business Leaders

Want solutions to business problems

Software Developers

Want to create better applications

Data Scientist

Want powerful, easy-to-use tools

Like most technologies, machine learning can raise ethical issues

- Recall the basic model
  - We start with data, we process that data using ML algorithms to produce a model
  - We then use that model to make decisions

- But what happens if the data is biased?

  - Suppose we have a bank lending model, but suppose the data that we use to create that model is from historic loan patterns and contains racial bias
  - If that is the case, our model will also contain that racial bias and we might not even know it because the data could be so large that we could not see the bias ourselves

- Suppose someone accuses you of having a biased model. How can you explain the model decision?

ML models are very different software, their behaviour cannot be easily revealed through code examination

- Models generated by ML are different from other kinds of software
  - Traditional software is written directly by people who could work out in great detail exactly what the software does
  - If you need to, somebody could look at that code directly to figure out why it behaves in a certain way
- With ML, models are typically generated using complex statistical techniques and the result is not ordinary computer code
  - You cannot just look at it to see why it is doing what it is doing
- Some models can be very hard to explain and there are scenarios where you might be required, legally required to explain your model

Machine Learning Use Cases

American Fidelity Assurance

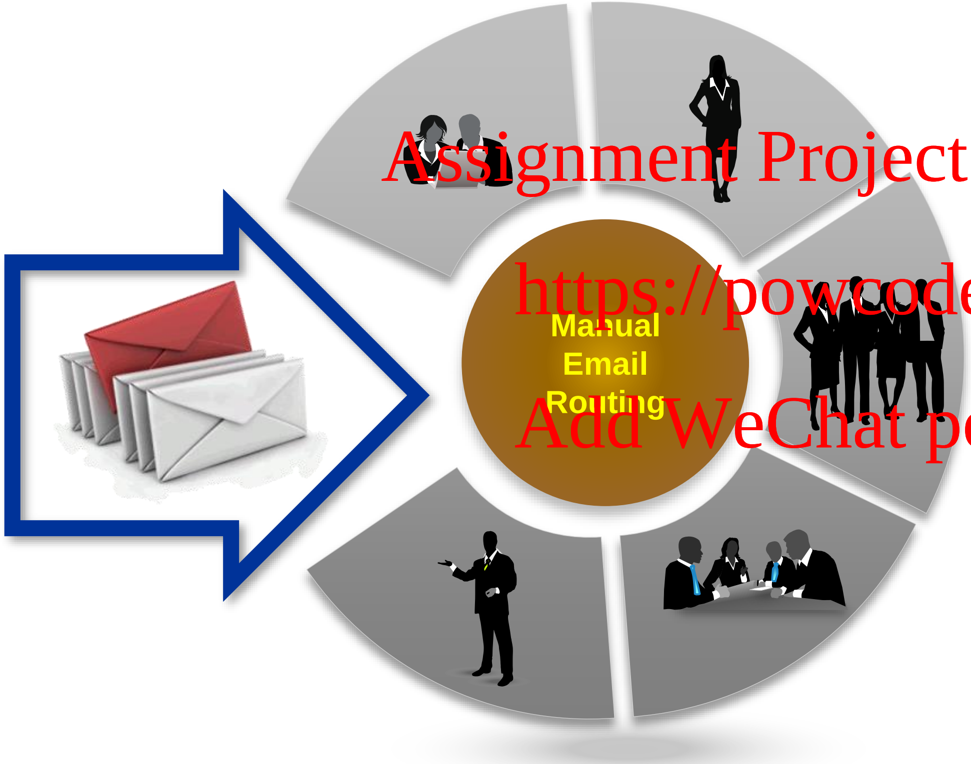
AMERICAN FIDELITY

a different opinion

American Fidelity Assurance handling large number of insurance policies needs to streamline its operations

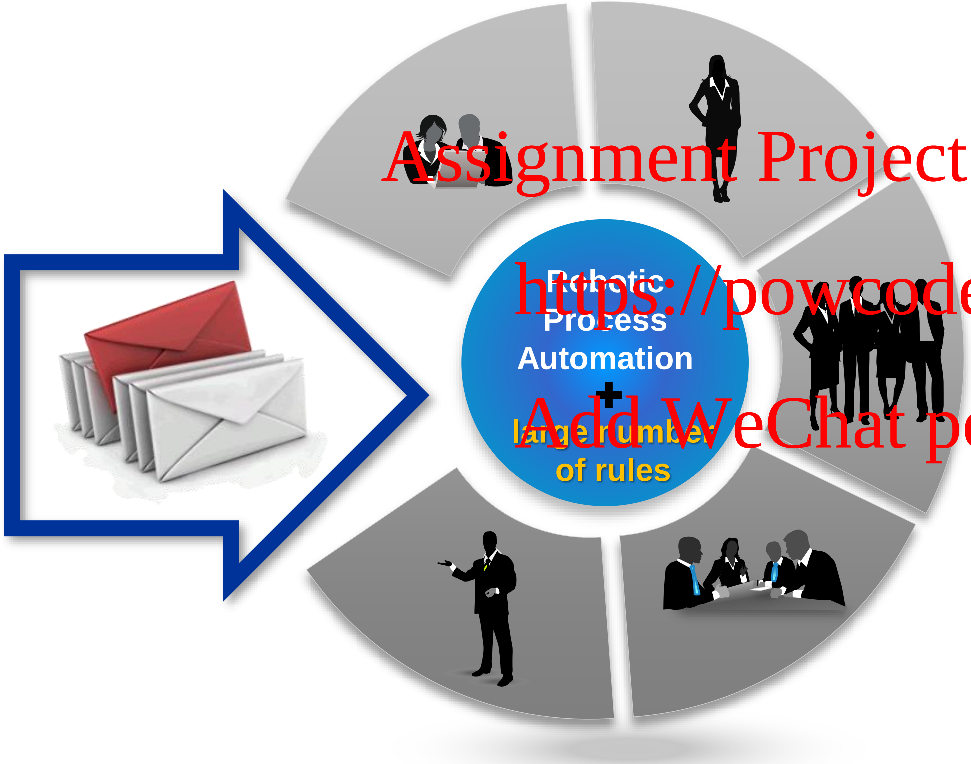
- American Fidelity Assurance (AFA) is an American private, family-owned life and health insurance company
- It provides voluntary supplemental health insurance products (cancer, disability, life and hospital indemnity) and tax deferred annuities to education employees, auto dealerships, health care providers and municipal workers across the United States
- Headquartered in Oklahoma City, AFA is a subsidiary of American Fidelity Corporation, which is owned by the founding Cameron family
- The company handled 2.5 million health insurance policies

