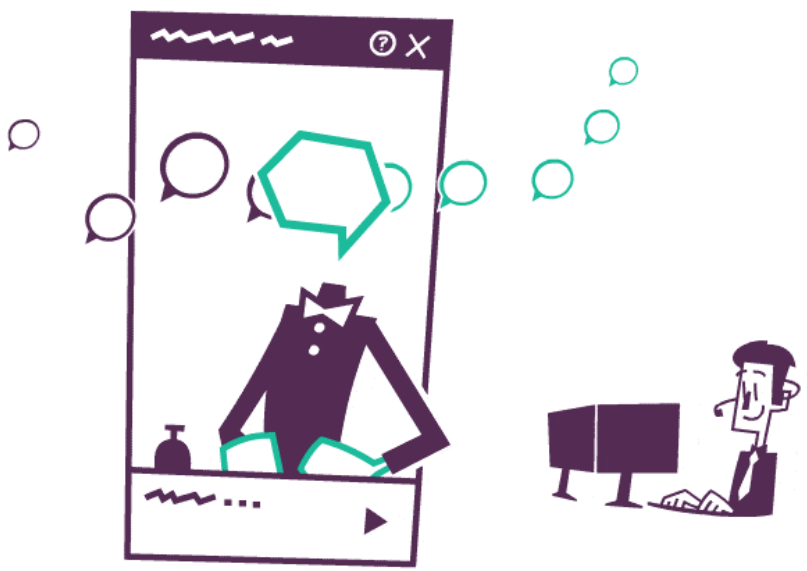


Machine Learning (ML)

Chapter I
Introduction to ML

Prof. Dr. Ibrahim Hamarash
Salahaddin University-Erbil

Example: A toy problem

Teaching a computer how to distinguish between pictures of **cats** from those with **dogs**.

Do you recall how you first learned about the difference between cats and dogs, and how they are different animals?

The answer is probably **NO**, as most humans learn to perform simple cognitive tasks like this very early on in the course of their lives.

One thing is certain, however: young children do not need some kind of formal scientific training, or a zoological lecture on felis catus and canis familiaris species, in order to be able to tell cats and dogs apart. **Instead, they learn by example**. They are naturally presented with many images of what they are told by a supervisor (a parent, a caregiver, etc.) are either cats or dogs, until they fully grasp the two concepts.

J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: a CAPTCHA that exploits interest-aligned manual image categorization," Proceedings of ACM Conference on Computer and Communications Security, pp. 366–374, 2007.

How do we know when a child can successfully distinguish between cats and dogs?

Intuitively, when they encounter new (images of) cats and dogs, and can correctly identify each new example.

or,

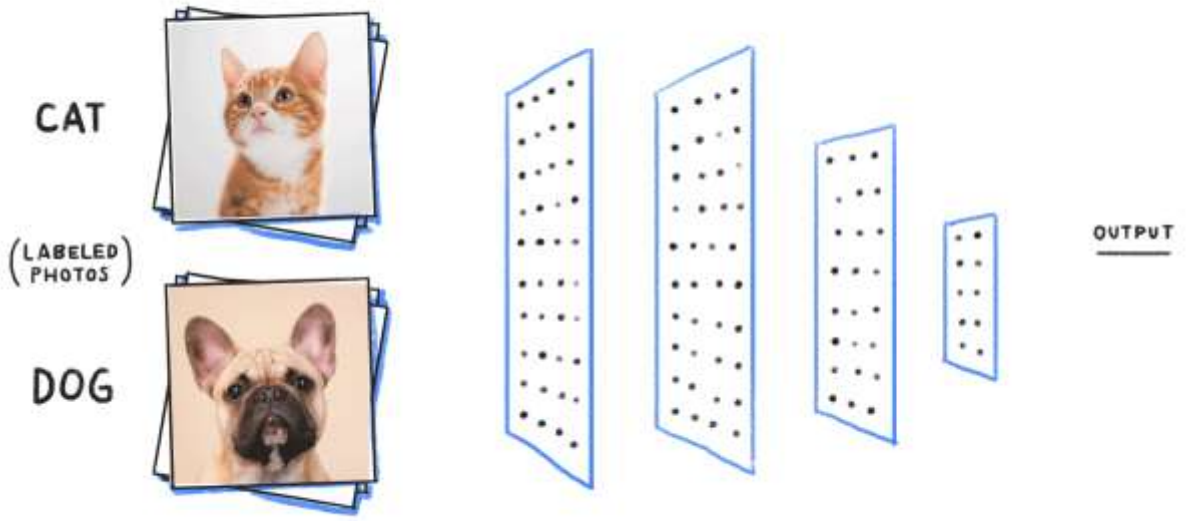
When they can generalize what they have learned to new, previously unseen, examples.

REMARK

Like human beings, computers can be taught how to perform this sort of task in a similar manner.

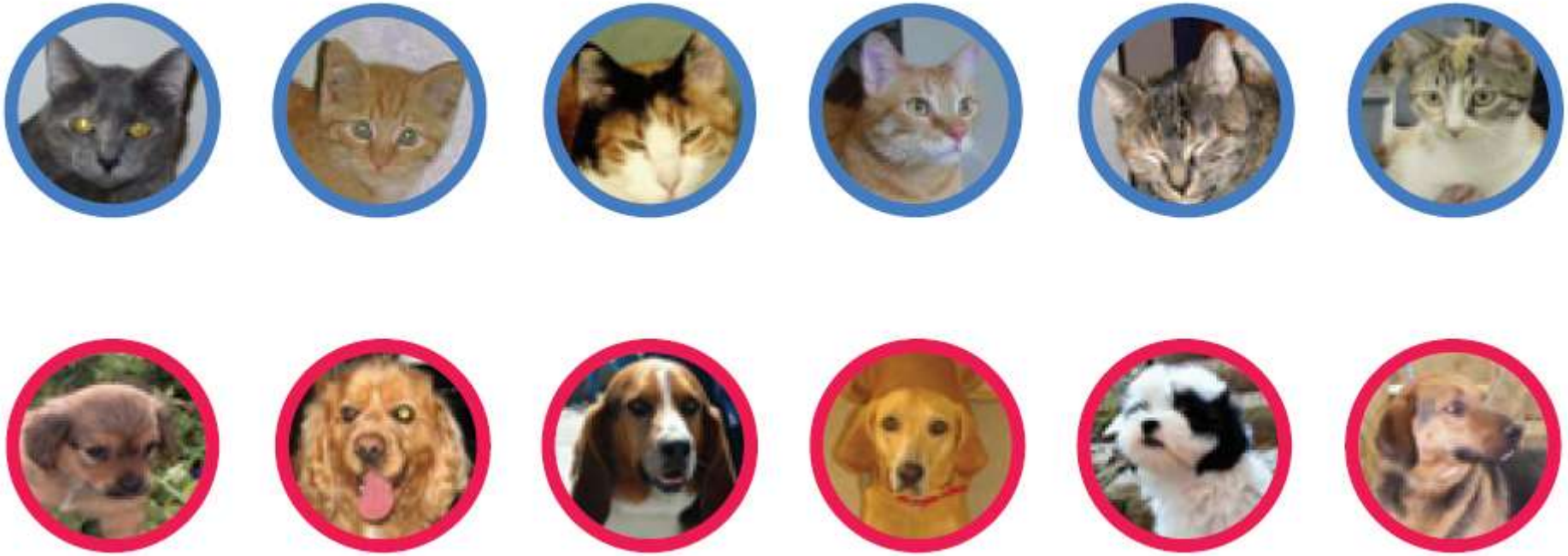
This kind of task where we aim to teach a computer to distinguish between different types or classes of things (here cats and dogs) is referred to as a **classification problem** in the jargon of machine learning, and is done through a series of steps:

