Machine Learning

NSC Workshop (November 2024)

Machine Learning

Natworpong Loyswai (Peng)
natworpong@shipsgp.com

Boonnithi Jiaramaneepinit (Up)
shiragap@shipsgp.com

Sakulthip Rassameecharerntham (Yok)
peony@shipsgp.com

Shipty

SHIPS シップス

Agenda (Part 1)

- Welcome and Introduction
- Setting Up Virtual Environment
- Installing and Exploring Jupyter Notebook
- Key Python Libraries for Machine Learning
- Lecture: Introduction to Machine Learning
- Lecture: Neural Networks
- Hands-On: Data Collection
- Hands-On: Model Training
- Hands-on: Evaluate and deploy the trained model
- Review, Q&A, and Closing Remarks

Agenda (Part 2)

- Welcome and Introduction
- Setting Up Virtual Environment
- Installing and Exploring Jupyter Notebook
- Key Python Libraries for Machine Learning
- Lecture: Introduction to Machine Learning
- Lecture: Neural Networks
- Hands-On: Data Collection
- Hands-On: Model Training
- Hands-on: Evaluate and deploy the trained model
- Review, Q&A, and Closing Remarks

Agenda (Part 3)

- Welcome and Introduction
- Setting Up Virtual Environment
- Installing and Exploring Jupyter Notebook
- Key Python Libraries for Machine Learning
- Lecture: Introduction to Machine Learning
- Lecture: Neural Networks
- Hands-On: Data Collection
- Hands-On: Model Training
- Hands-on: Evaluate and deploy the trained model
- Review, Q&A, and Closing Remarks

Setting Up Virtual Environment

Installing and Exploring Jupyter Notebook

Key Python Libraries for Machine Learning

Lecture: Introduction to Machine Learning

Kimchi - a UNESCO Cultural Heritage

Rising Prices of Kimchi

Reflect

What do you think are the factors affecting the prices of kimchi?

Basics of Systems Maps

1. Useful when there are multiple factors affecting an issue - we call these factors, “elements”
2. In a systems map, all elements are connected by relationships - we represent these by “arrowhead lines”
3. The loops indicate the direction and degree of feedback.
4. Changing elements or changing relationships between elements changes the system outcome

Basics of Systems Maps

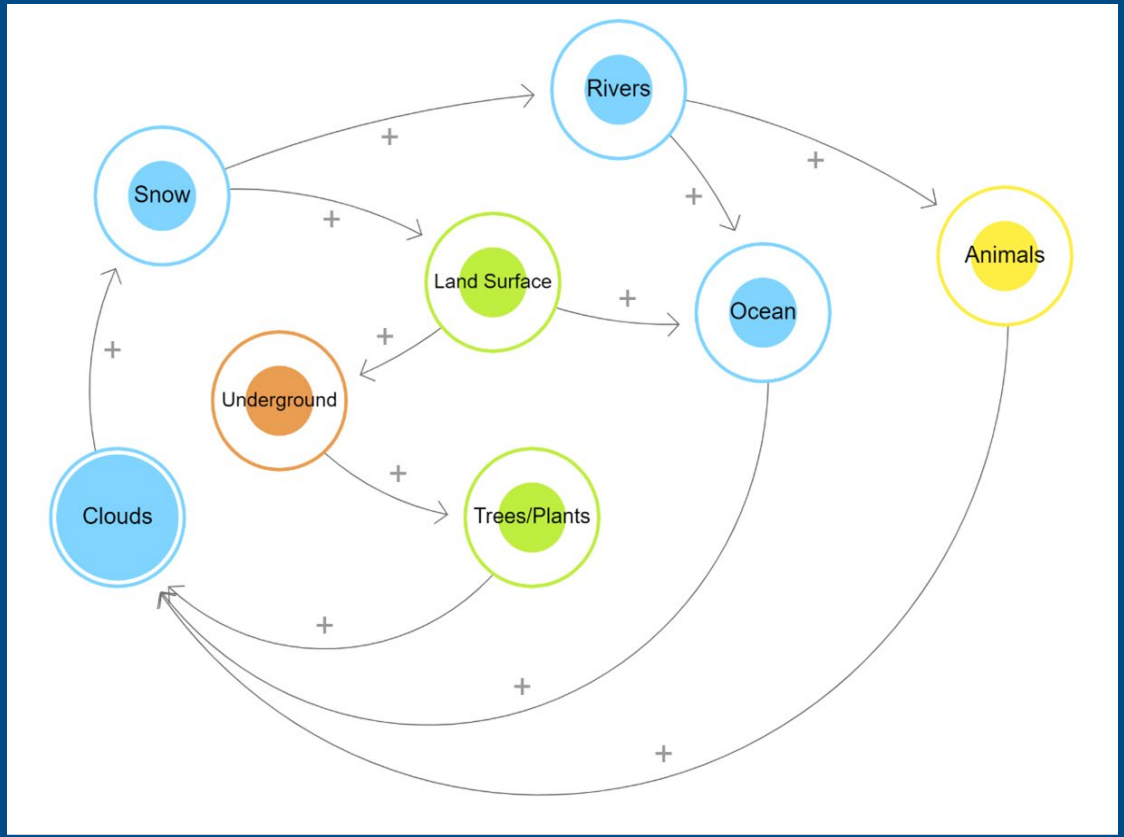
5. Positive Relationships are represented by a “+” sign

   a) E.g. An increase in element “X” leads to an increase in element “Y”

6. Negative Relationships are represented by a “-” sign

   a) E.g. An increase in element “X” leads to a decrease in element “Y”

Water Cycle Systems Map



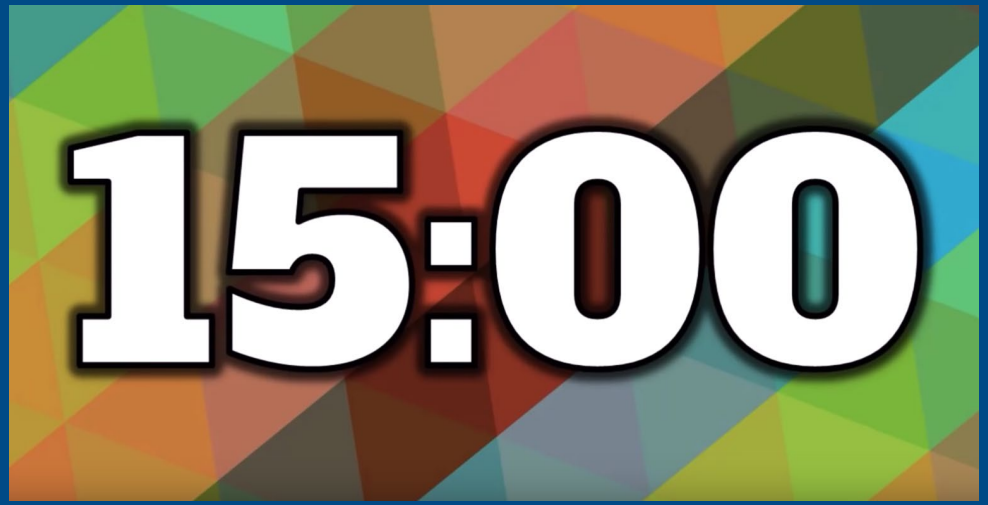
Basics of Systems Maps

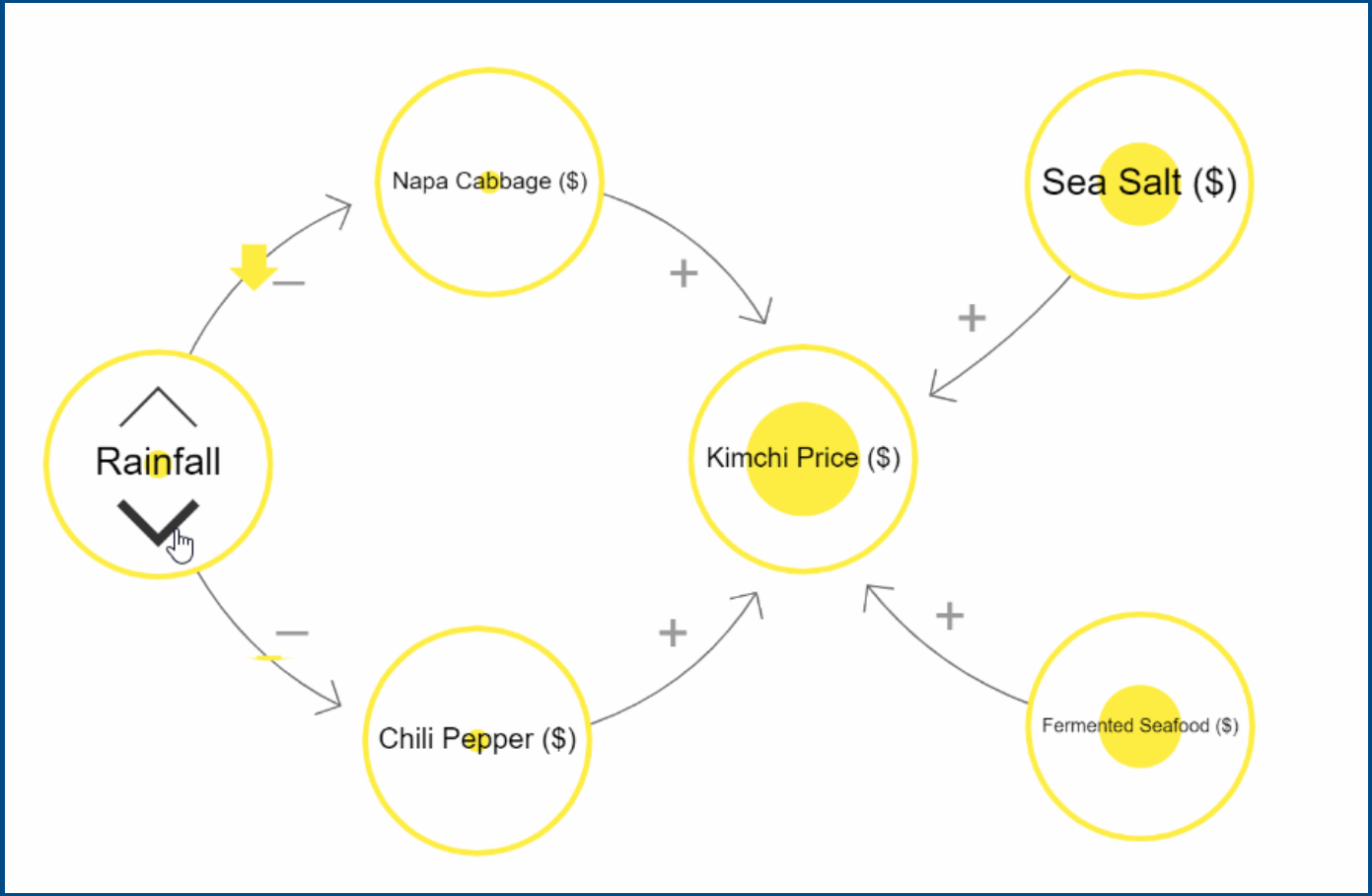
1. Amount of Rainfall
2. Price of Napa Cabbage
3. Price of Chili Pepper
4. Price of Fermented Seafood
5. Price of Sea Salt

Draw How These Elements Are Connected

Use Loopy to draw the systems map online
https://ncase.me/loopy/

Let’s Start




Napa Cabbage ($)
Sea Salt ($)
Rainfall
Kimchi Price ($)
Chili Pepper ($)
Fermented Seafood ($)
+
+
+
+
−
−

Project Vegita

16-year-old Kim Han Seo from Korea uses AI to predict cabbage prices from weather data

How does Project Vegita work?

1. Systems Map of Kimchi Price

a. It was determined that rainfall, temperature and cabbage prices were the main factors.

b. With our AI skills, our best leverage would be to predict cabbage price.

How does Project Vegita work?