Emails in the centralized queue are examined manually to decide to which departments they should be forwarded



- AFA gets a lot of customer emails on all sorts of subjects - from claims to address changes – in a centralized queue
- A key task, then, is to identify the primary topic of the email and forward it to the department best suited to address it

Building intelligent email routing using RPA rules resulted in too many rules being defined & to maintain



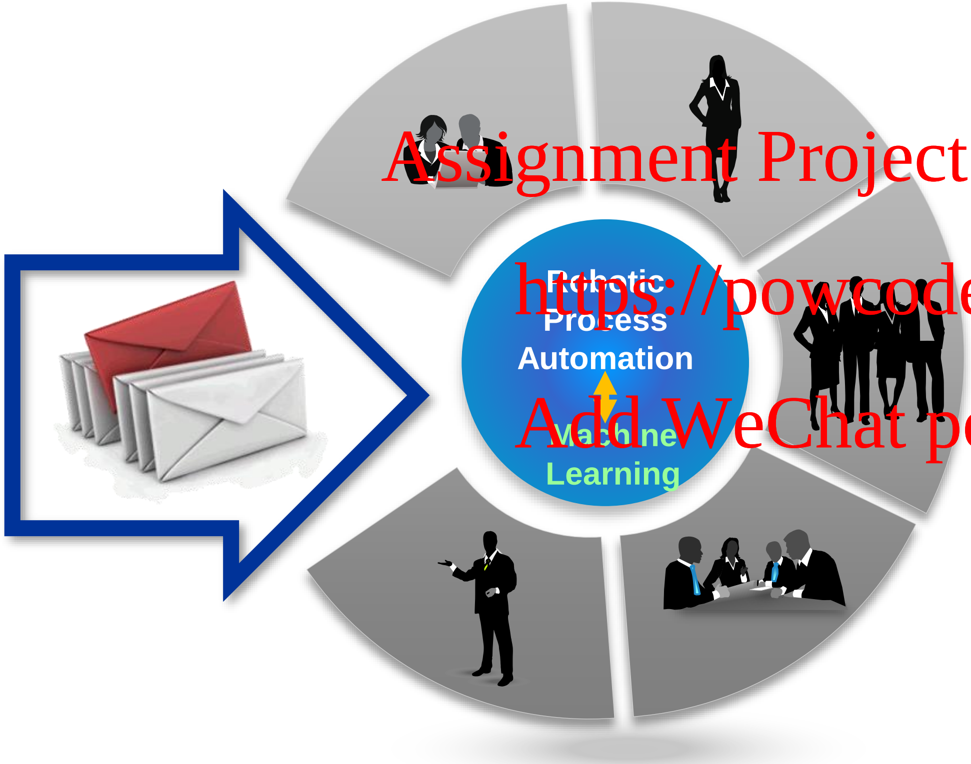
- RPA was believed to have a role to play in the sending emails automatically to the right department
- But what was the best way to decide which department should receive the email?
- AFA tried using RPA-based rules to classify keywords, but that approach resulted in too many rules

Historical email routing data is used as the training dataset in the machine learning solution

- The alternative was to try machine learning as a way to classify emails
- That technology - at least the supervised learning form of it - requires a substantial amount of labeled data with the correct outcome for purposes of training the model
- AFA already had a database of customer emails and outcomes - the department that eventually responded back to the customer
- It served as an excellent training dataset
  - The algorithm is based on an analysis of 10,000 actual emails to see which departments responded to various words and phrases and create the right routing rules

Machine learning technology is used to pick the best algorithm and provide an API for the RPA system to call

https://www.youtube.com/watch?v=JDbg4y_NPAQ



- Existing email data is analyzed to discover the model that fits the training data best
- Testing data is then used to prove that the ML model is able to accurately predict which department to receive the email
- The RPA system only needs to route the email to the machine learning API to make the prediction and then send the email accordingly

The combination of RPA and machine learning significantly improved the email distribution process

- The RPA robot starts by opening each email, extracting all text, and sending this information to the ML model through the generated API

- The ML model classifies each customer email into the best fitting category

- Once the emails are classified, the ML model returns them to the RPA robot, which automatically routes them to the classified department

- This system even allows for employees to get involved if they are needed
  - e.g., a lot of back-and-forth in a particular email conversation could indicate customer frustration and the RPA robot can immediately send the email to an employee to faster, more effective resolution
- This has a significant impact on the business and customer experience

Delivering boutique customized experience at scale

"I heard someone talking about a 1950's store owner in a small town who knew each customer. That store owner knew you, he knew your preferences, and he was able to provide an amazing customer experience. The robots allow us to create that boutique, customized experience, unique to every customer – at scale"

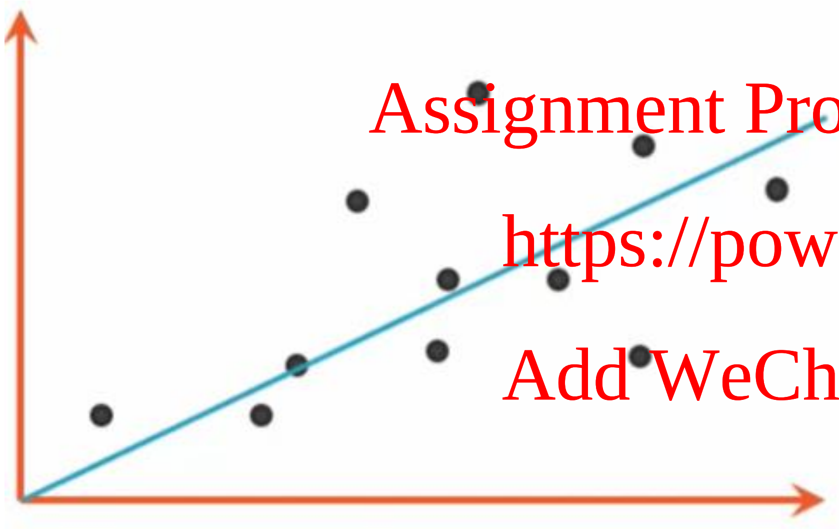
Shane Jason Mock
Vice President of Research and Development, AFA