1. Data collection
2. feature design
3. Model training
4. Model validation



1. **Data collection**

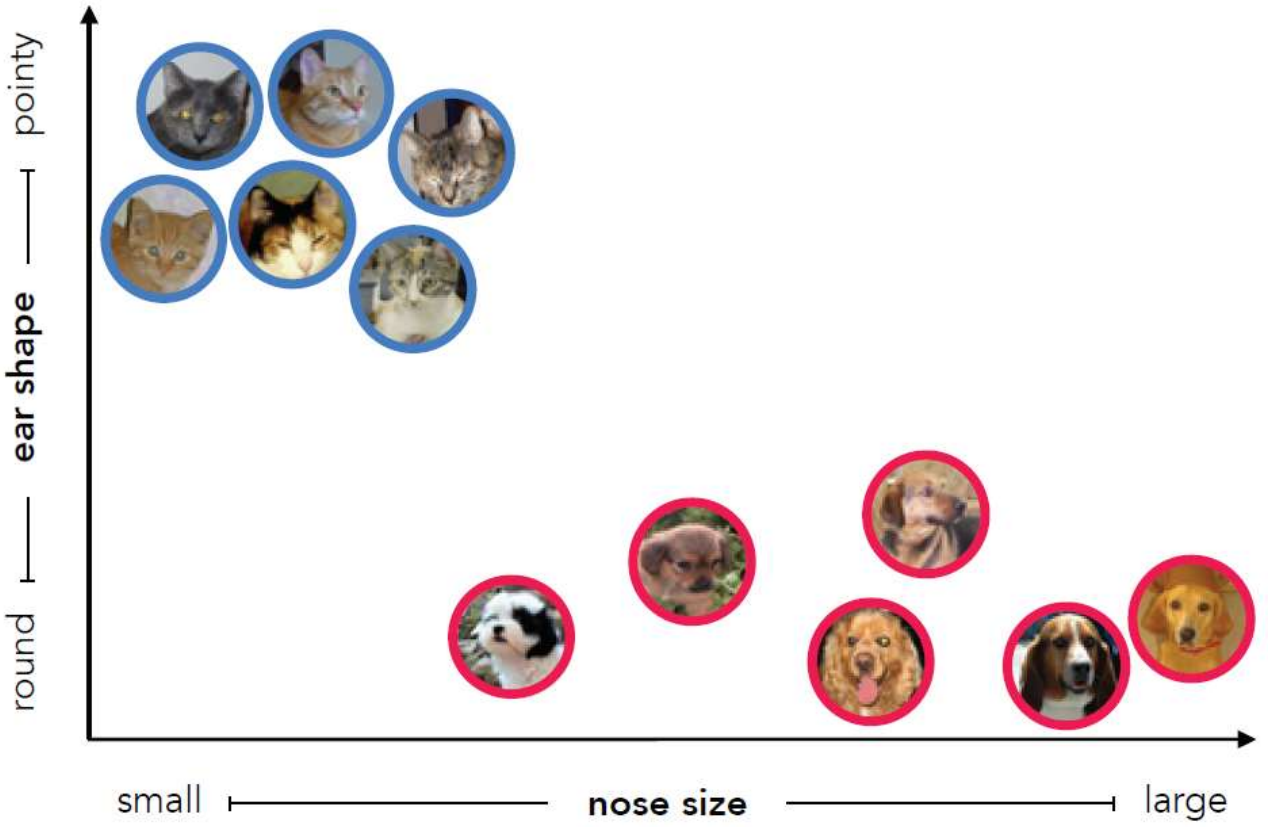
Like human beings, a computer must be trained to recognize the difference between these two types of animals by learning from a batch of examples, typically referred to as a **training set of data**.



2. Feature design

Think for a moment about how we (humans) tell the difference between images containing cats from those containing dogs.

We use color, size, the shape of the ears or nose, and/or some combination of these **features** in order to distinguish between the two.



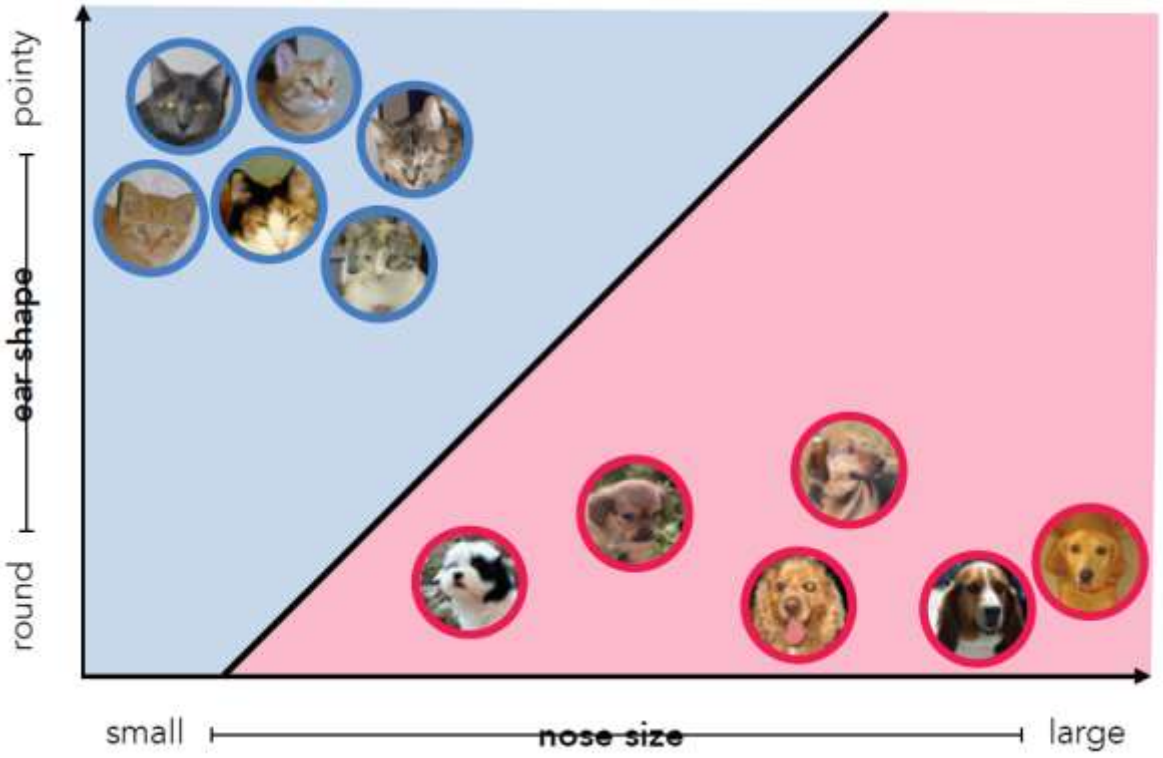
3. Model training

With our feature representation of the training data the machine learning problem of distinguishing between cats and dogs is now a simple geometric one:

have the machine find a line or a curve that separates the cats from the dogs in our carefully designed feature space?

Supposing for simplicity that we use a line, we must find the right values for its two parameters – a slope and vertical intercept – that define the line's orientation in the feature space.

The process of determining proper parameters relies on a set of tools known as **mathematical optimization.**



4. Model validation

To validate the efficacy of our trained learner we now show the computer a batch of previously unseen images of cats and dogs, referred to generally as a **validation set of data**, and see how well it can identify the animal in each image.

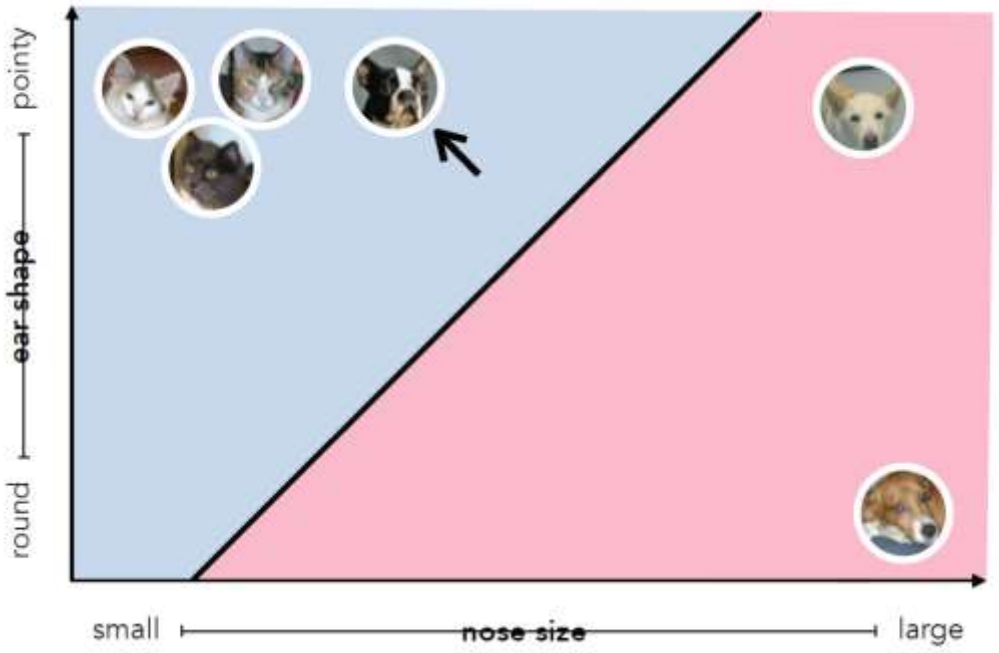
We show a sample validation set for the problem at hand, consisting of three new cat and dog images. To do this, we take each new image, extract our designed features (i.e., nose size and ear shape), and simply check which side of our line (or classifier) the feature representation falls on.