2. Collect Data

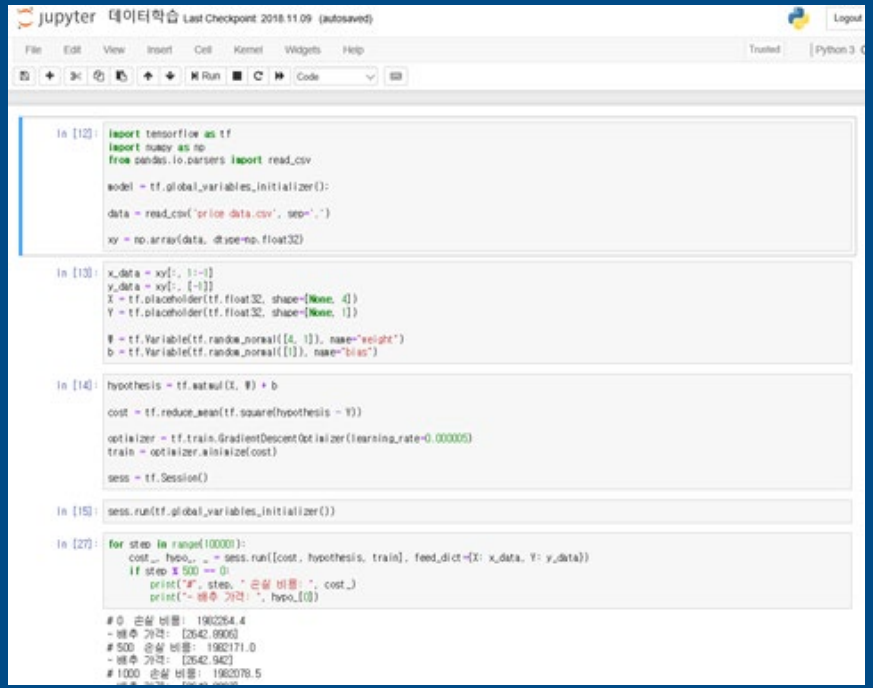
a. > 3,000 daily temperatures, rainfall and cabbage prices were collected from the Korea Meteorological Administration and the Ministry of Agriculture and Forestry sites from the years 2010 to 2017.

	A	B	C	D	E	F
1	year	avgTemp	minTemp	maxTemp	rainFall	avgPrice
2	20100101	-4.9	-11	0.9	0	2123
3	20100102	-3.1	-5.5	5.5	0.8	2123
4	20100103	-2.9	-6.9	1.4	0	2123
5	20100104	-1.8	-5.1	2.2	5.9	2020
6	20100105	-5.2	-8.7	-1.8	0.7	2060
7	20100106	-7.3	-11.4	-2.5	0.3	2060
8	20100107	-6.7	-11.2	-1.2	0	2140
9	20100108	-5.6	-11.4	1.4	0	2140
10	20100109	-3.1	-8.8	1.8	0.1	2140
11	20100110	-1.3	-5	2.8	0	2140
12	20100111	-1.5	-5.1	2.1	0	2140
13	20100112	-5	-8.2	-1.7	0.4	2140
14	20100113	-8.3	-11.4	-4.6	0.7	2140
15	20100114	-6.4	-12.7	0	0.1	2140
16	20100115	-2.6	-7.9	3.9	0.1	2140
17	20100116	-4.1	-9.9	2.9	0	2140

How does Project Vegita work?

3. AI Learning

a. A “Linear Regression Algorithm” was applied using Python and the model was trained 100,000 times.



How does Project Vegita work?

4. Using AI

a. The trained model is used to predict the price of cabbage based on the input of the particular day's temperature and rainfall

```
avg_temp = float(input('AvgTemp: '))
min_temp = float(input('Mintemp: '))
max_temp = float(input('MaxTemp: '))
rain_fall = float(input('Rainfall: '))
```

```
AvgTemp: 8
Mintemp: 4
MaxTemp: 14
Rainfall: 0
```

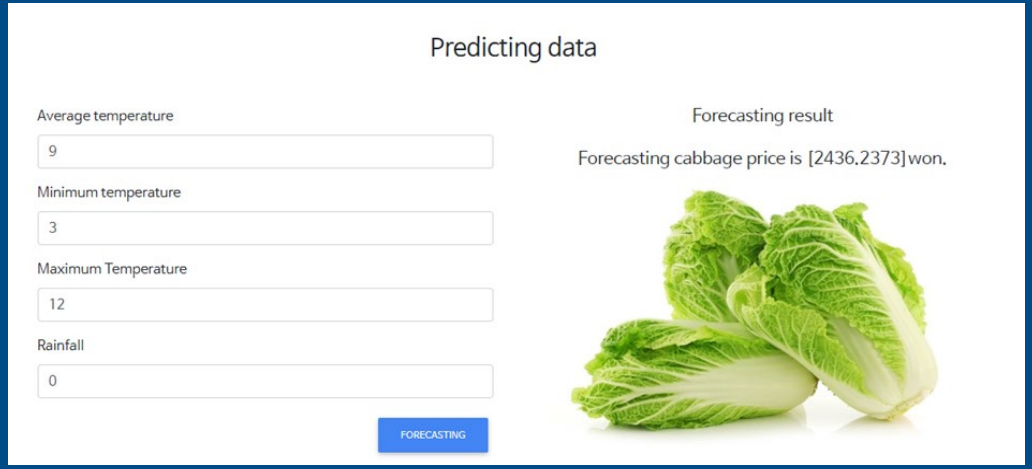
```
with tf.Session() as sess:
    sess.run(model)

    save_path = "./saved.cpkt"

    saver.restore(sess, save_path)
    data = ((avg_temp, min_temp, max_temp, rain_fall), )

    arr = np.array(data, dtype=np.float32)


    x_data = arr[0:4]
    dict = sess.run(hypothesis, feed_dict={X: x_data})

    print(dict[0])
```

```
INFO:tensorflow:Restoring parameters from ./saved.cpkt
[-11.351378]
```

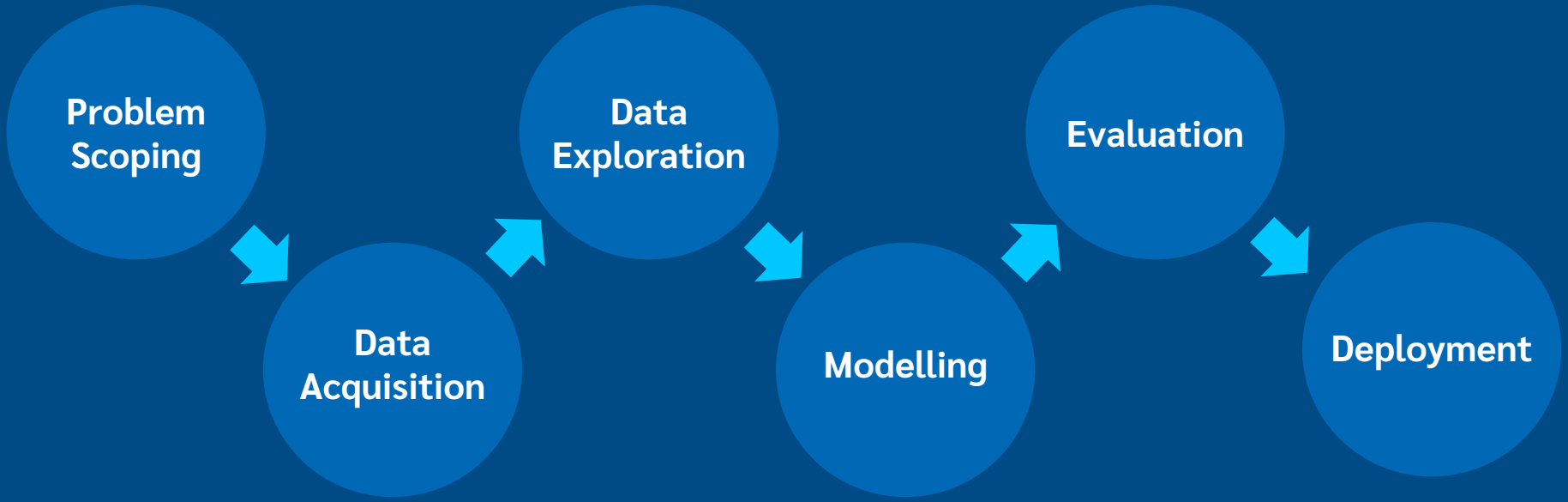
How does Project Vegita work?

5. Checking Data with Real Life

a. The forecasted data is checked with actual prices to improve the model over time.



AI Project Cycle



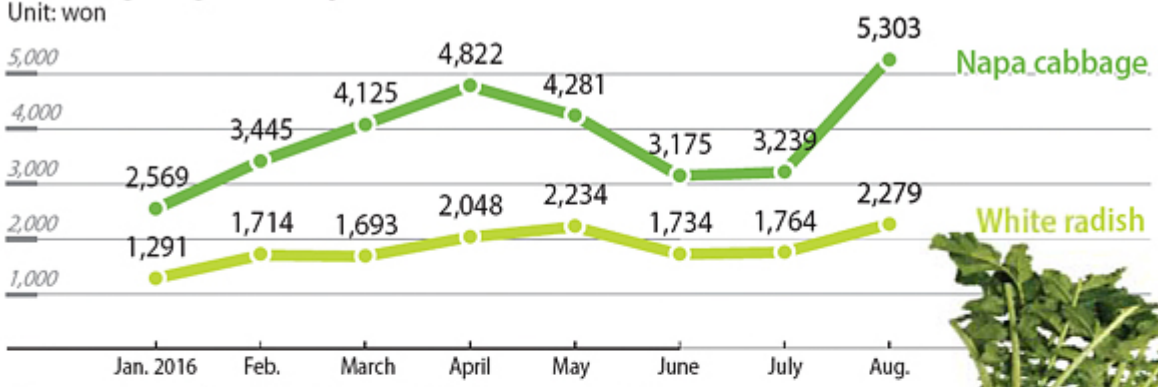
1. Problem Scoping

4Ws Problem Canvas

Rising Prices of Kimchi

Soaring vegetable prices

Unit: won



*Average price per head of cabbage and 1.5 kilograms of radishes

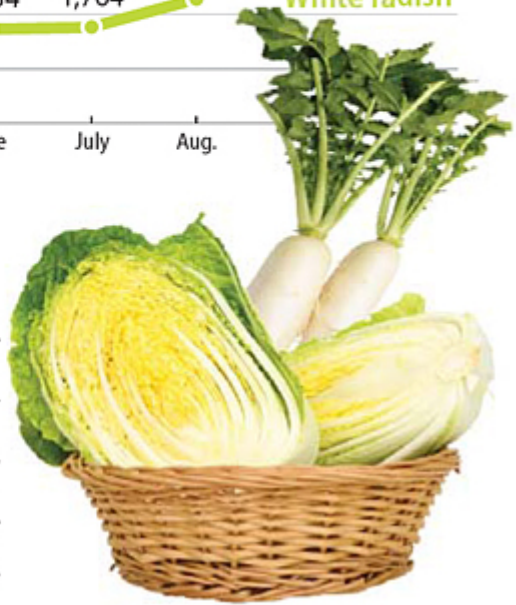
Price by types of stores

Unit: won

	Napa cabbage	White radish
Traditional markets	4,784	1,697
Department stores	5,530	2,251
Large discount chains	3,702	1,967

*Average price per head of cabbage and 1.5 kilograms of radishes

Source: Korea Consumer Agency



4Ws Problem Canvas

WHO ?	**WHAT ?**
WHERE ?	**WHY ?**

4Ws Problem Canvas – WHO?

Who is having the problem?

1. Who are the stakeholders?

(Think about the ones having the problem and benefitting from the solution)

2. What do you know about them?

4Ws Problem Canvas – WHO?

Who is having the problem?

1. **Who are the stakeholders?**
   - Home Makers: Mothers, Daughter-in-laws
   - Farmers
2. **What do you know about them?**
   - Home Makers have their own family kimchi recipes handed down generation to generation
   - Farmers plant Napa cabbage for a living

4Ws Problem Canvas – WHAT?

What is the nature of the problem?

1. **What is the problem?**
2. **How do you know it is a problem?**

   (Is there evidence?)

4Ws Problem Canvas – WHAT?

What is the nature of the problem?

1. What is the problem?

- Rise in cabbage prices due to poor harvest
- Unpredictability of cabbage prices

2. How do you know it is a problem?

- News articles on soaring cabbage prices near kimchi making season

4Ws Problem Canvas – WHERE?

Where does the problem arise?

1. What is the context/situation the stakeholders experience the problem?

(Did you observe the problem in context? Can you describe it?)

4Ws Problem Canvas – WHERE?