Lessons Learned

- Robots are good at moving data across multiple systems
  - They do it faster and with fewer errors than humans
  - Every 1 hour spent on building bots automates approximately 10 hours' worth of tasks
- Scanning 9,000 emails that took 45 staff hours previously would take 3 seconds by bots

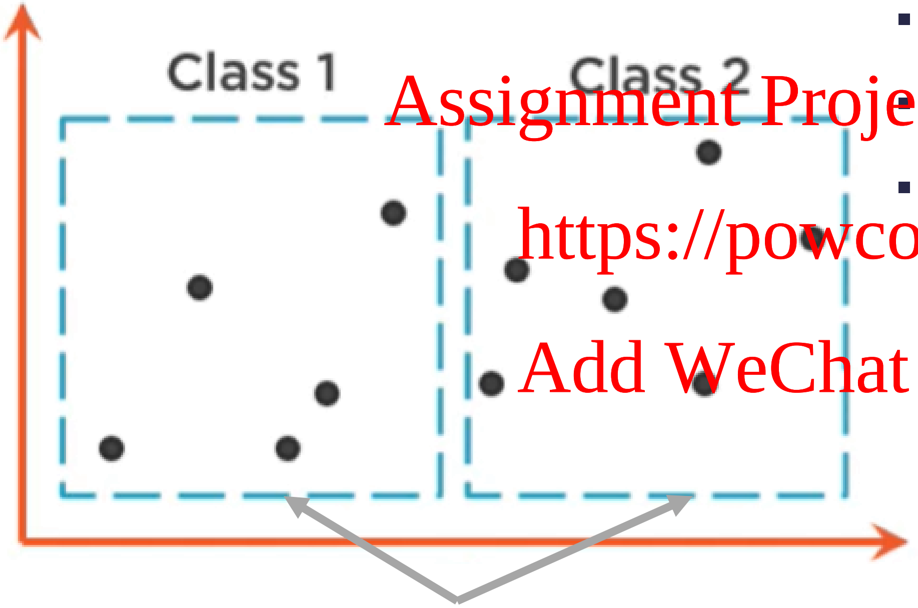
Machine Learning Models

Machine Learning Models: Regression



- For supervised learning
- Predict continuous numeric values
- Example questions
  - How many units of this product will we sell next month?
  - Given past stock data, what is the price tomorrow?
  - Given the characteristics of a car, what is the likely mileage?
  - Given location and attributes of a house, what is the price?

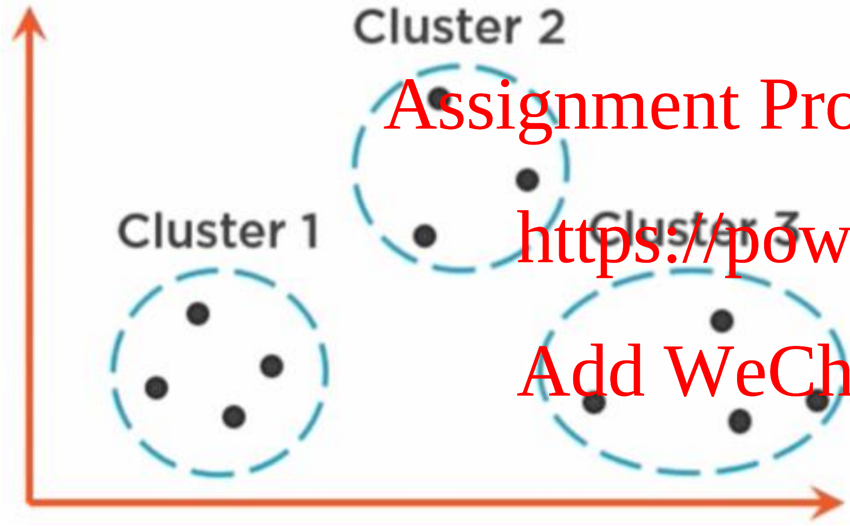
Machine Learning Models: Classification



Can be more than two classes

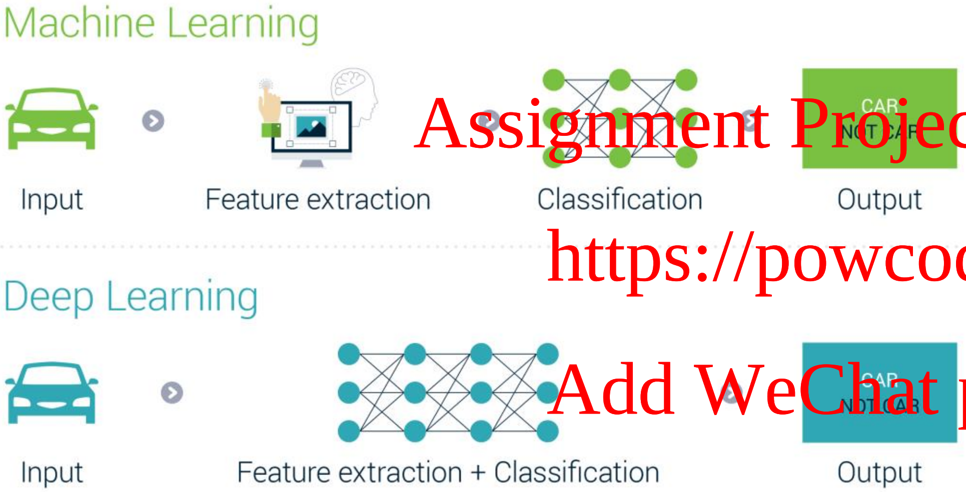
- For supervised learning
- Predict for discrete categories
- Example questions
  - Is this credit card transaction fraudulent?
  - Is this email message spam or ham?
  - Should I buy, sell, or hold this stock?
  - Is it a cat, dog, or mouse?
  - Is the customer’s sentiment positive, negative, or neutral?