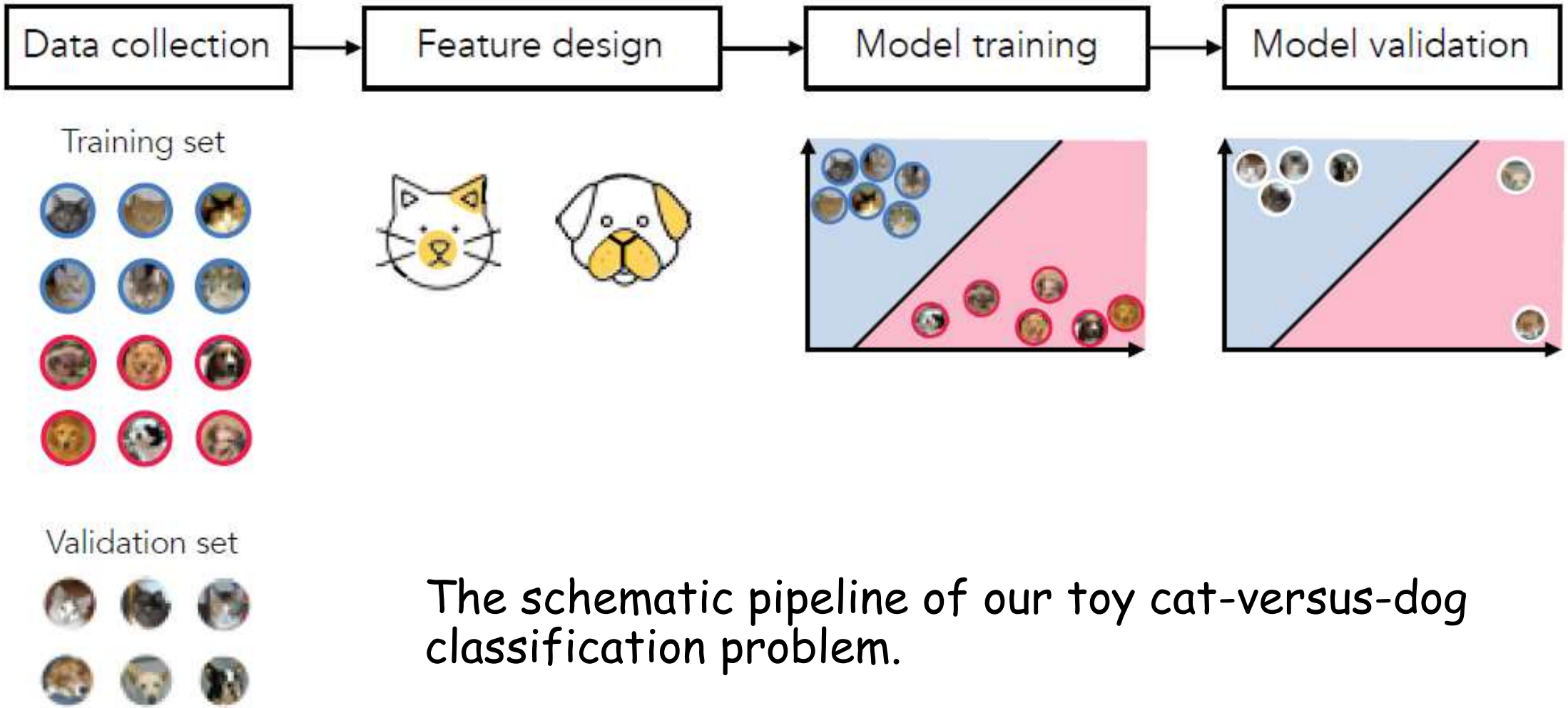
REMARK

In this instance, as can be seen in here, all of the new cats and all but one dog from the validation set have been identified correctly by our trained model.

The misidentification of the single dog is largely the result of our choice of features, which we designed based on the training set and to some extent our decision to use a linear model (instead of a nonlinear one).

This dog has been misidentified simply because its features, a small nose and pointy ears, match those of the cats from our training set. Therefore, while it first appeared that a combination of nose size and ear shape could indeed distinguish cats from dogs, we now see through validation that our training set was perhaps too small and not diverse enough for this choice of features to be completely effective in general.

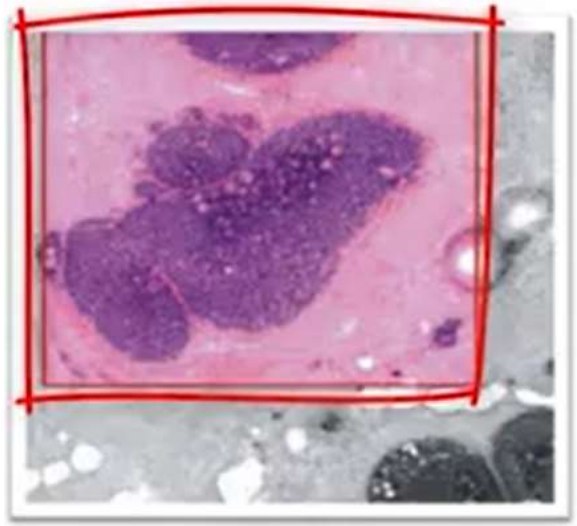




The schematic pipeline of our toy cat-versus-dog classification problem.

Example: Cell Type Prediction

Is this benign or malignant cell?



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

Ex. Cell Type Prediction, cont.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

Ex. Cell Type Prediction, cont.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Modeling

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	Benign

Prediction

Accuracy = 89%

Example: Recommendation: Personalization, Netflix perspective

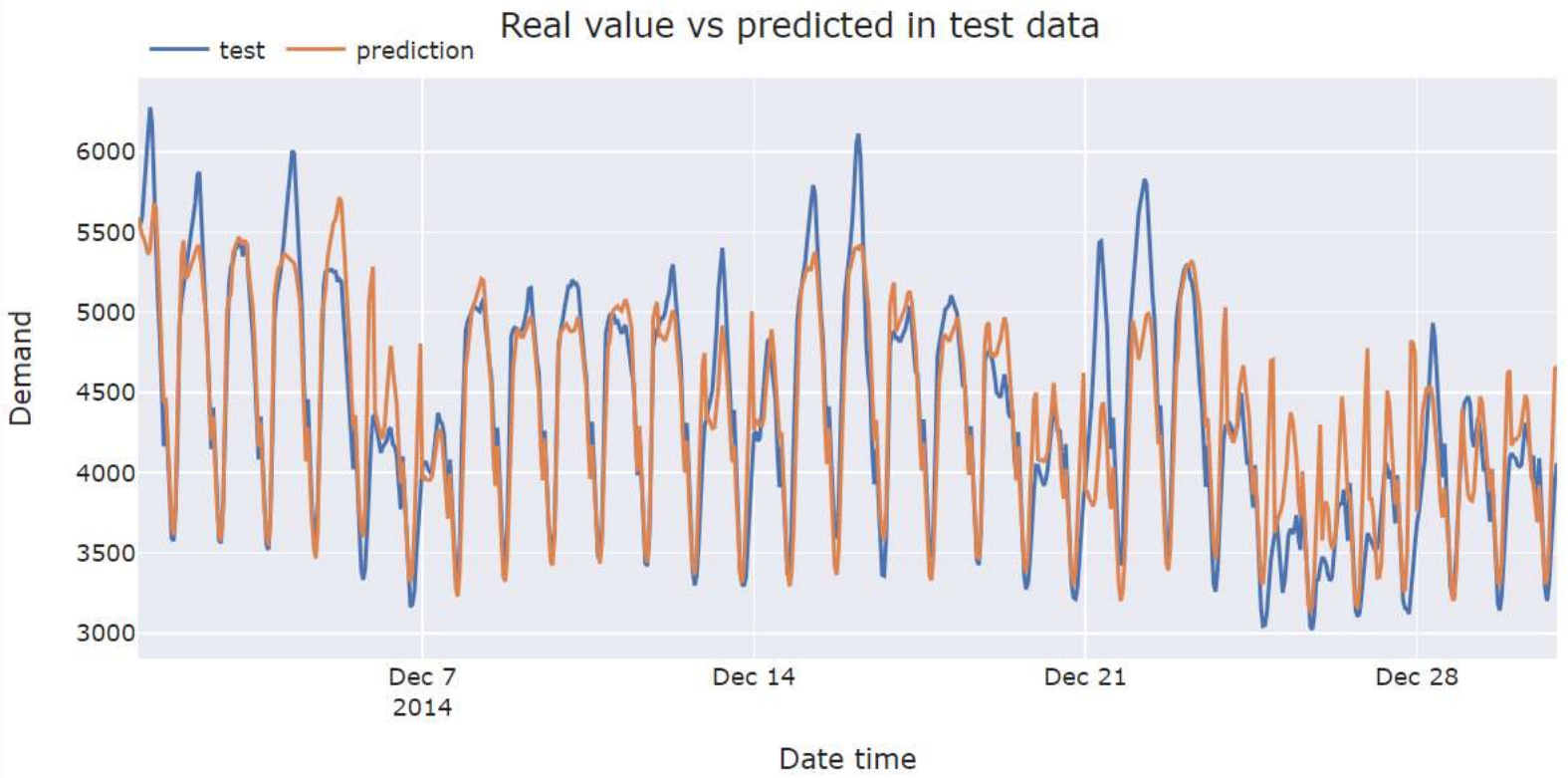
The Netflix Prize

In 2000, Netflix introduced personalized movie recommendations and in 2006, launched **Netflix Prize**, a machine learning and data mining competition with a $1 million dollar prize money.