Where does the problem arise?

1. What is the context/situation the stakeholders experience the problem?

- During the kimchi making seasons
- During cabbage harvest seasons

4Ws Problem Canvas – WHY?

Why do you believe it is a problem worth solving?

1. **What would be of key value to the stakeholders?**
2. **How would it improve their situation?**

4Ws Problem Canvas – WHY?

Why do you believe it is a problem worth solving?

1. **What would be of key value to the stakeholders?**
   - To buy cabbage at its lowest prices
2. **How would it improve their situation?**
   - Allow them to make more kimchi and lower household expenses

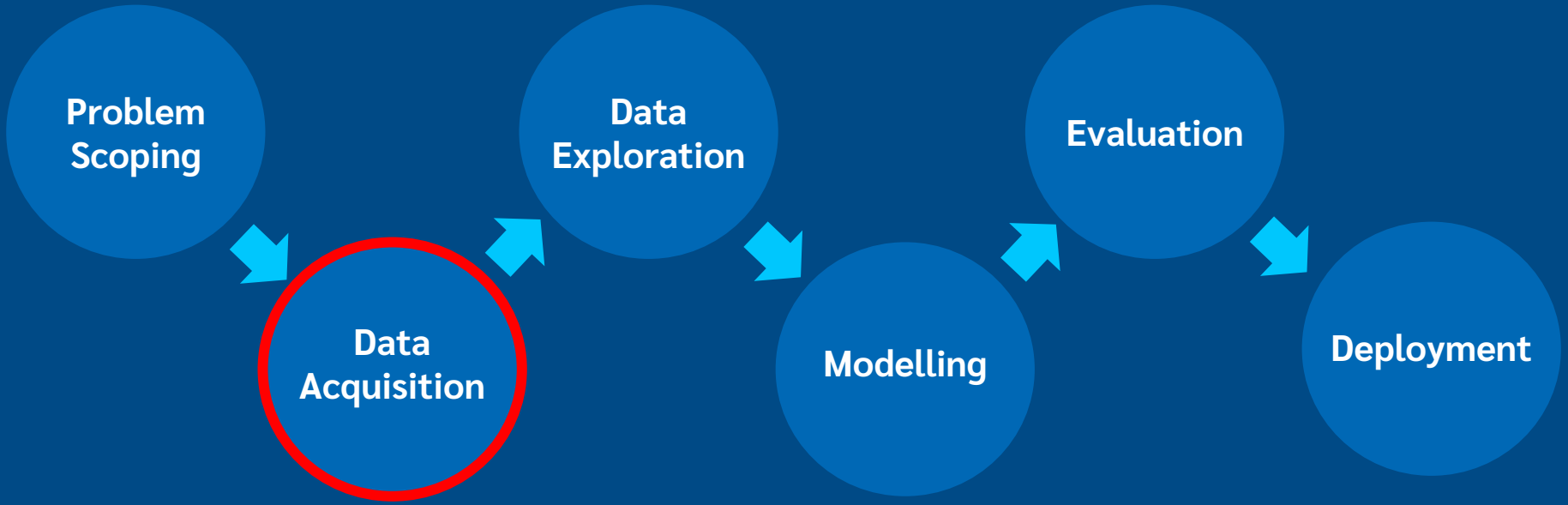
Problem Statement Template

Our	[stakeholders]	Who
has a problem that	[issue, problem, need]	What
when / while	[context, situation].	Where
An ideal solution would	[benefit of solution for them]	Why

Problem Statement Template

Our	home makers	Who
has/have a problem that	cabbage prices are unpredictable	What
when / while	they want to make kimchi	Where
An ideal solution would	help them predict the best time to buy cabbages	Why

AI Project Cycle



2. Data Acquisition

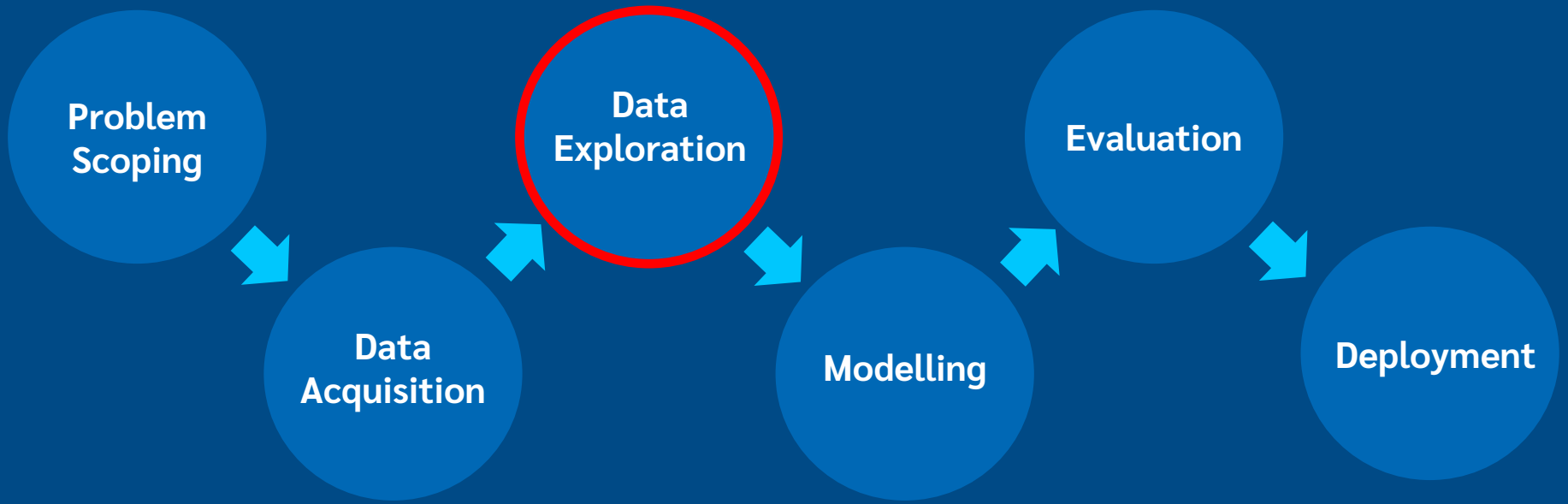
How do we know what data to get?

GOAL: How might we reduce illegal poaching of elephant tusks?

GOAL: How might we predict whether a songs makes it to the billboard?

GOAL: How might we improve the exam scores of students?

AI Project Cycle



3. Data Exploration

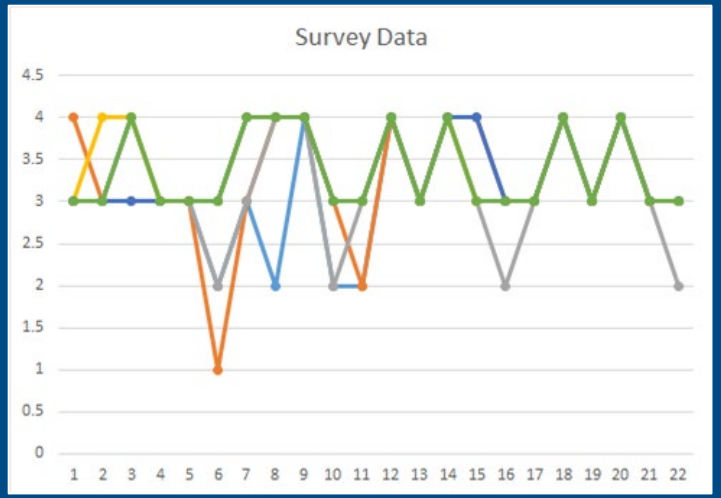
Why do we need to explore & visualize data?

Why Explore Data?

- We want to quickly get a sense of the trends, relationships and patterns contained within the data
- It helps us define strategy for which model to use at a later stage

Why Explore Data Through Visualization?

	A	B	C	D	E	F	G
1	1 = Strongly disagree; 4 = Strongly agree						
2	learning objectives were	materials and resources (e.g., videos,	the ideas, knowledge and skills learnt from the	The presentation was clear.	The questions were adequately addressed.	The facilitator was skillful at facilitating the participants' learning.	The duration of t programme was sufficient to mee the objectives of the programme.
3	3	4	3	3	3	3	Just Right
4	3	3	3	4	3	3	Just Right
5	4	3	4	4	3	4	Too Short
6	3	3	3	3	3	3	Just Right
7	3	3	3	3	3	3	Just Right
8	2	1	2	3	3	3	Too Short
9	3	3	3	4	4	4	Just Right
10	2	4	4	4	4	4	Just Right
11	4	4	4	4	4	4	Just Right
12	2	3	2	3	3	3	Just Right
13	2	2	3	3	3	3	Just Right
14	4	4	4	4	4	4	Just Right
15	3	3	3	3	3	3	Just Right
16	4	4	4	4	4	4	Just Right
17	3	4	3	3	4	3	Just Right
18	3	3	2	3	3	3	Just Right
19	3	3	3	3	3	3	Just Right
20	4	4	4	4	4	4	Just Right
21	3	3	3	3	3	3	Just Right
22	4	4	4	4	4	4	Just Right
23	3	3	3	3	3	3	Just Right



Visual representation is easier to understand and communicate to others

Examples of visualization

Visual representation helps us to understand the data

Examples of visualization

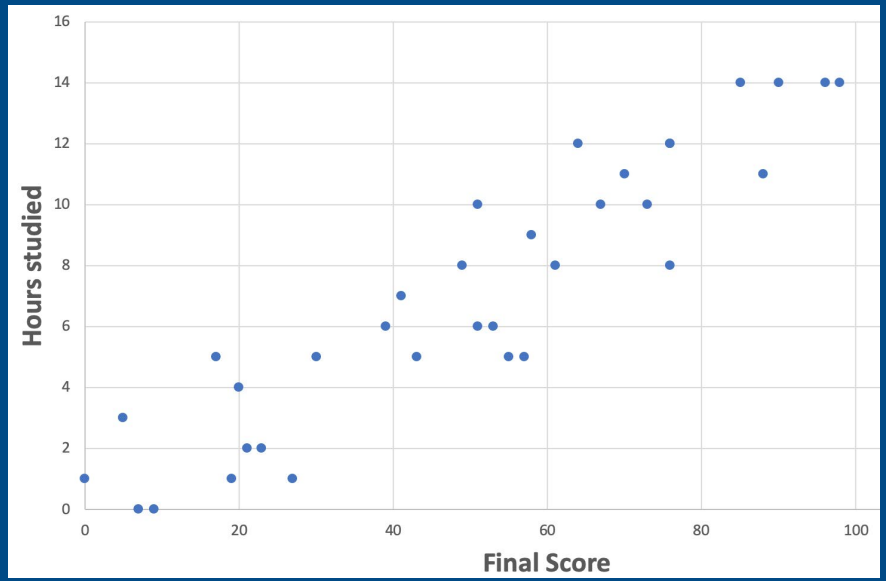
Visual representation helps us to understand the data

Name	Final Score	Hours Studied
RUBY	55	5
EMILY	19	1
GRACE	51	10
JESSICA	49	8
CHLOE	64	12
SOPHIE	67	10
LILY	5	3
AMELIA	70	11
EVIE	88	11
MIA	53	6
ELLA	58	9
CHARLOTTE	23	2
LUCY	85	14
MEGAN	61	8
ELLIE	73	10
ISABELLE	76	12
ISABELLA	17	5
HANNAH	9	0