Machine Learning Models: Clustering



- For unsupervised learning
- Self discovery of patterns and groupings in data
- Example questions
  - **Document discovery**: find all documents related to homicide cases
  - **Social media advertisement targeting**: find all users who are interested in family office policies

Deep learning uses the same algorithms but different architectures to solve different problems



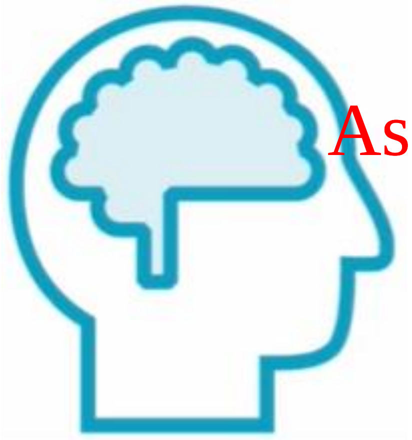
Images: Convolutional Neural Networks

Time series: Recurrent Neural Networks

- Deep learning models extracts the right features from the data by themselves
  - There are several layers involved in the neural network
  - Each layer extracts different features from the data
- Not that different neural networks have different algorithms
- The differences lie in the architecture of the neural network

The Machine Learning Process

The first problem is asking the right question



Right Question

What do you really care about?

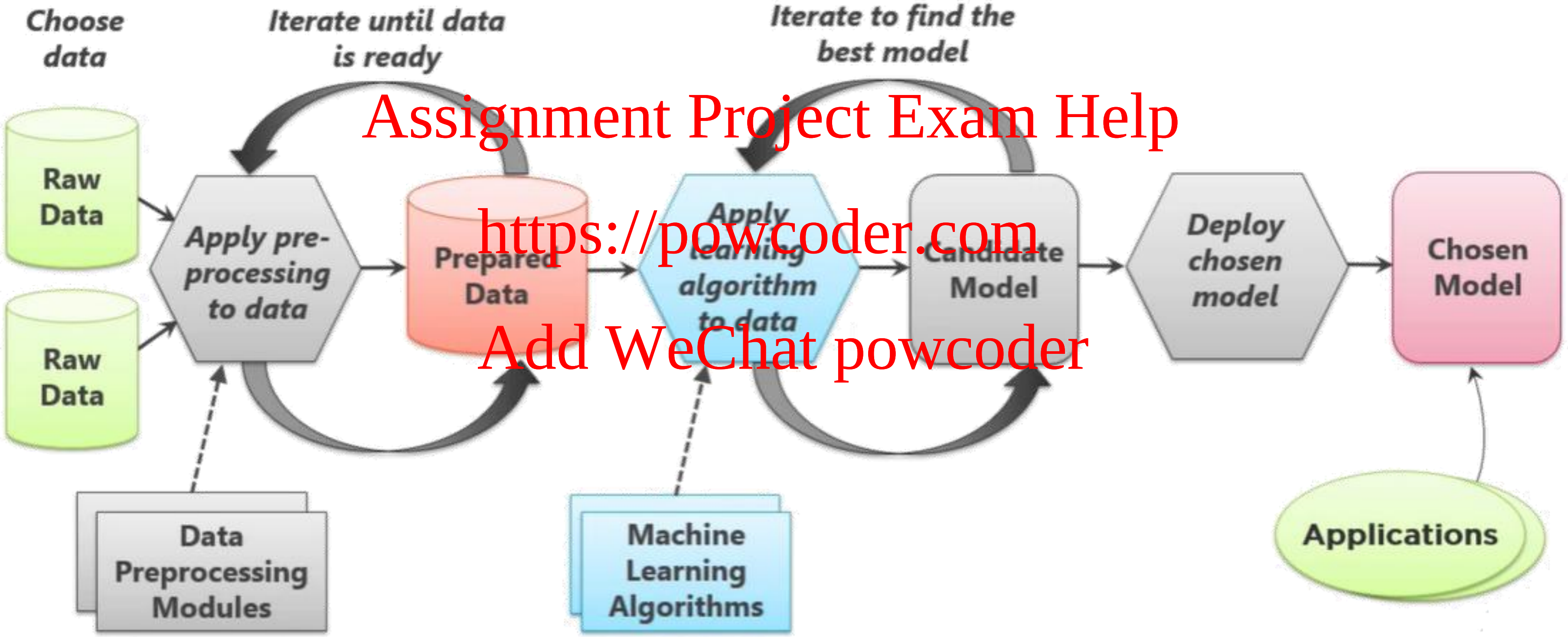
Relevant Data

Do you have the relevant data to answer the question?

Measure of Success

Do you know how you will measure success?

Understanding machine learning is about understanding the machine learning process



More often than not machine learning deals with labelled financial data

Training Data

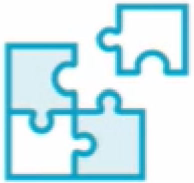
The prepared data used to create a model