Machine Learning Approach

The solution and approach that Netflix uses is a Machine Learning one, where they aim to create a scoring function by training a model using historical information of which homepages they have created for their members — including what they actually see, how they interacted with and what they played.

Example: Daily Electricity Demand Forecast

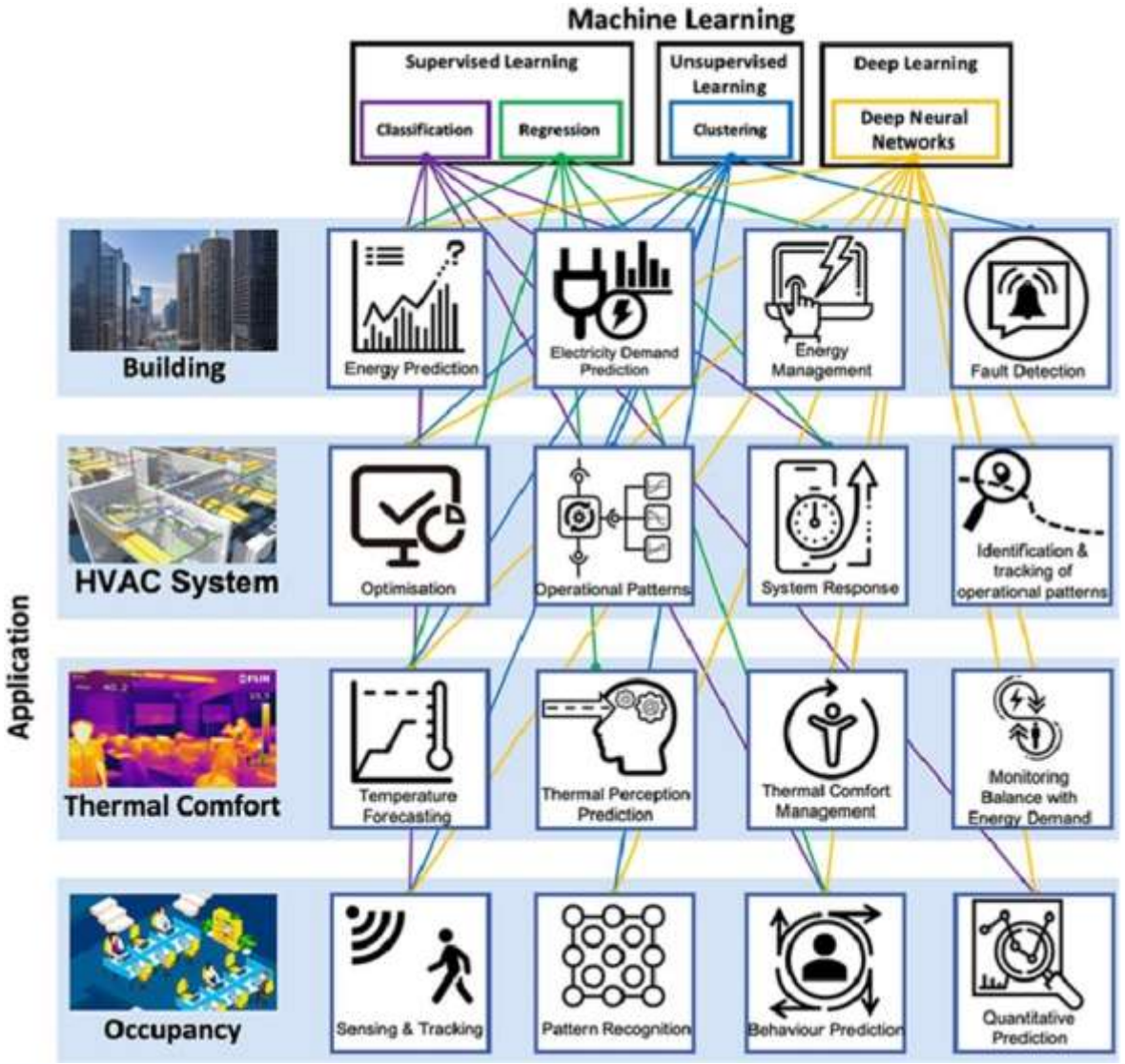


Variables:

- Time: date and time of the record.
- Date: date of the record.
- Temperature: temperature in Melbourne, the capital of Victoria.
- Holiday: indicates if the day is a public holiday.
- Demand: electricity demand (MW).

Example:

Enhancing Building Energy Efficiency



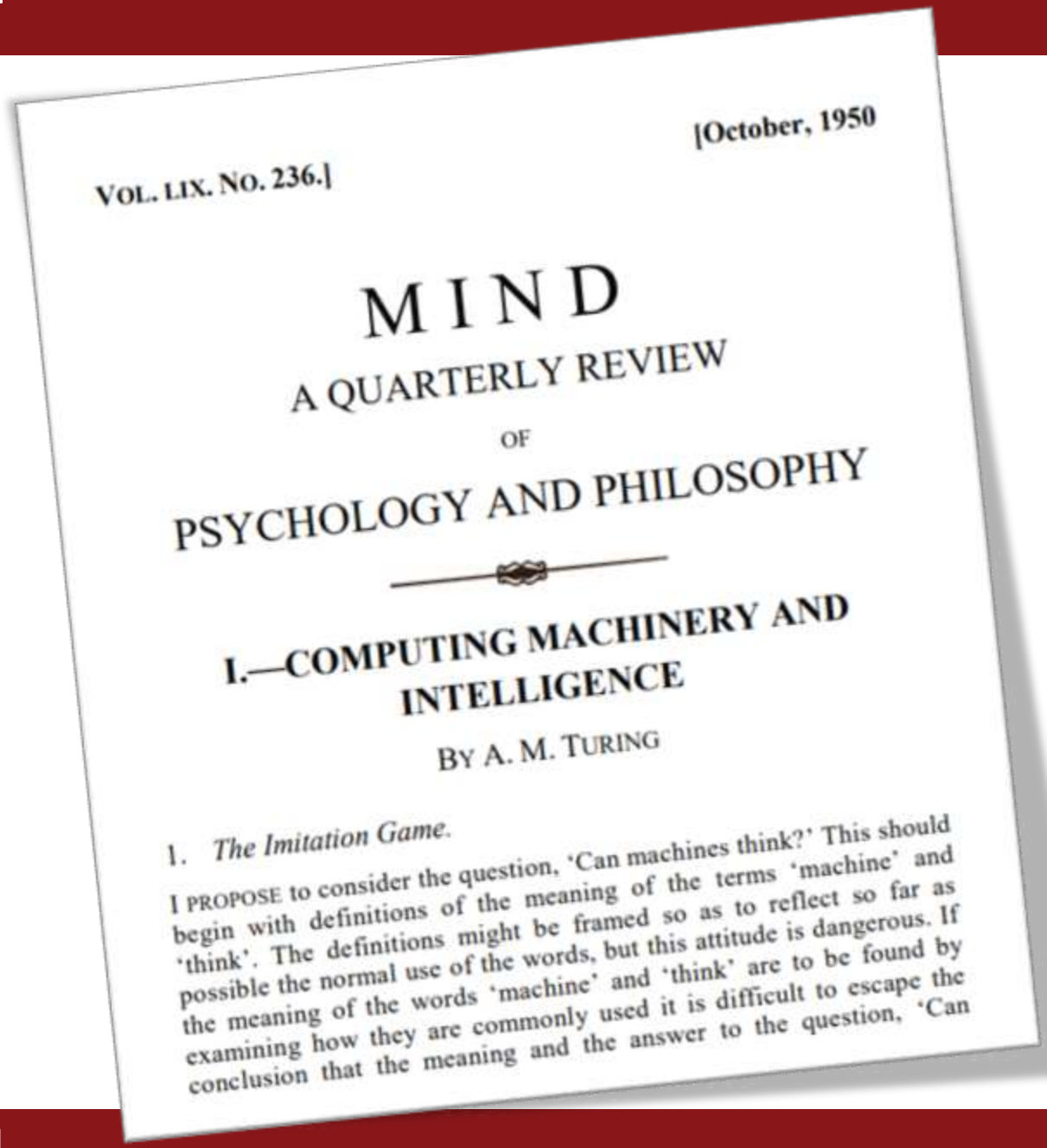
https://doi.org/10.1016/j.egyai.2022.100198

https://www.projectpro.io/article/applications-of-machine-learning-in-energy-sector/770

The first one?