Examples of visualization

Visual representation helps us to understand the data

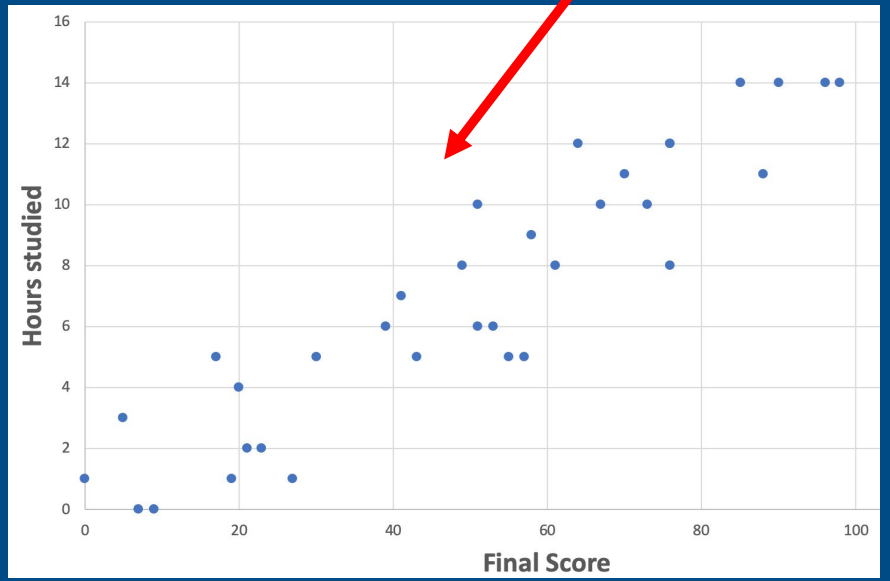
Name	Final Score	Hours Studied
RUBY	55	5
EMILY	19	1
GRACE	51	10
JESSICA	49	8
CHLOE	64	12
SOPHIE	67	10
LILY	5	3
AMELIA	70	11
EVIE	88	11
MIA	53	6
ELLA	58	9
CHARLOTTE	23	2
LUCY	85	14
MEGAN	61	8
ELLIE	73	10
ISABELLE	76	12
ISABELLA	17	5
HANNAH	9	0



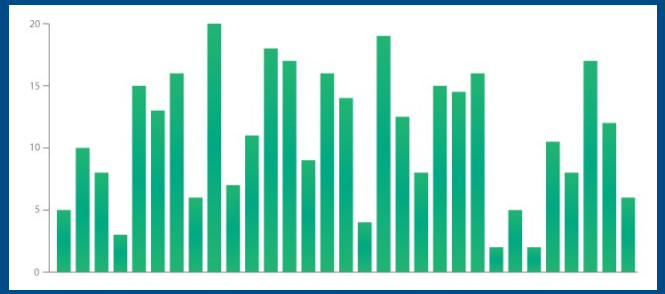
Examples of visualization

Visual representation helps us to understand the data

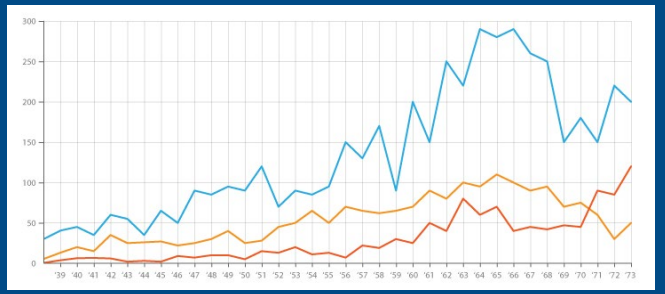
Name	Final Score	Hours Studied
RUBY	55	5
EMILY	19	1
GRACE	51	10
JESSICA	49	8
CHLOE	64	12
SOPHIE	67	10
LILY	5	3
AMELIA	70	11
EVIE	88	11
MIA	53	6
ELLA	58	9
CHARLOTTE	23	2
LUCY	85	14
MEGAN	61	8
ELLIE	73	10
ISABELLE	76	12
ISABELLA	17	5
HANNAH	9	0



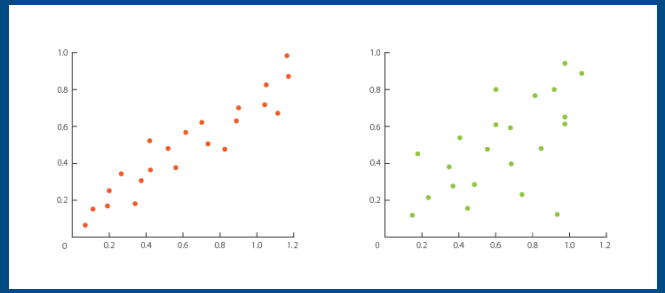
Common ways to visualize data



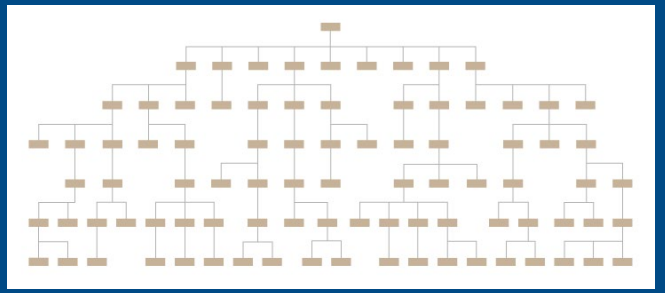
Bar Chart



Line Chart



Scatterplot

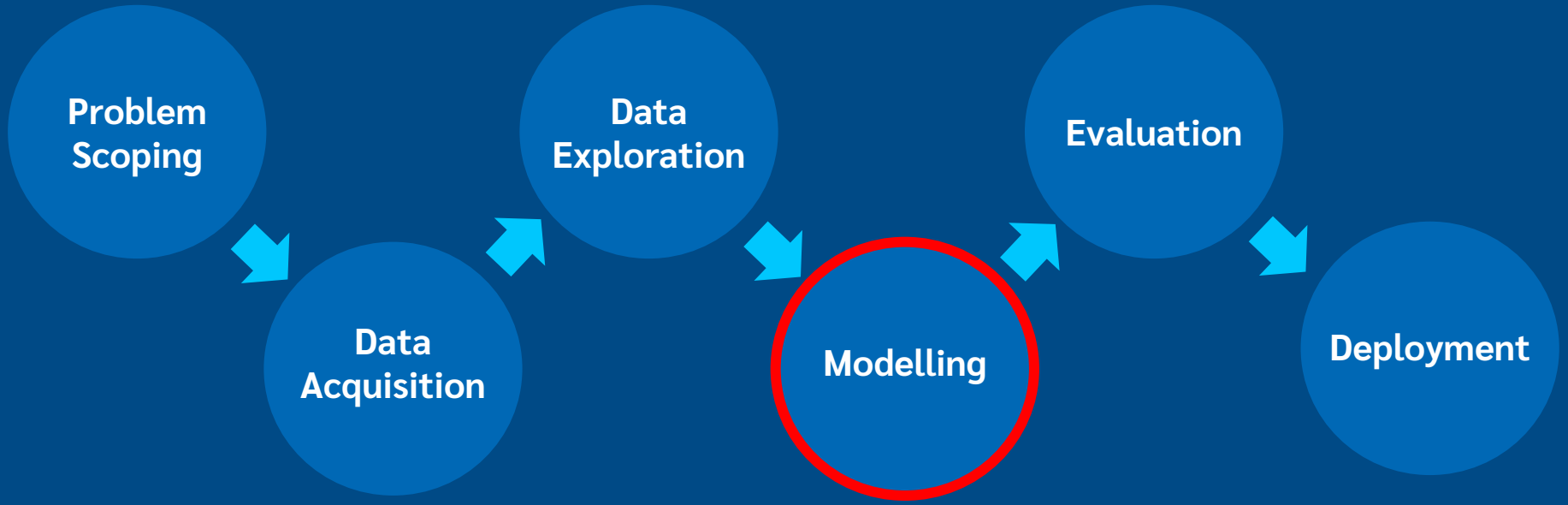


Tree Diagram

Source: datavizcatalogue.com

4. Modelling

AI Project Cycle



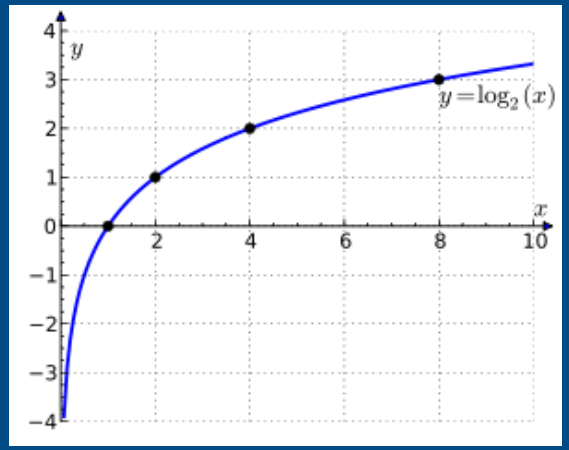
Types of Approaches When Building Models in AI

1. Rule Based Approaches

2. Learning Approaches

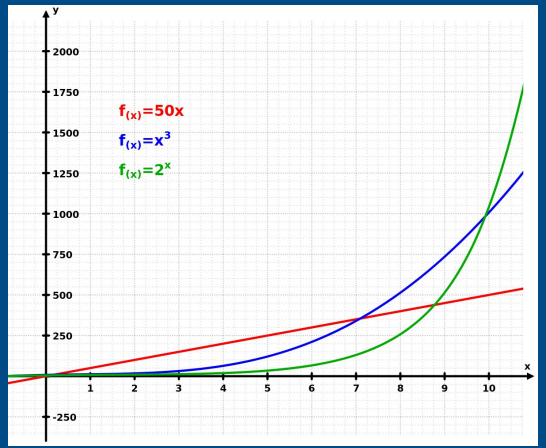
- Machine Learning
  - Decision Trees
  - Deep Learning

Examples of Mathematical Models



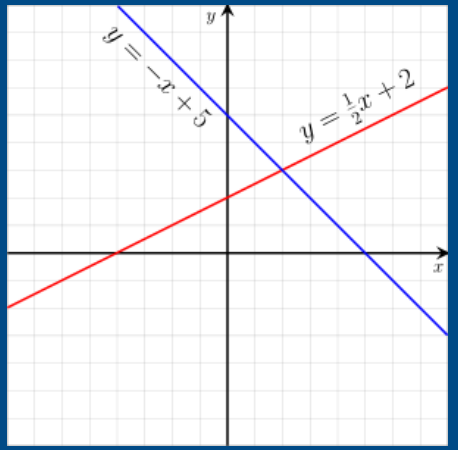
Source: Wikipedia

Logarithmic



Source: Wikipedia

Exponential

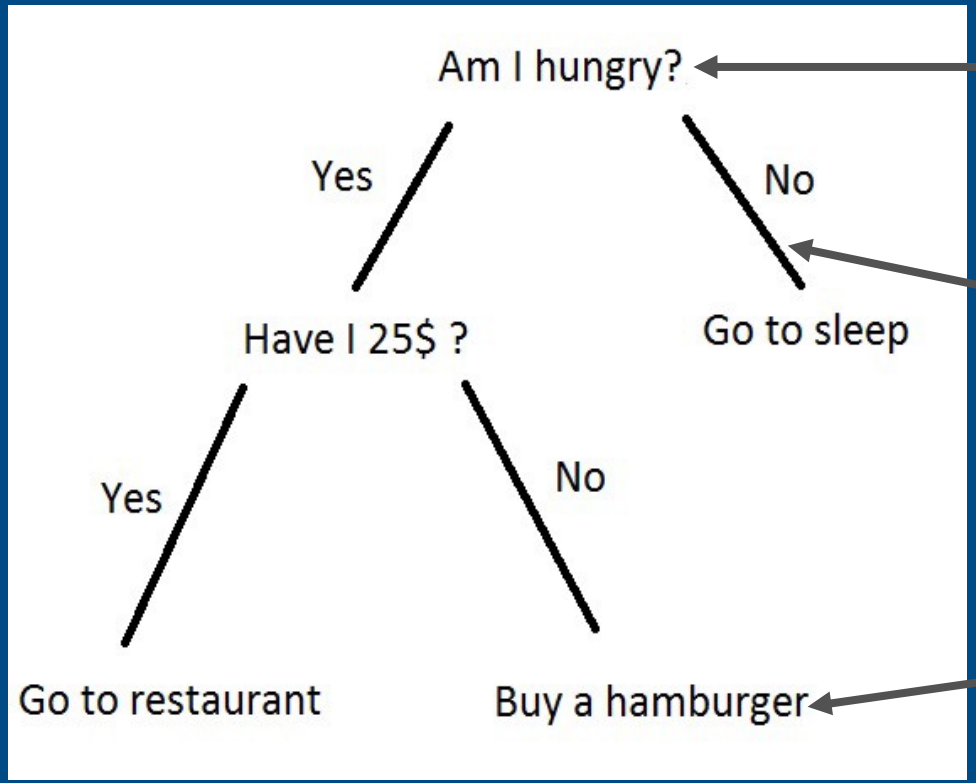


Source: Wikipedia

Linear

A **rule-based** approach means that we define the relationship
A **machine learning** approach means that we use machines to figure out the relationship for us, given the data.