Creating a model is called training a model

Supervised Learning

The value you want to predict is in the training data

The data is labeled

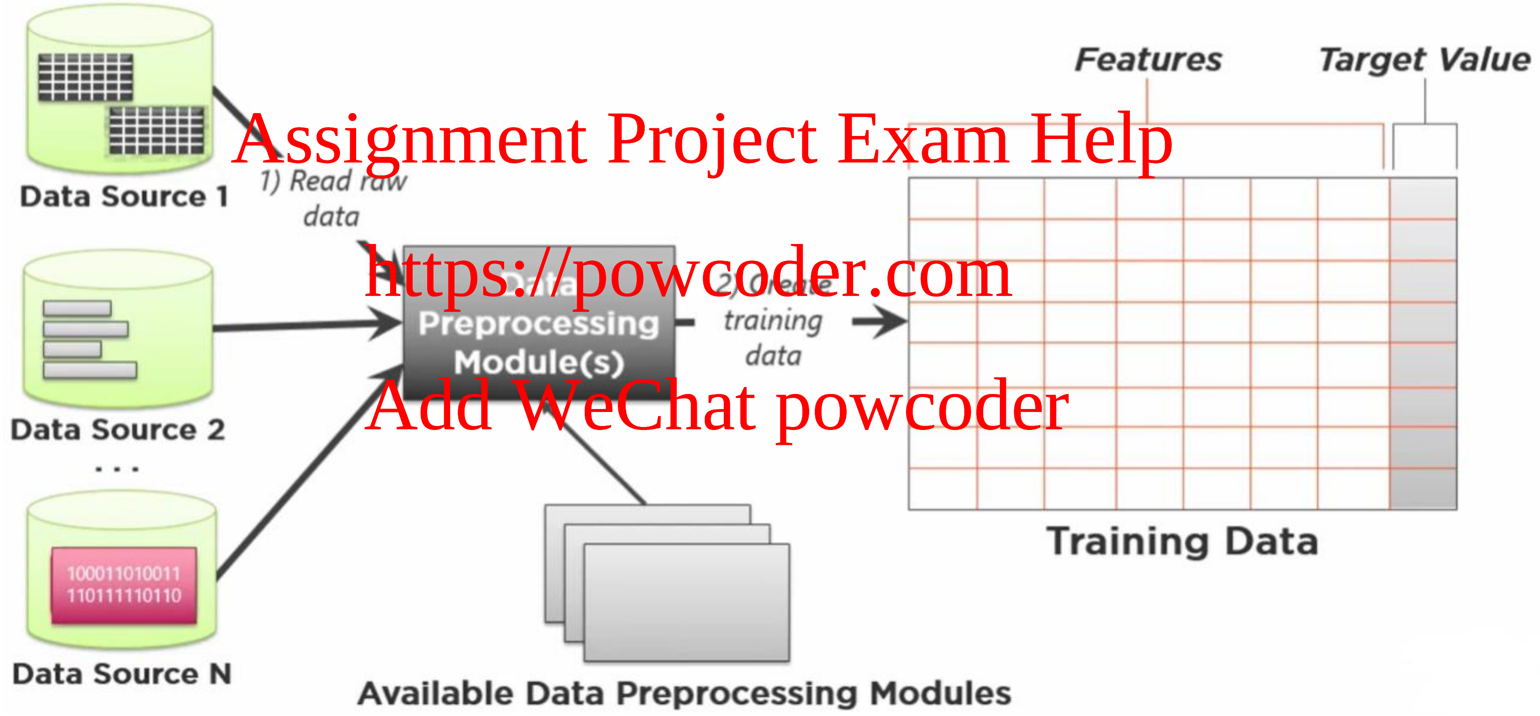


Unsupervised Learning

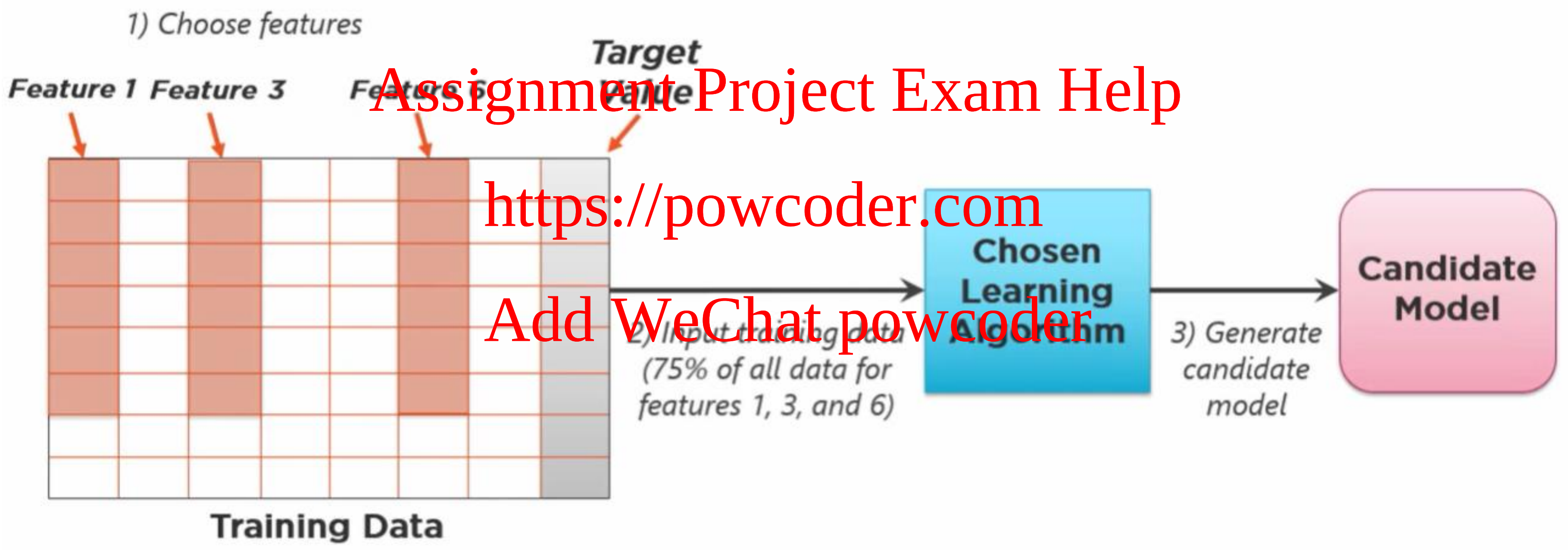
The value you want to predict is not in the training data