Turing's paper considers the question "**Can machines think?**"

"What we want is a machine that can learn from experience"



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can

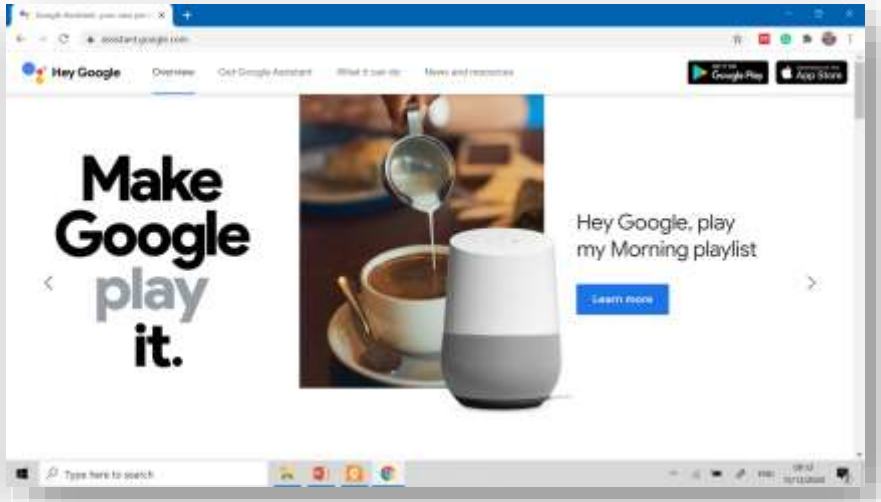
Arthur Samuel
IBM Laboratories

Coined the phrase "Machine learning" in 1952.

Wrote the first computer learning program to play the game of Checker

Machine Learning in daily life

Machine Learning in daily life, Context-Aware Systems



Machine Learning: Definition

ML is a subfield of computer Science that gives computers the ability to learn without being **explicitly** programmed.

Programming explicit rules

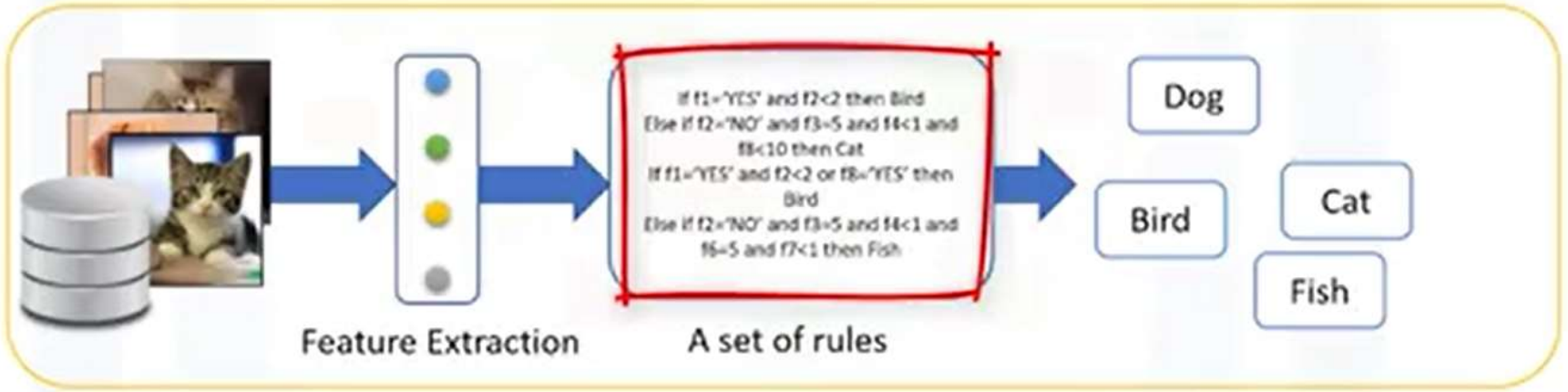
```
if(speed<4){
   status=WALKING;
}
```

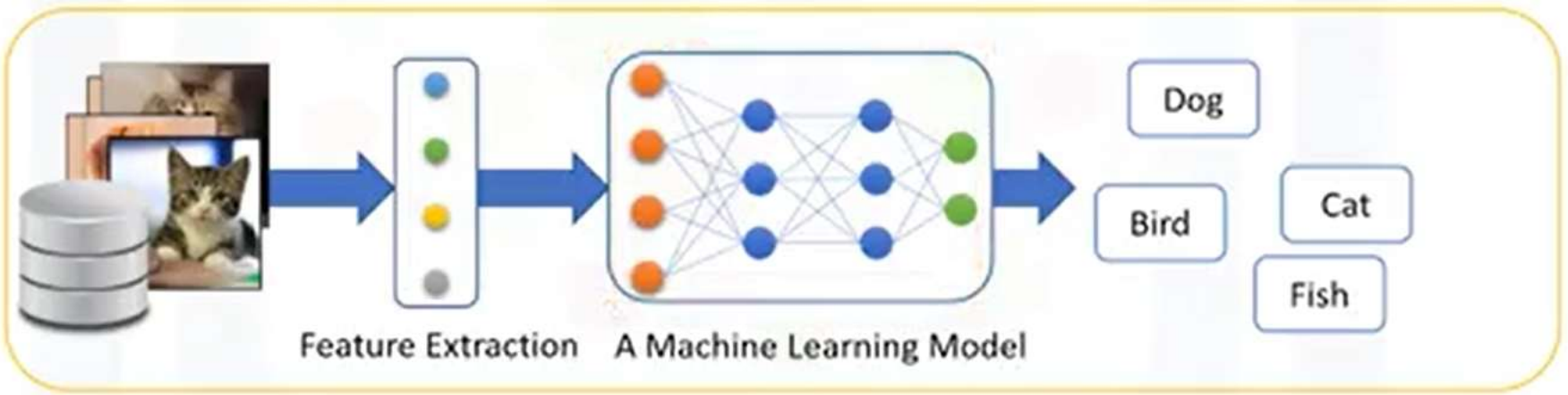
```
if(speed<4){
     status=WALKING;
} else {
     status=RUNNING;
}
```

```
if(speed<4){
     status=WALKING;
} else if(speed<12){
     status=RUNNING;
} else {
     status=BIKING;
}
```

```
// Now what?
```