Decision Trees



Root

Branch
(condition)

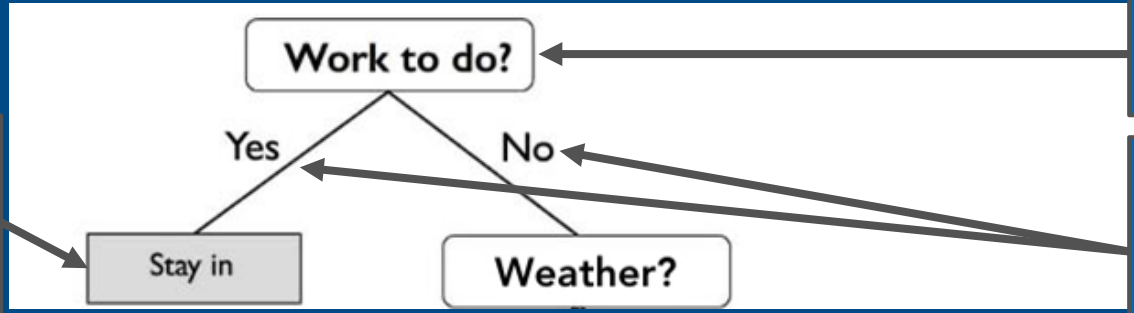
Leaf
(decision)

Source: skerritt.blog

Source: commons.wikimedia.org

Source: https://dev.to/nexttech/classification-and-regression-analysis-with-decision-trees-jgp

Leaf
(decision)

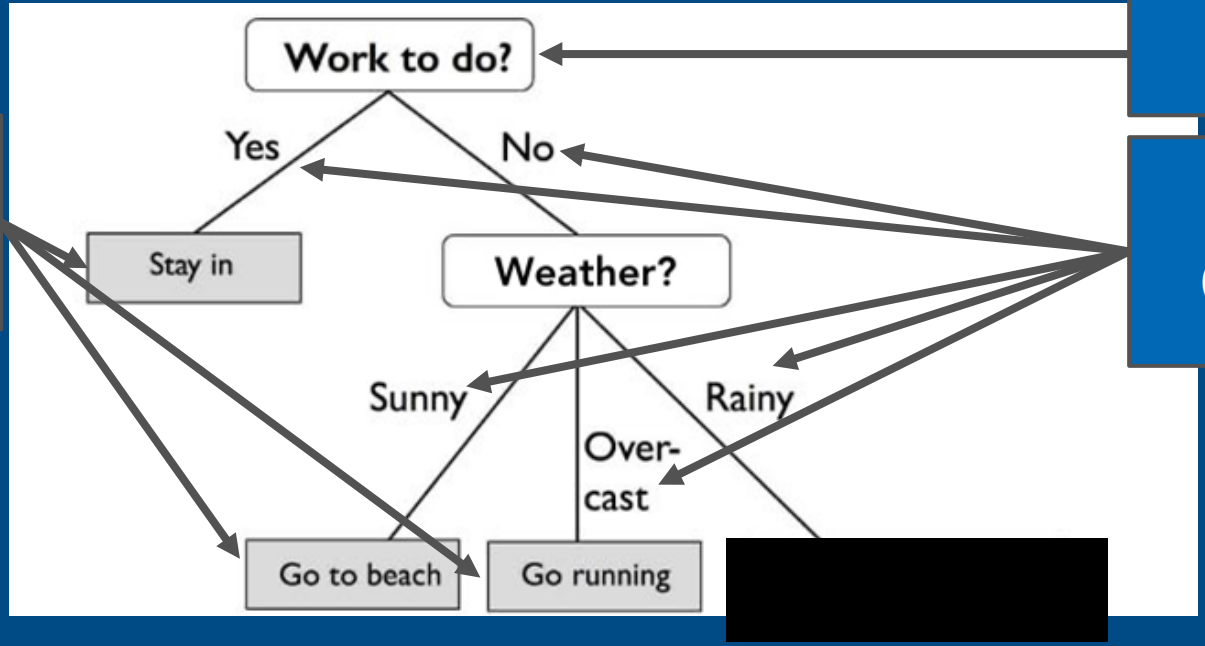


Root

Branch
(condition)

Source: https://dev.to/nexttech/classification-and-regression-analysis-with-decision-trees-jgp

Leaf
(decision)

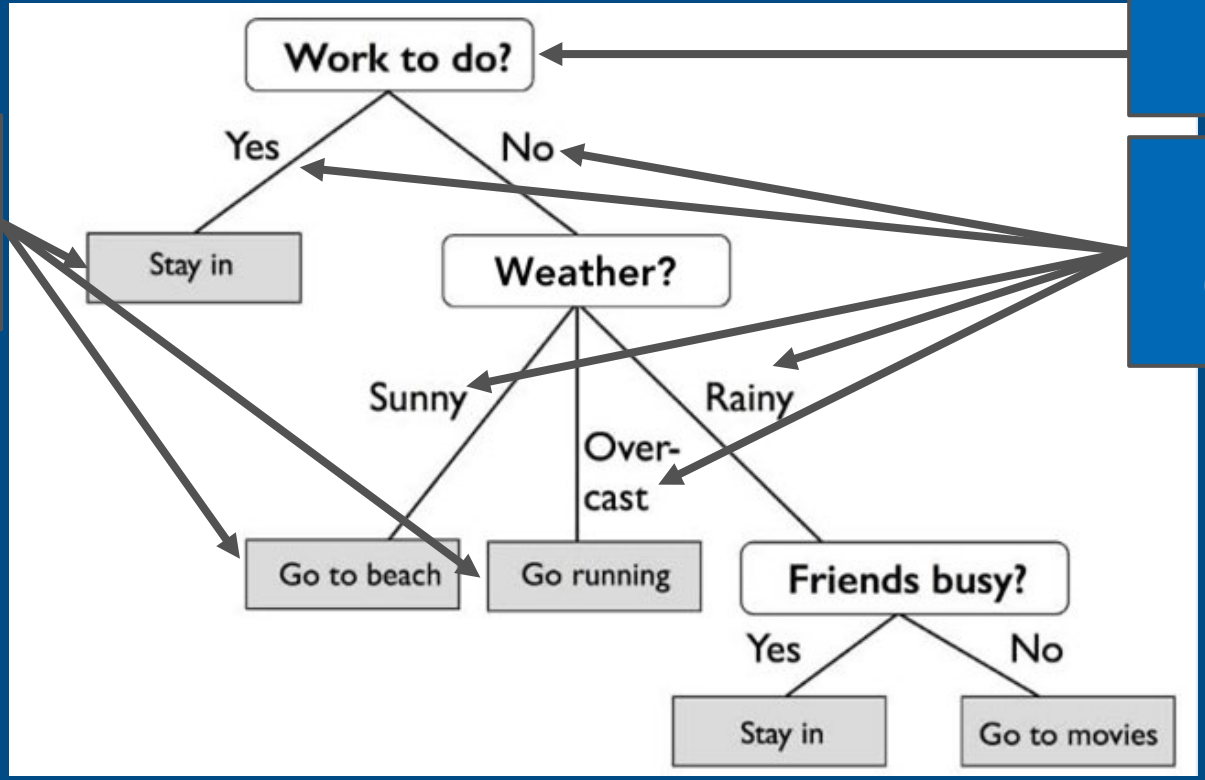


Root

Branch
(condition)

Source: https://dev.to/nexttech/classification-and-regression-analysis-with-decision-trees-jgp

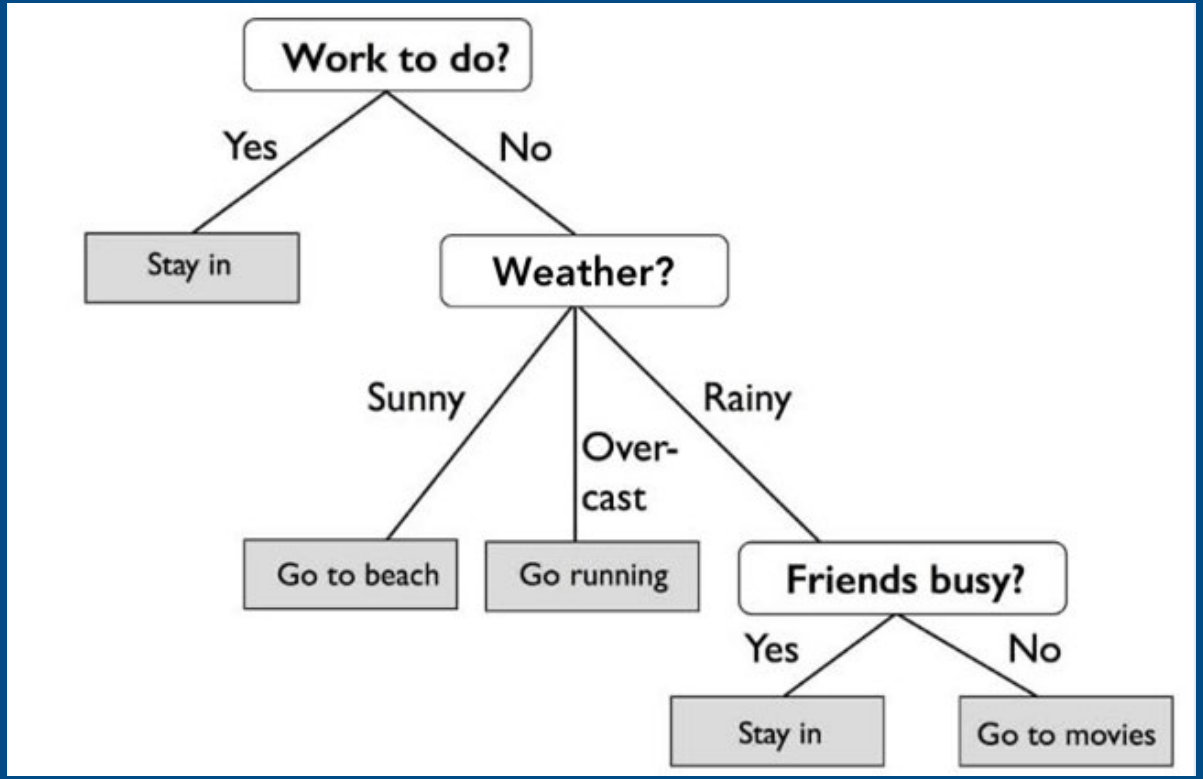
Leaf (decision)



Root

Branch (condition)

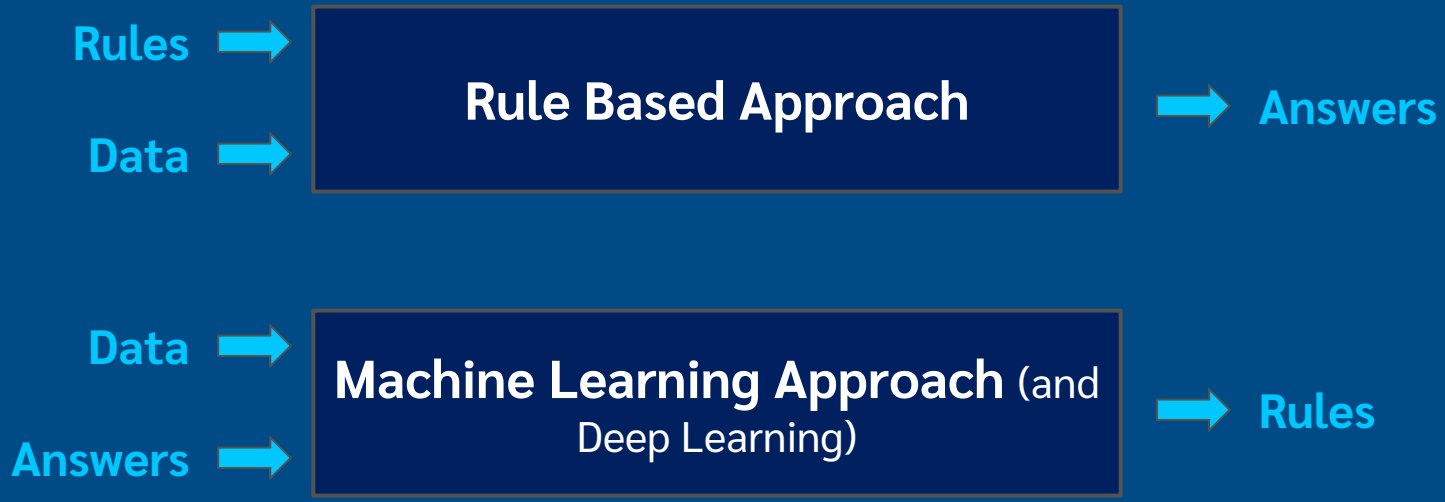
Source: https://dev.to/nexttech/classification-and-regression-analysis-with-decision-trees-jgp