The data is **un**labeled

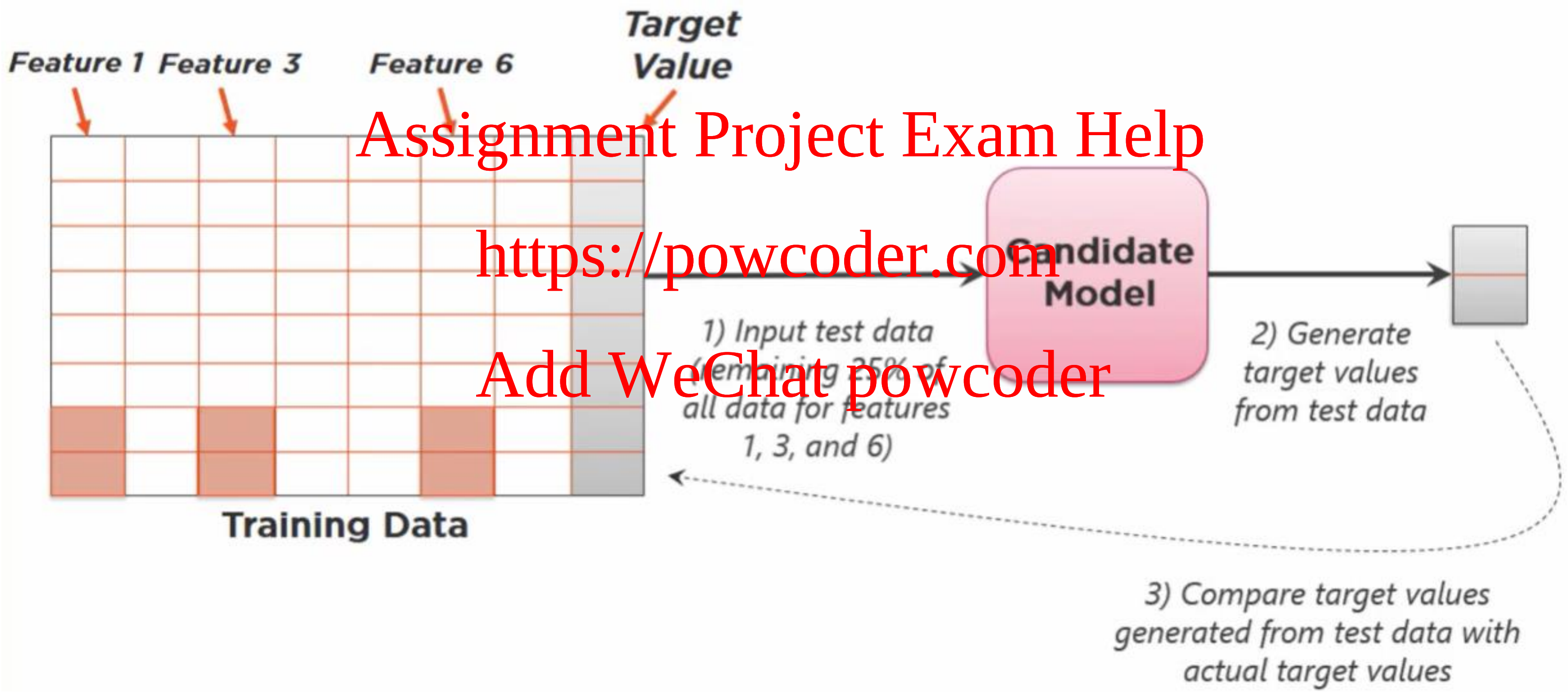
Data are pre-processed to make ready for model training and to optimise the performance of the model to be trained



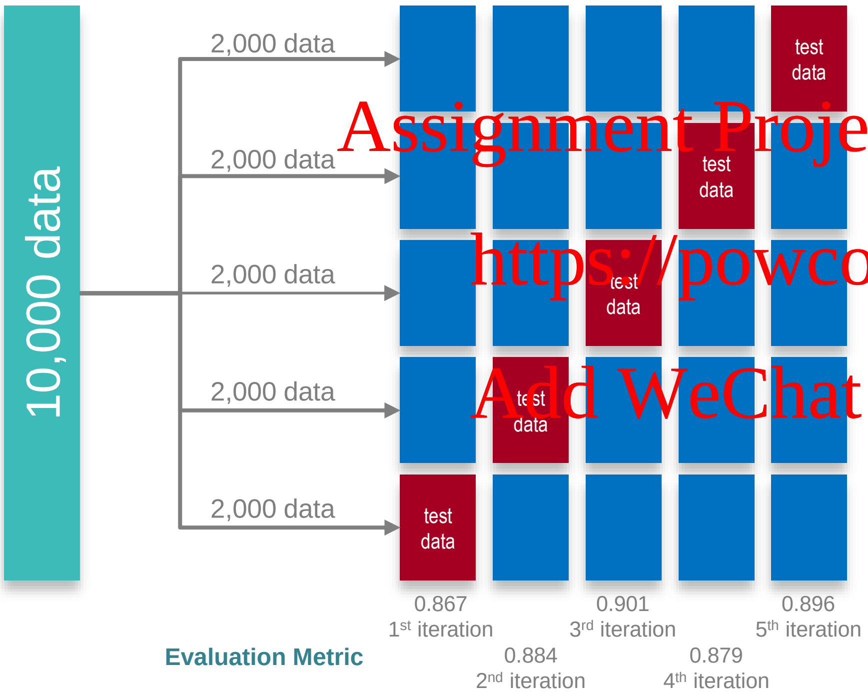
Feature engineering is performed and holdout data is split into two portions: one to train and one to test the model



The remaining (25%) data in the holdout dataset is used to validate the model trained using 75% of the data