Explicit and ML





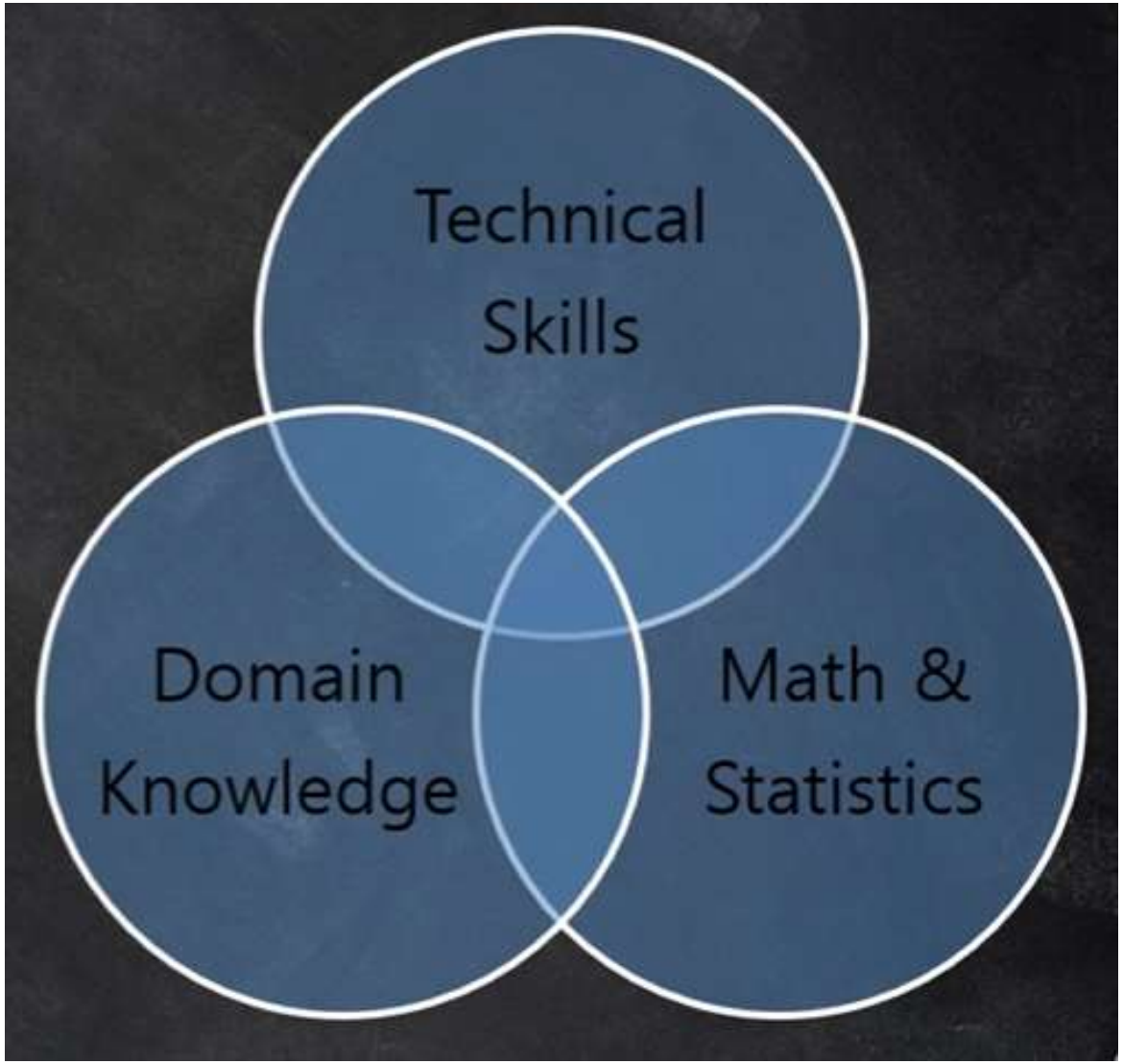
Machine Learning, Another definition

Machine learning is defined as a set of methods that can automatically **detect patterns in data**, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).

Related Terms to ML: Data Science

Is the process of obtaining, transforming, analyzing and communicating data to answer a question.

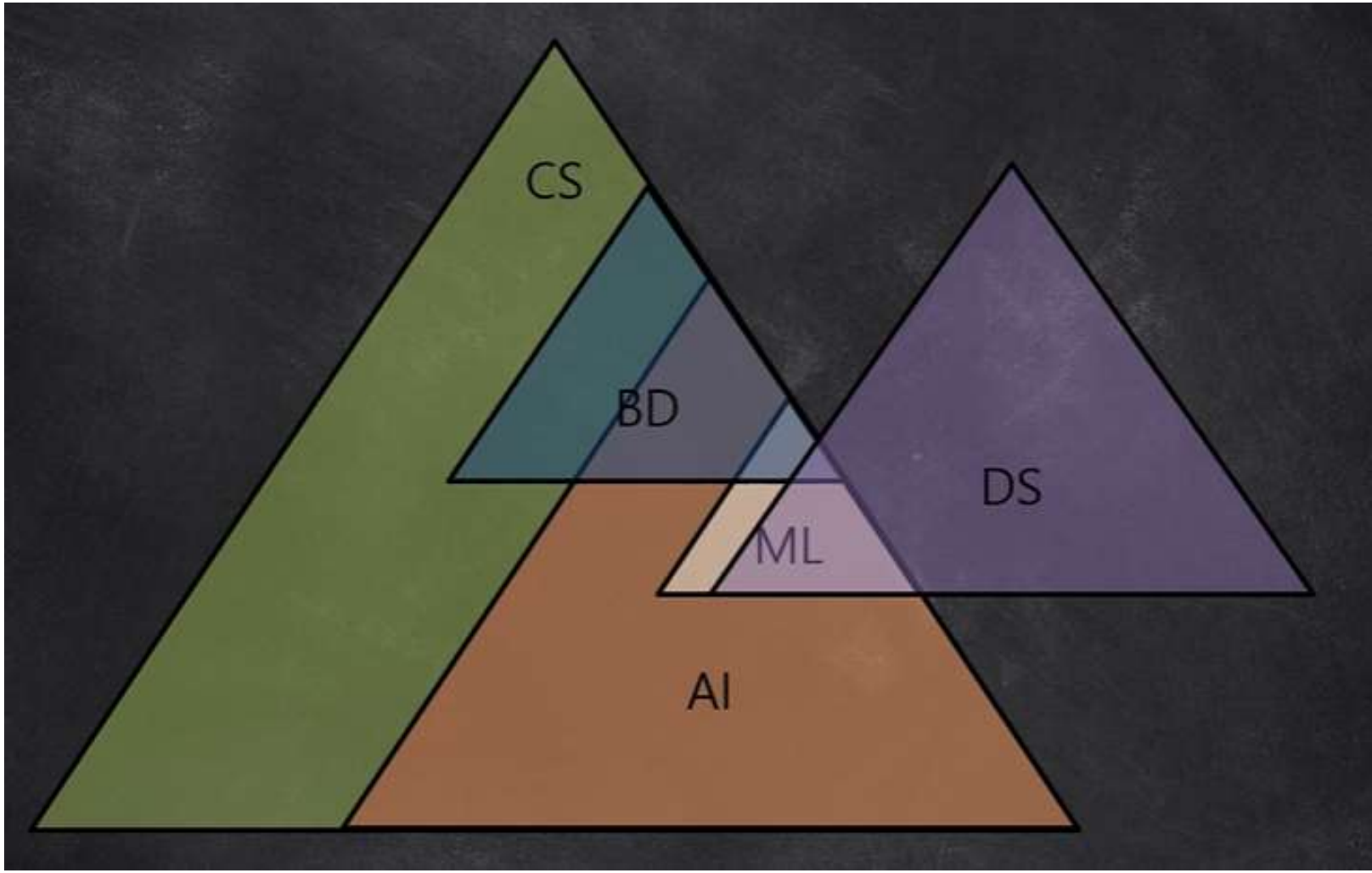
Johns Hopkins Data Science Specialization



The Backyard Data Scientist

Dave Valentine

http://www.tbdatascientist.com/

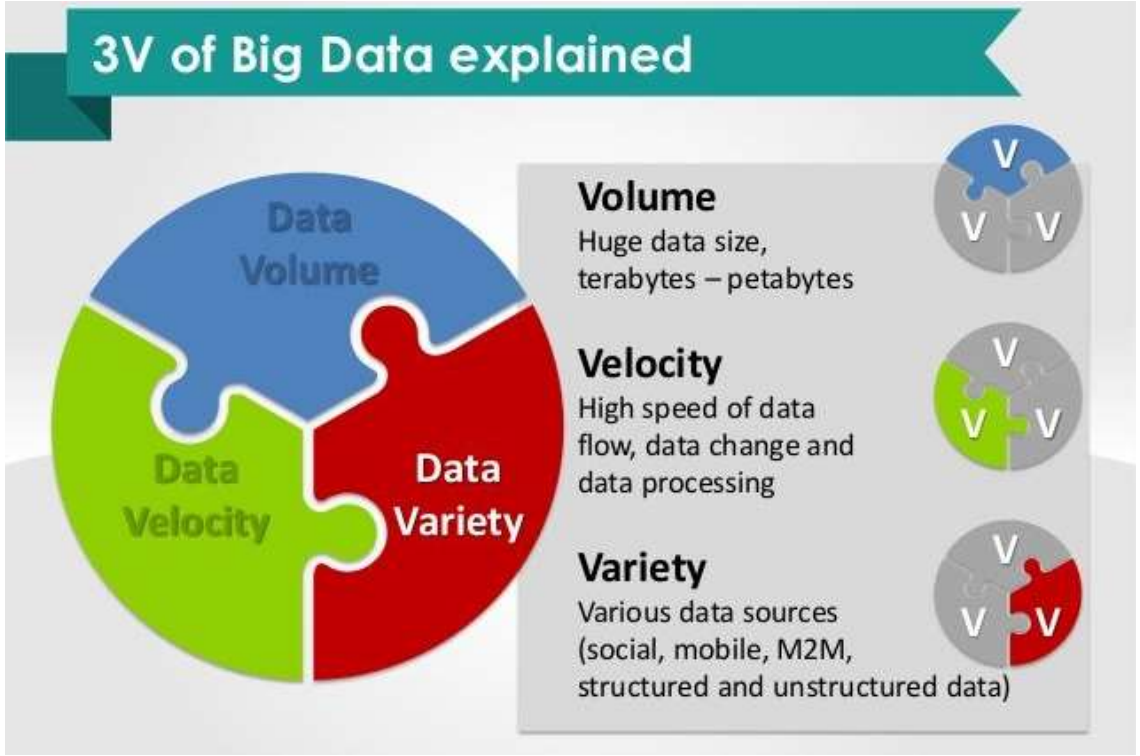


Terms related to ML: BIG DATA

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

www.gartner.com

A field within computer sciences that is concern with the storage, retrieval, transmission, and processing of extremely high volume, high velocity and high variety data.