Source: https://dev.to/nexttech/classification-and-regression-analysis-with-decision-trees-jgp

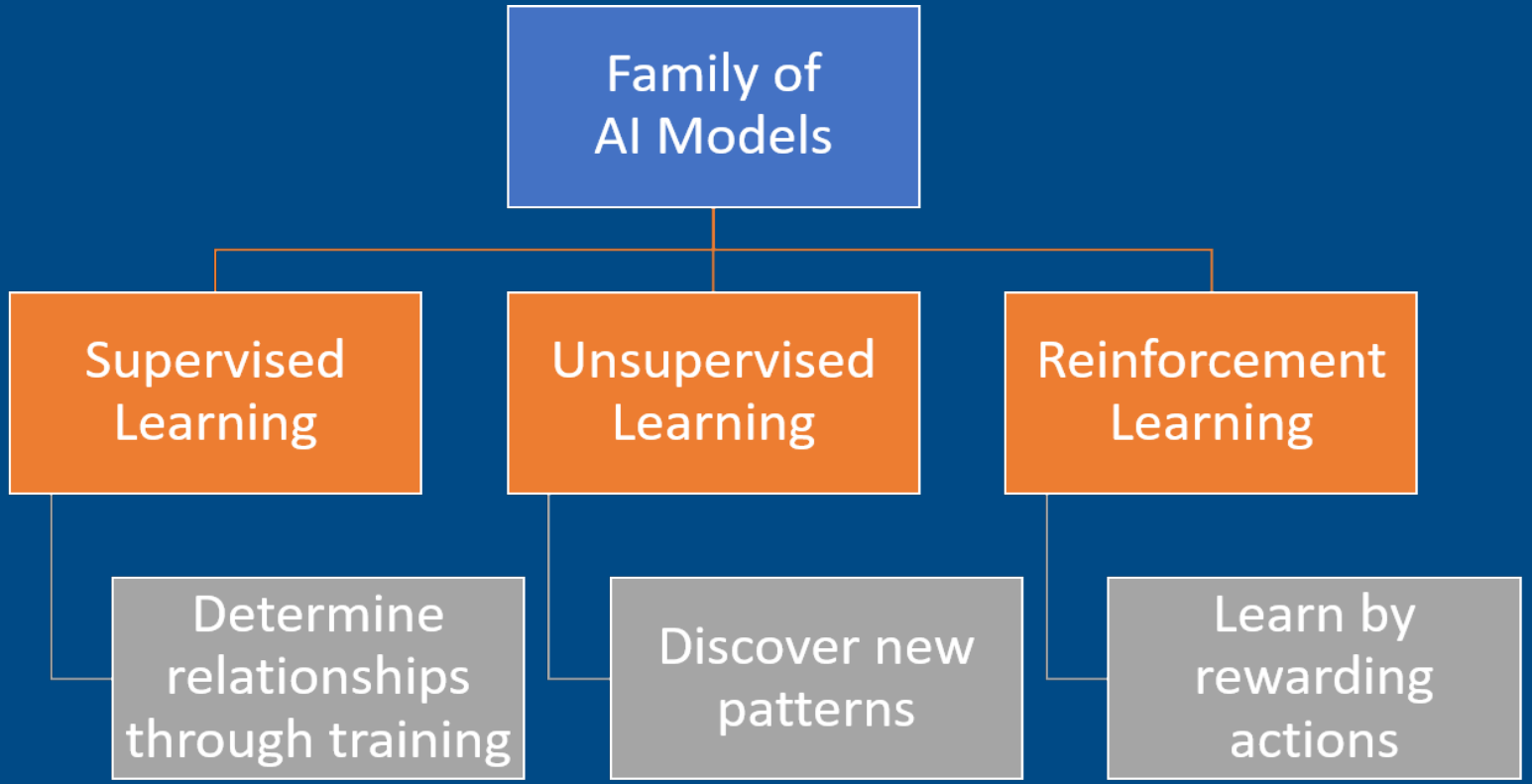
How is machine learning different from a rule-based approach?

Types of models in AI



What Machine Learning models are out there?

1. In general, there are 3 families of models:
    - **Supervised Learning**
    - **Unsupervised Learning**
    - **Reinforcement Learning**
2. We will only be focusing on **Supervised** and **Unsupervised Learning**


Family of AI Models
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Determine relationships through training
Discover new patterns
Learn by rewarding actions

Supervised Learning

1. Goal: To find specific relationships or structure in the input data that allow us to effectively produce correct output data

2. Data is labelled

3. Algorithms predict the output from the input data

- Classification (map input to output labels)
- Regression (map input to a continuous output)

Unsupervised Learning

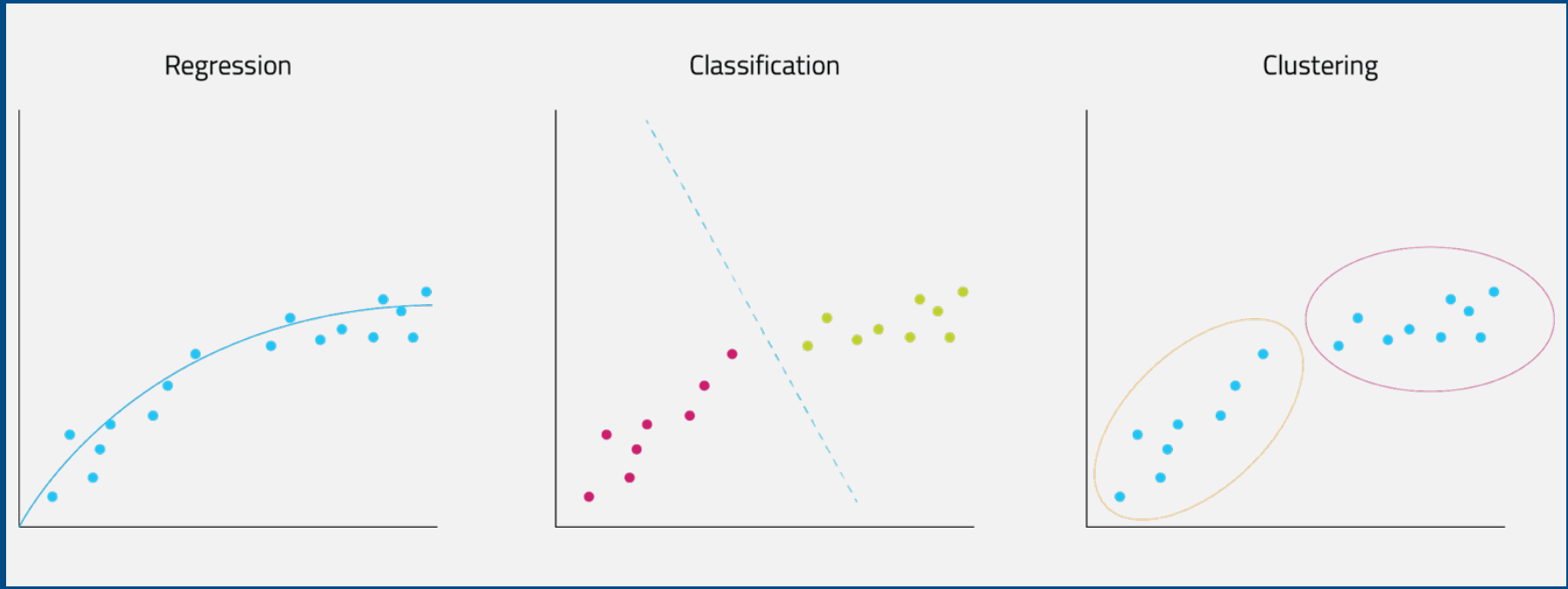
1. Goal: To learn the inherent structure of our data without using explicitly-provided labels

2. Data is unlabeled

3. Algorithms learn the inherent structure from the input data:

- Clustering (learn relationships between individual feature)
- Dimensionality Reduction (methods used to represent data using less columns or features)

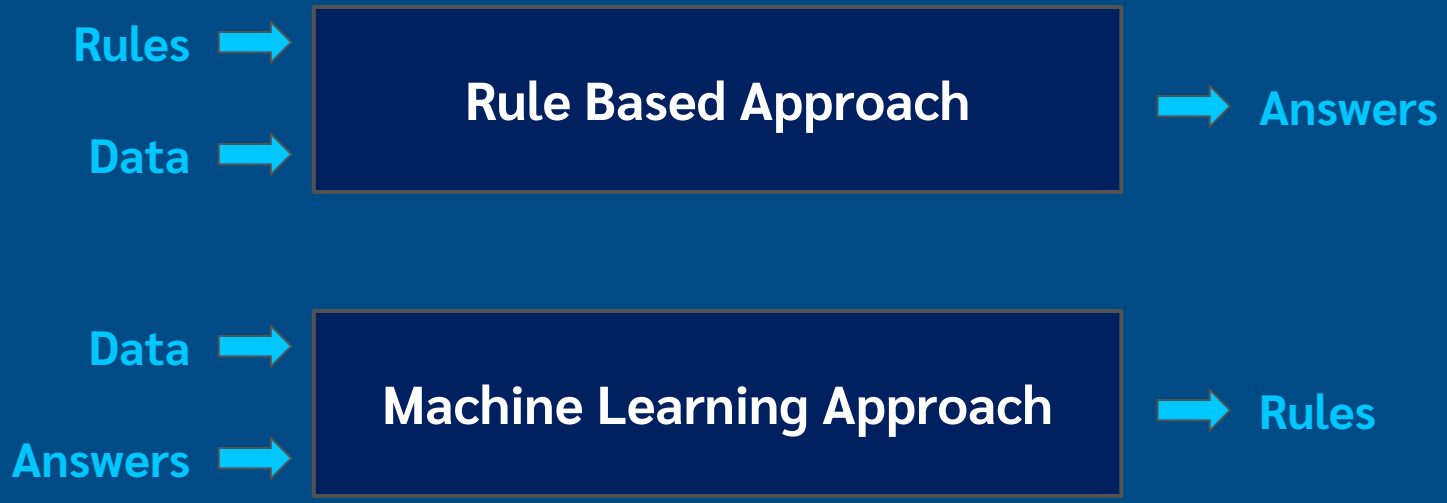
Let’s review the common ML algorithms



Source: https://www.moogsoft.com/blog/aiops/understanding-machine-learning-aiops-part-2/regression-class-clustering-graph