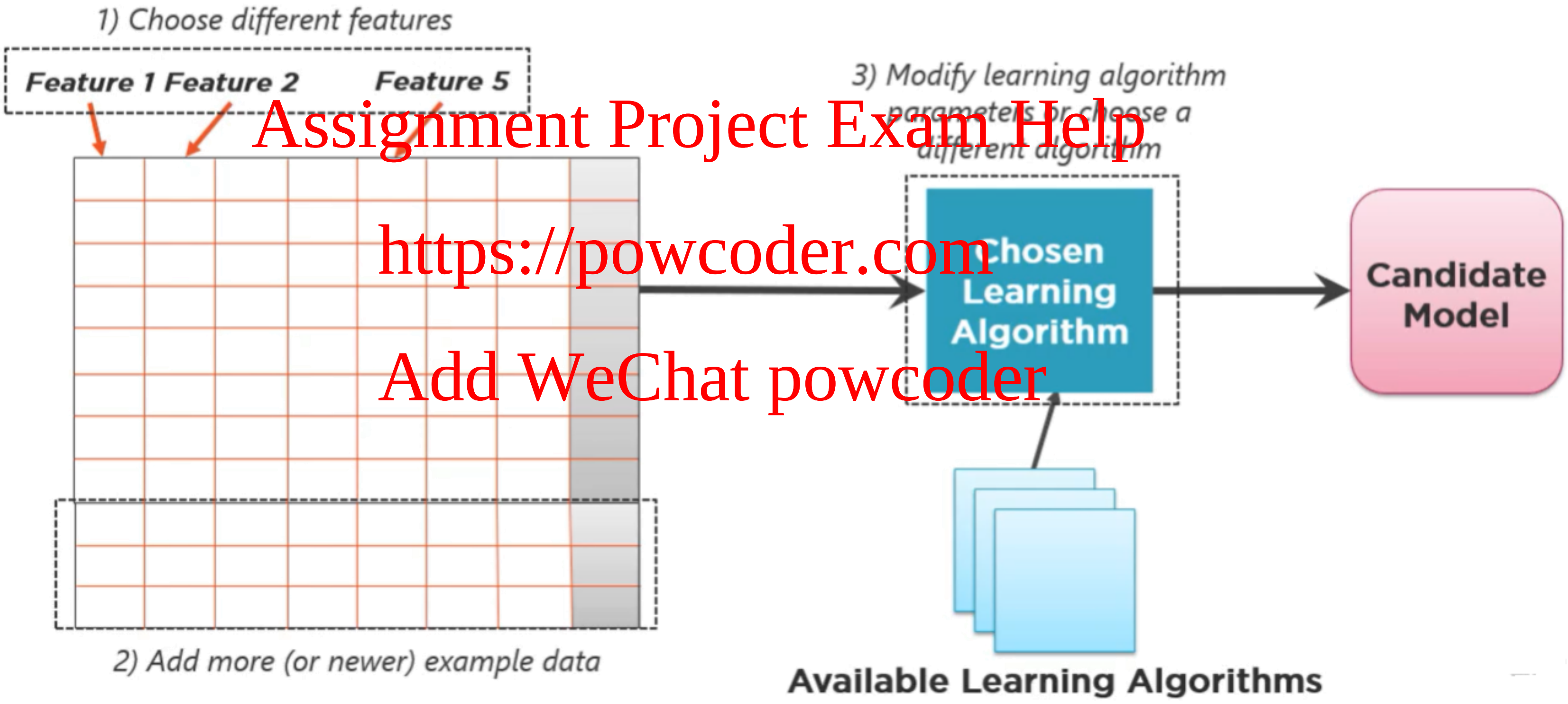


All data in the hold-out dataset can be used for both training and testing through k-fold cross-validation



- Data is divided into k subsets and the holdout method is repeated k times
- Each time, one of the k subsets is used as the test set and the other k-1 subsets are combined to be used to train the model

Model performance can be further improved through investigating the columns and rows in the dataset



Candidate models can be fine-tuned using hyperparameter optimization: random search and grid search

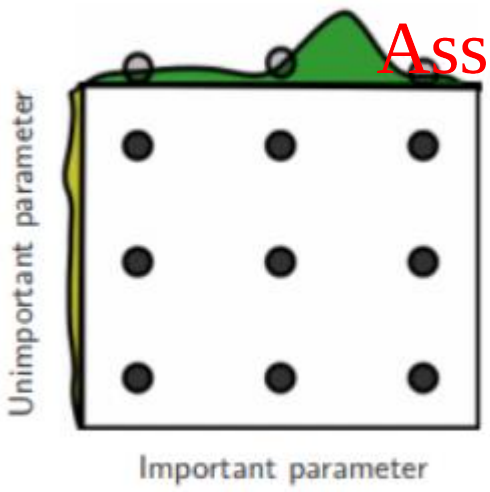
- Hyperparameters are specified parameters that can control a machine learning algorithm's behavior by tuning

- They are different from model parameters in that hyperparameters are parameters set and supplied to the model before training while model parameters are values that are learnt during training by the machine

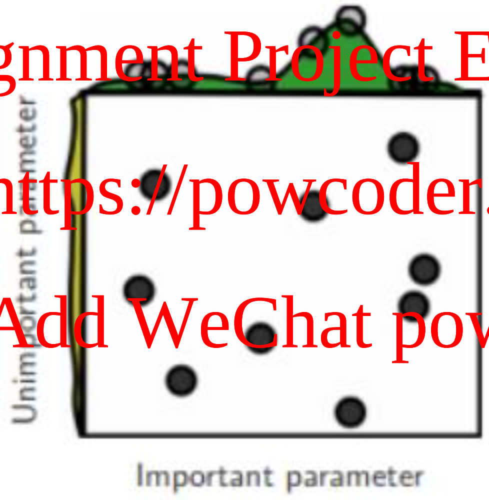
- Different models are tested and hyperparameters are tuned to get better predictions

- There are tools available to optimize hyperparameters: Random Search and Grid Search
- These two methods make the process of hyperparameter optimization easier as they sort through different combinations of parameters and hyperparameters to output the best combination of values

Random search is preferred for hyperparameter tuning if searching space is high as it has lesser time complexity