Supervised and Unsupervised Learning

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

attributes

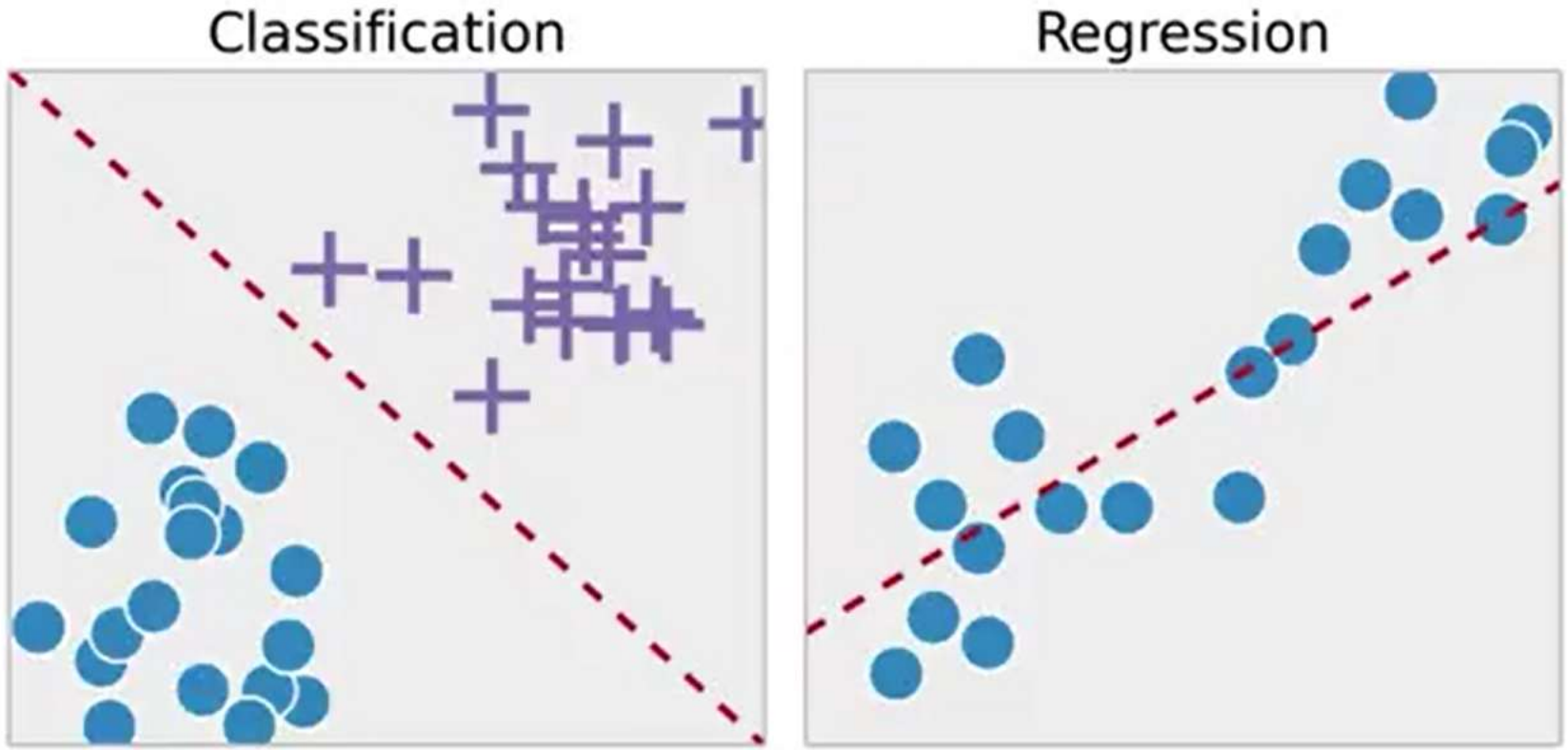
observation

numerical

feature

Category

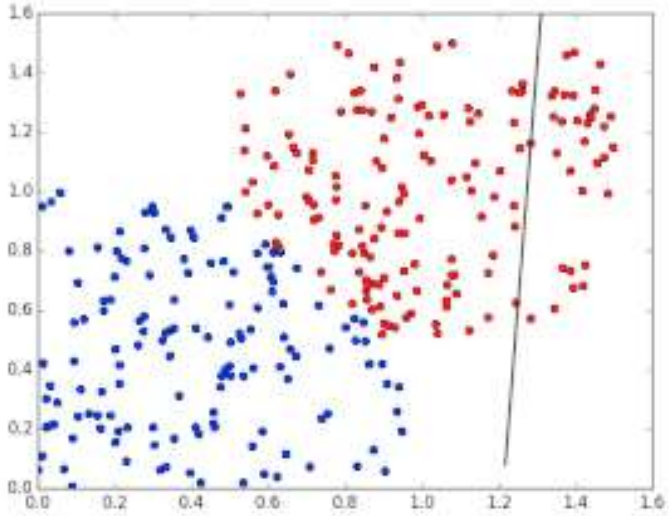
Supervised Techniques
Classification and Regression