Lecture: Neural Networks

Let’s review the general approaches in AI models



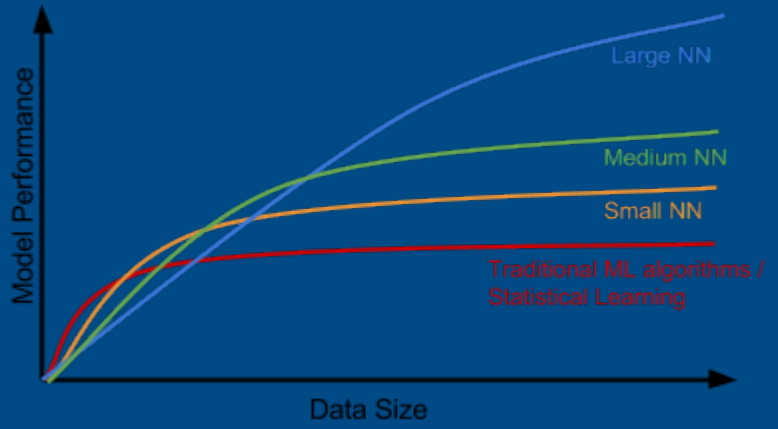
Neural Networks

1. Neural networks are loosely modelled after how neurons in the brain behave.

2. They can automatically extract features without input from the programmer.

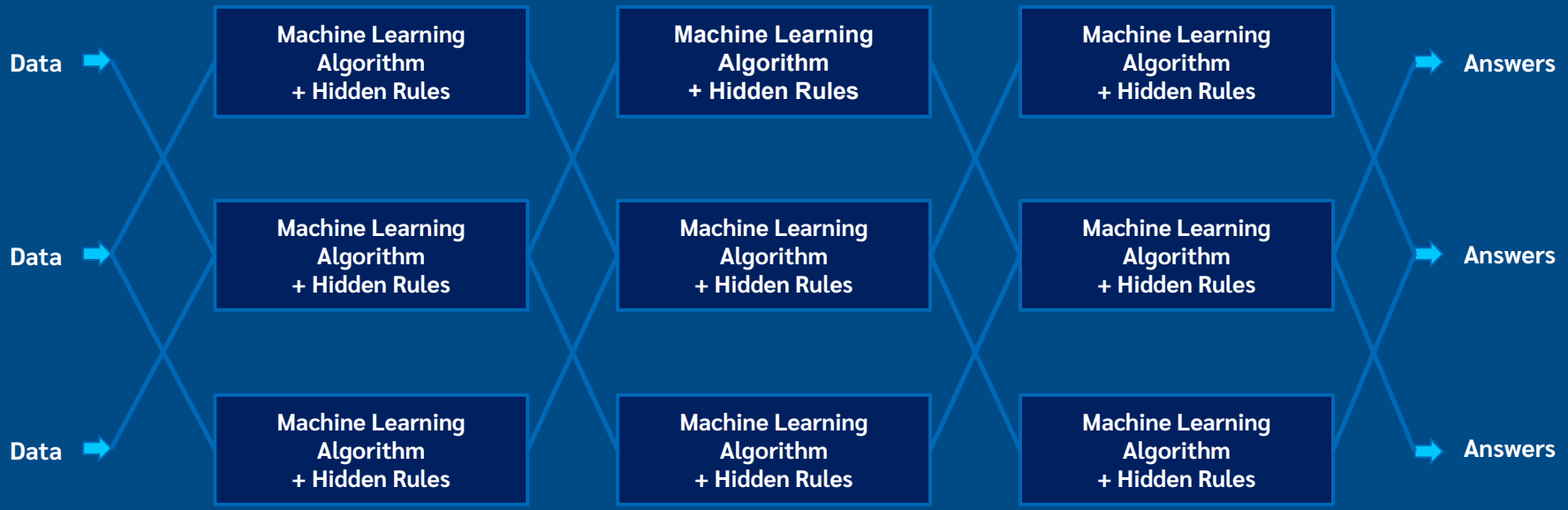
3. Every neural network node is essentially a machine learning algorithm.

4. It is useful when solving problems for which the data set is very large.



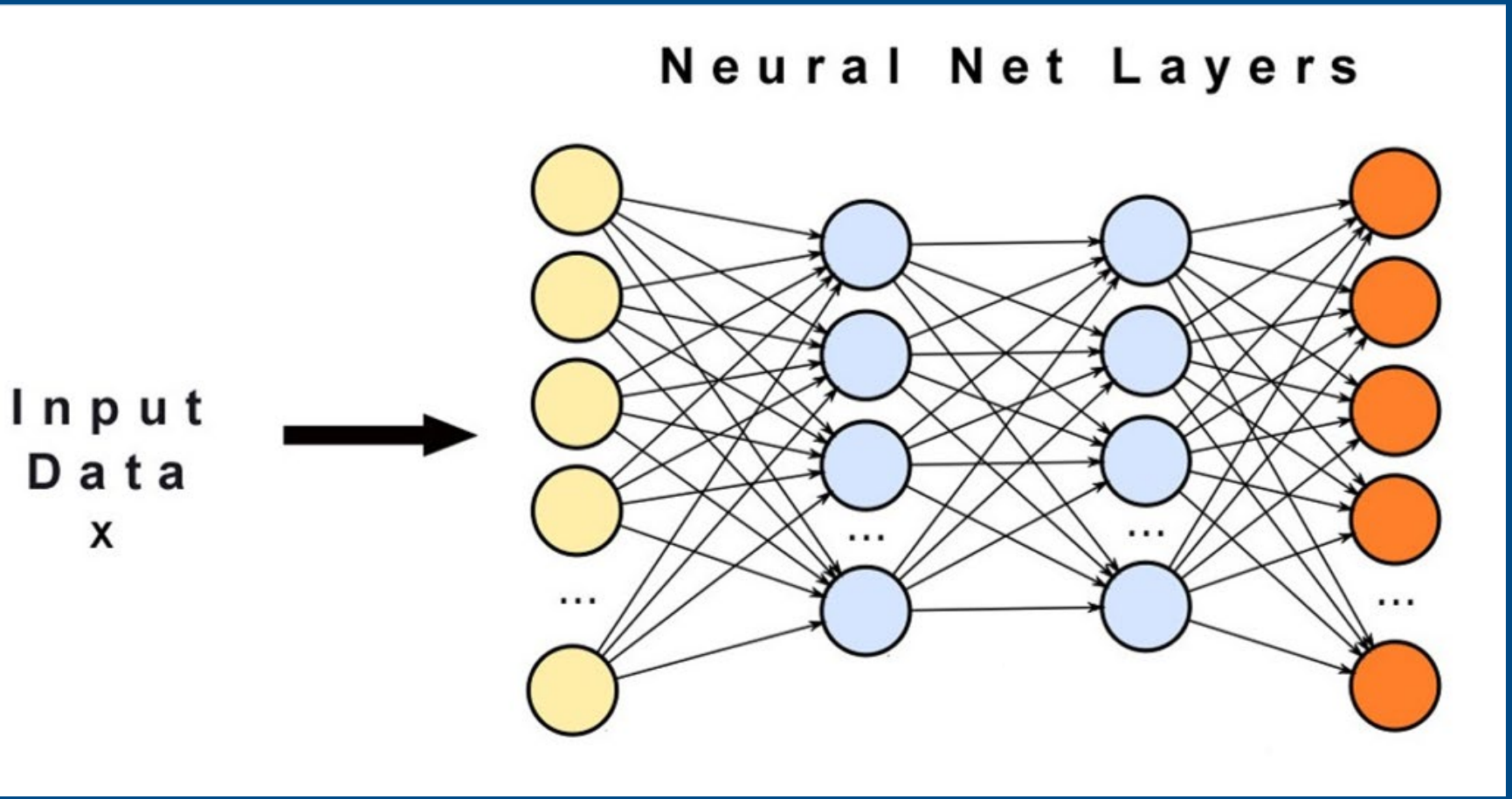
Source: https://towardsdatascience.com/machine-learning-a-gentle-introduction-17e96d8143fc