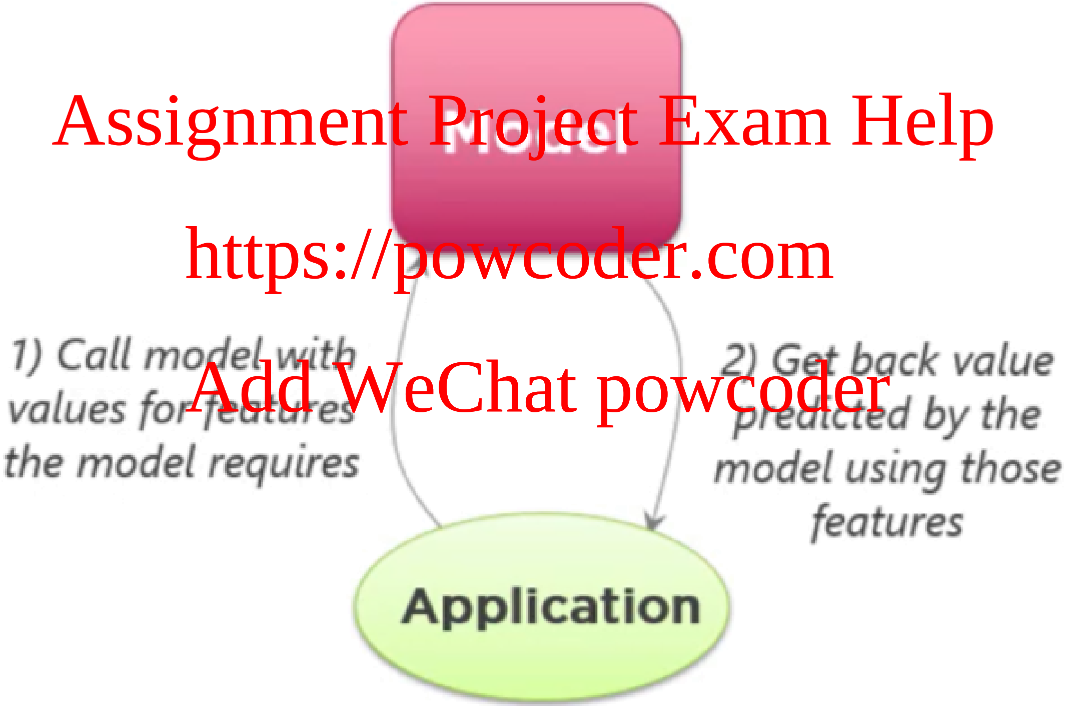
9 trials only test 3 distinct values of the important hyperparameter



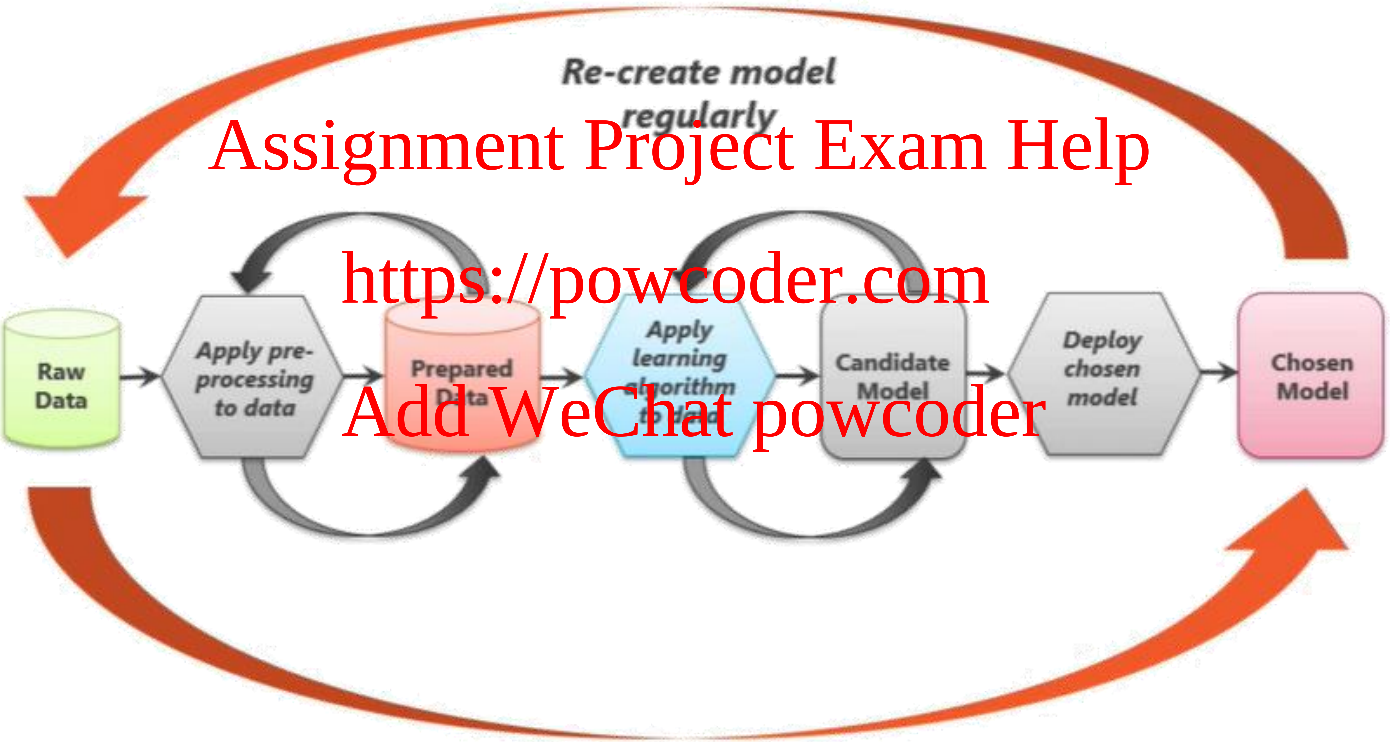
9 trials explore different distinct values of the important hyperparameter

- Random Search is the preferred approach when there are many parameters
  - The searching space is high meaning that there are more than 3 dimensions
  - Often outperforms Grid Search
- Grid Search performs an exhaustive search looking through all the combinations of specified hyperparameters
  - Can be very computationally expensive

A model can be deployed into production by calling the model from an application



The machine learning process is iterative in nature



References

References

"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", 2nd edition

Aurélien Géron

O'Reilly Media, October 15, 2019

ISBN-10 : 1492032646

ISBN-13 : 978-1492032649

Chapter 1

References

- "Machine Learning Glossary" (https://developers.google.com/machine-learning/glossary?utm_source=google-ai&utm_medium=card-image&utm_campaign=training-hub&utm_content=ml-glossary)

THANK YOU