Classification is the process of predicting discrete class labels or categories

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

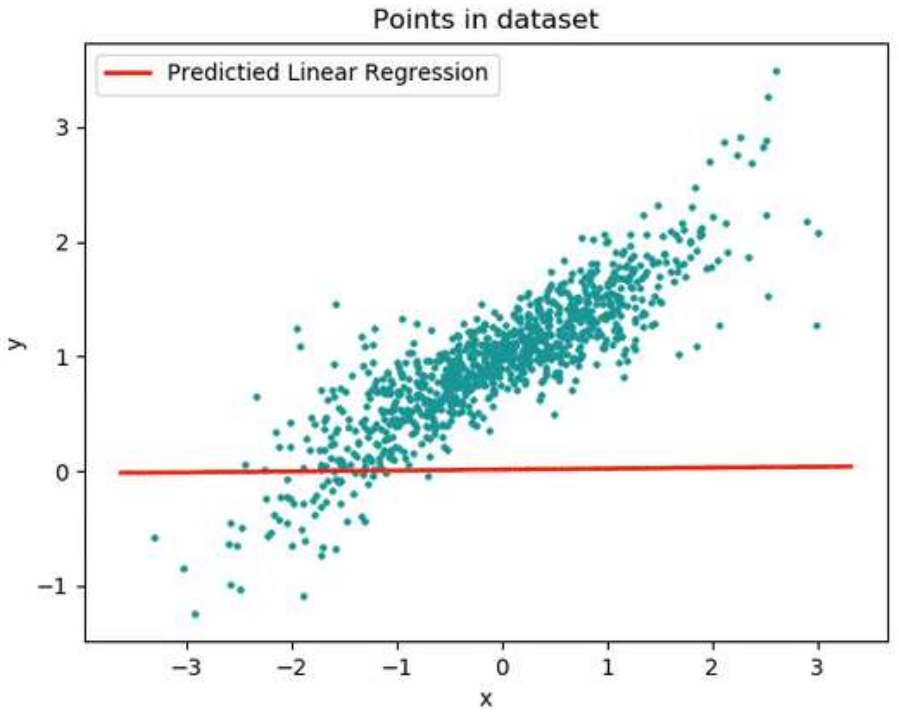
Categorical Values



Regression is the process of predicting continuous values

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values



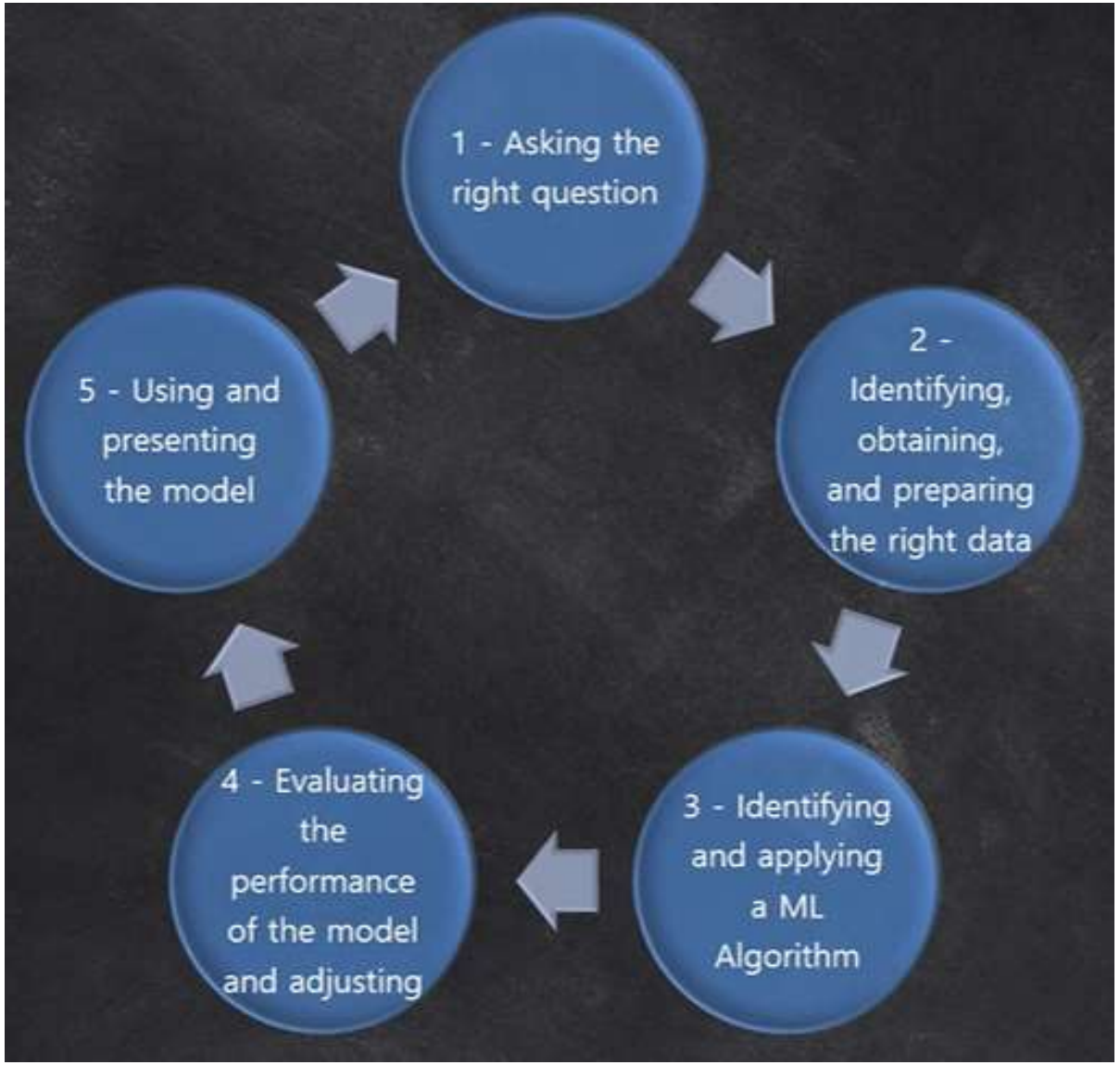
Unsupervised learning

The model works on its own to discover information.

Applications: Dimension reduction, Density estimation, Market basket analysis, clustering

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtincomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Machine Learning Five Steps Process



ML 5 Process, cont.

1. Asking the right question

- What question are we asking?
- What resources does addressing this question involve?
  People: Business, technical, analytical,
- What data sources will you use?
  Structured, unstructured, ..
- What rules are relevant to our question? (regulations, rules, ..)
- How will we measure the performance?
- How do we measure success? (False positive, False negative

ML 5 Process, cont.

2. Identifying, obtaining and preparing the right data

Identifying data
What data does your outcome depends on (Quality, speed, cost)

Obtaining data
(in house, third party, creating, public data)

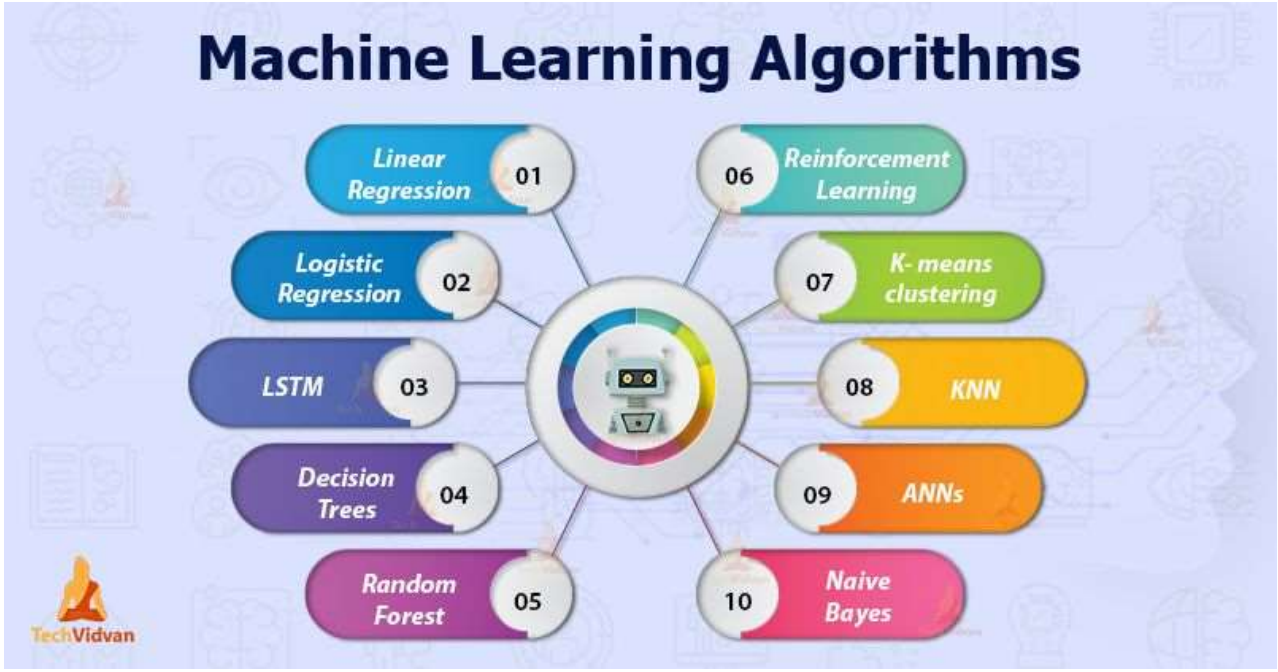
Preparing data
do not under estimate it.

ML 5 Process, cont.

3. Identifying and applying a ML algorithm

Decision trees
Naïve Bayesian Classifier
Neural Network
KNN

… and many more



Criteria
platform, cost, speed, ..

ML 5 Process, cont.

4. Evaluating the performance of the model and adjusting

Evaluating performance

Use the criteria you established in step #1
Test with hold-back data (Cross validation)

Adjusting Performance

- Pull in different data to get different features
- Clean it
- Tune the algorithm
- Adjust or tune the question

ML 5 Process, cont.

5. Using and presenting the model

- Before apply your model in your application, review it.
- Human judgement and oversight is always needed, in the interpretation of results of data science projects.

Common platforms and tools

Note: Despite the popularity of machine learning, the majority of ML scantiest work with PC size data, with PC remains the most popular platform.

- Ms Excel
- Matlab
- Python
- R
- SQL
- RapidMiner
- Weka
- …… etc.

Machine Learning Techniques

1. Regression: Predicting Continuous values
2. Classification: Predicting the item class, category of case
3. Clustering: Finding the structure of data, summarization
4. Association: Associating frequent co-occurring items/events
5. Anomaly detection: Discovering abnormal or unusual cases
6. Sequence mining: predicting next event
7. Dimension reduction: reducing the size of data
8. Recommendation systems: recommending items
9. Generative ML: image generation, text-to-image synthesis, data augmentation, and creative tasks

…… etc.

Final Remark

ML is like panning for gold

You have to have the right input
You have to have the right tool to extract it

Reading list

- *Jiang, H., (2021). Machine Learning Fundamentals, Cambridge Uni Press.*
- *Alexey G., (2021). Machine Learning Bookcamp, Manning Publisher.*
- *C. Sharmeela, et.al. 2023. IoT, Machine Learning and Blockchain Technologies for Renewable Energy and Modern Hybrid Power Systems, River Publishers.*
- *Osvaldo Simeone, 2023. Machine Learning for Engineers, Cambridge University Press.*
- *Kamal I.M Al Mala, 2023. Machine and Deep Learning Using MATLAB, Jon Wiley and Sons.*
- *Fenner, E.F., (2020). Machine Learning with Python for Everyone, Pearson-Addison-Wesly.*

Questions?