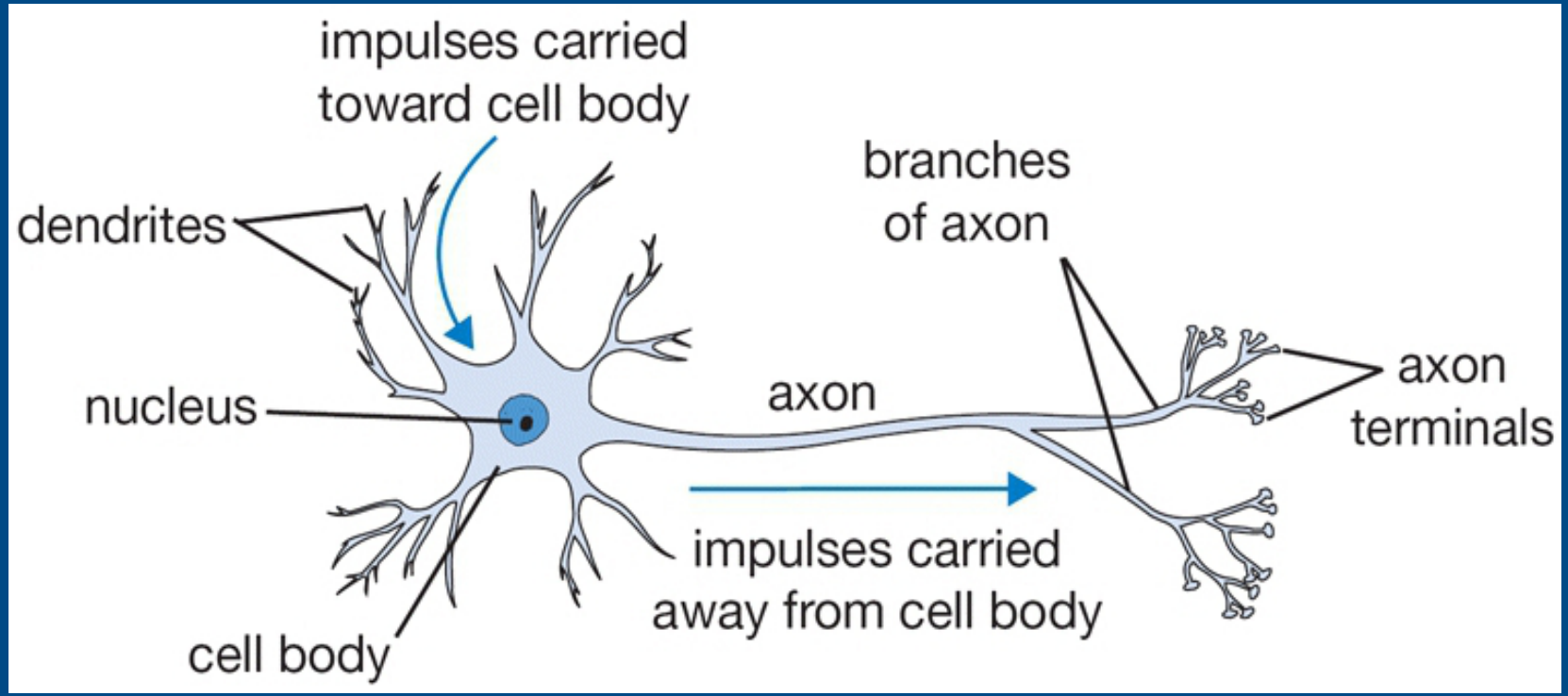
Neural Networks

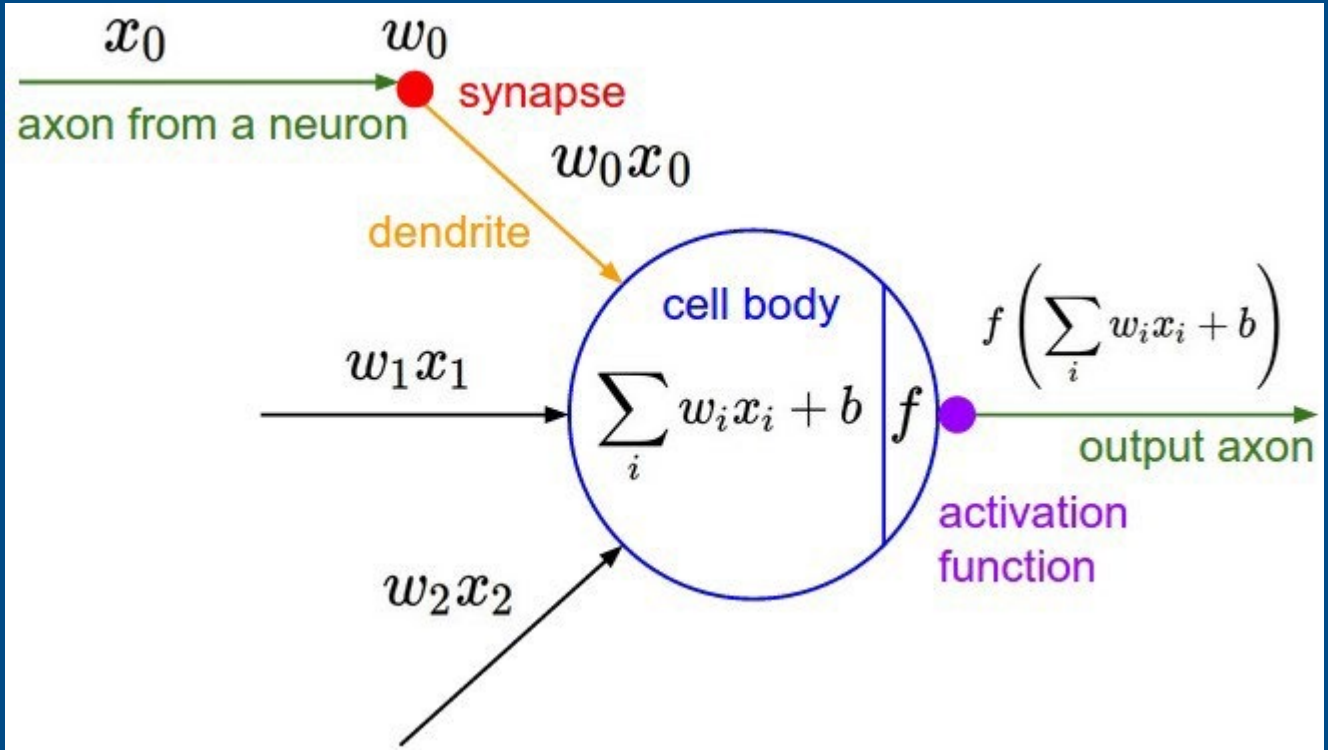


What the neural network code looks like

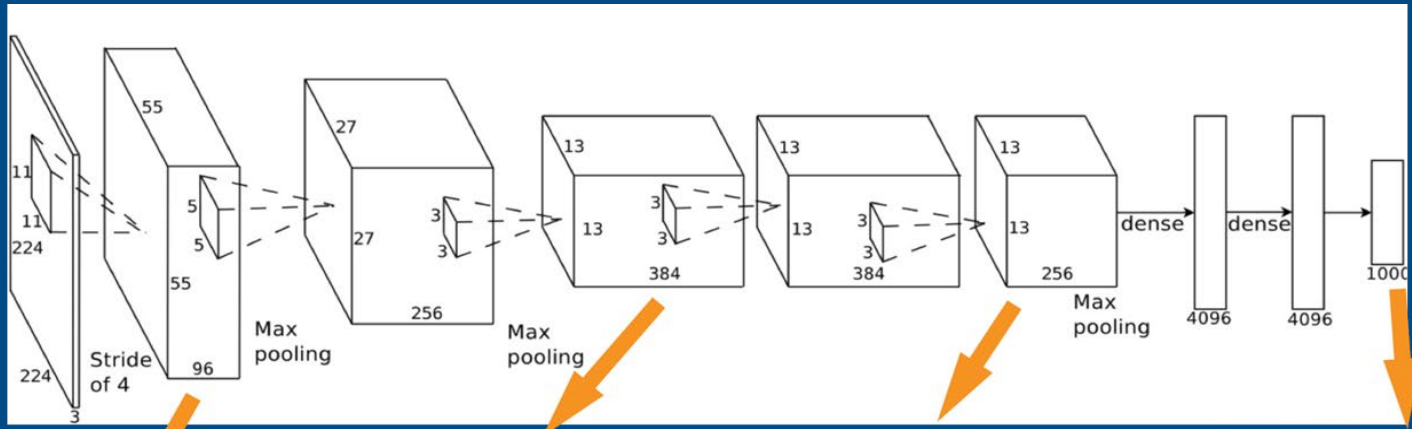
```
model = Sequential()

model.add(Convolution2D(nb_filters, kernel_size[0], kernel_size[1],
                        border_mode='valid',
                        input_shape=input_shape))
model.add(Activation('relu'))
model.add(Convolution2D(nb_filters, kernel_size[0], kernel_size[1]))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=pool_size))
model.add(Dropout(0.25))
```


Neural Net Layers
Input
Data
x

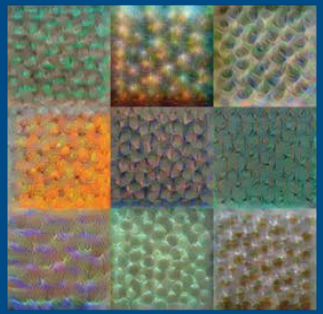

impulses carried
toward cell body
dendrites
branches
of axon
nucleus
axon
axon
terminals
impulses carried
away from cell body
cell body


x_0
w_0
synapse
axon from a neuron
$w_0 x_0$
dendrite
cell body
$w_1 x_1$
$\sum_i w_i x_i + b$
f
$f\left(\sum_i w_i x_i + b\right)$
output axon
activation function
$w_2 x_2$

```
model = Sequential()

model.add(Convolution2D(nb_filters, kernel_size[0], kernel_size[1],
                        border_mode='valid',
                        input_shape=input_shape))
model.add(Activation('relu'))
model.add(Convolution2D(nb_filters, kernel_size[0], kernel_size[1]))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=pool_size))
model.add(Dropout(0.25))
```



Conv 1: Edge + Blob



Conv 3: Texture

Numerical

Data-Driven

Conv 5: Object Parts

Cock

Ship

Dining tables

Grocery store

Fe8: Object Classes

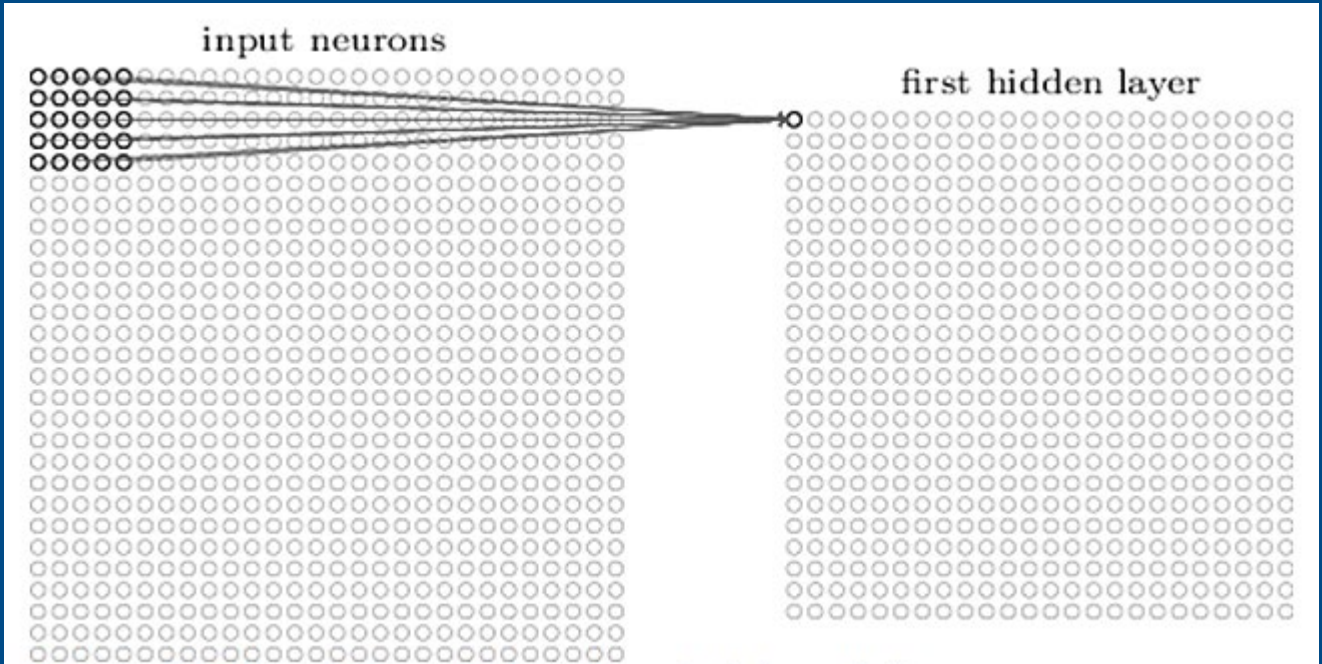
What exactly happens ?

What We See

```
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
52 70 95 23 04 60 11 42 69 24 68 56 01 32 56 71 37 02 36 91
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 67 48
```

What Computers See

Some Math Happens In The Layer

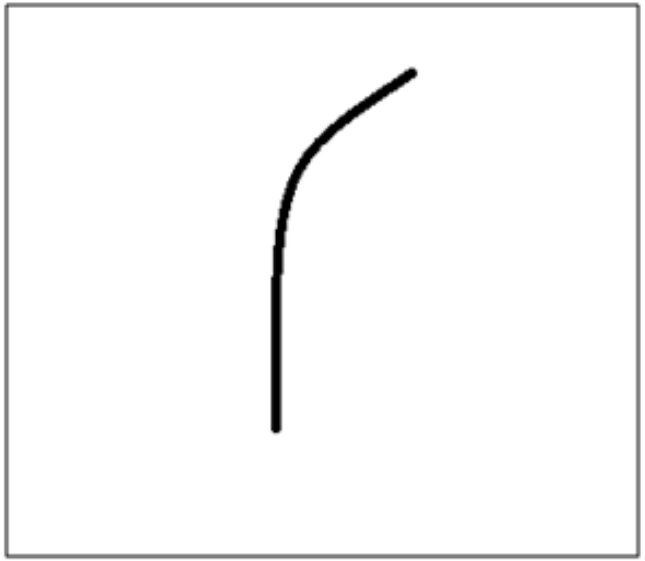


Visualization of 5 x 5 filter convolving around an input volume and producing an activation map

Feature Filter

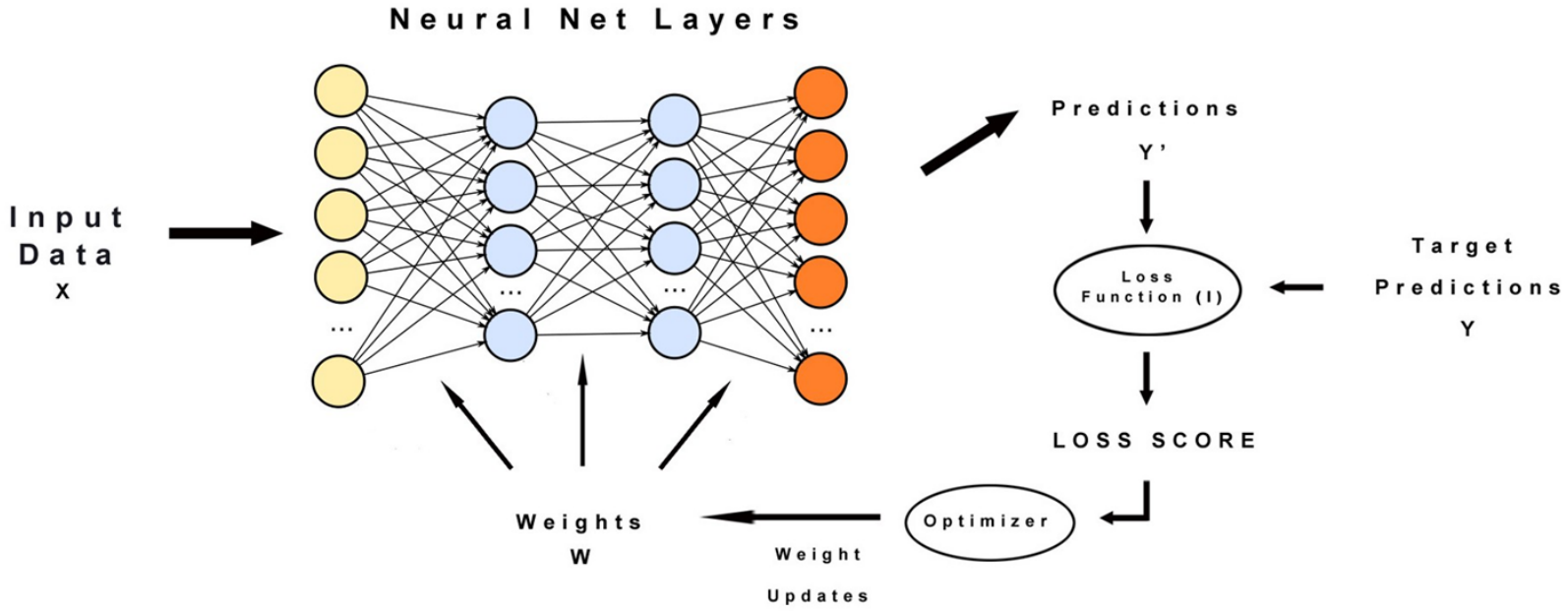
0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter



Visualization of a curve detector filter

LOGICAL FLOW OF A NEURAL NETWORK



Common Applications of Neural Networks

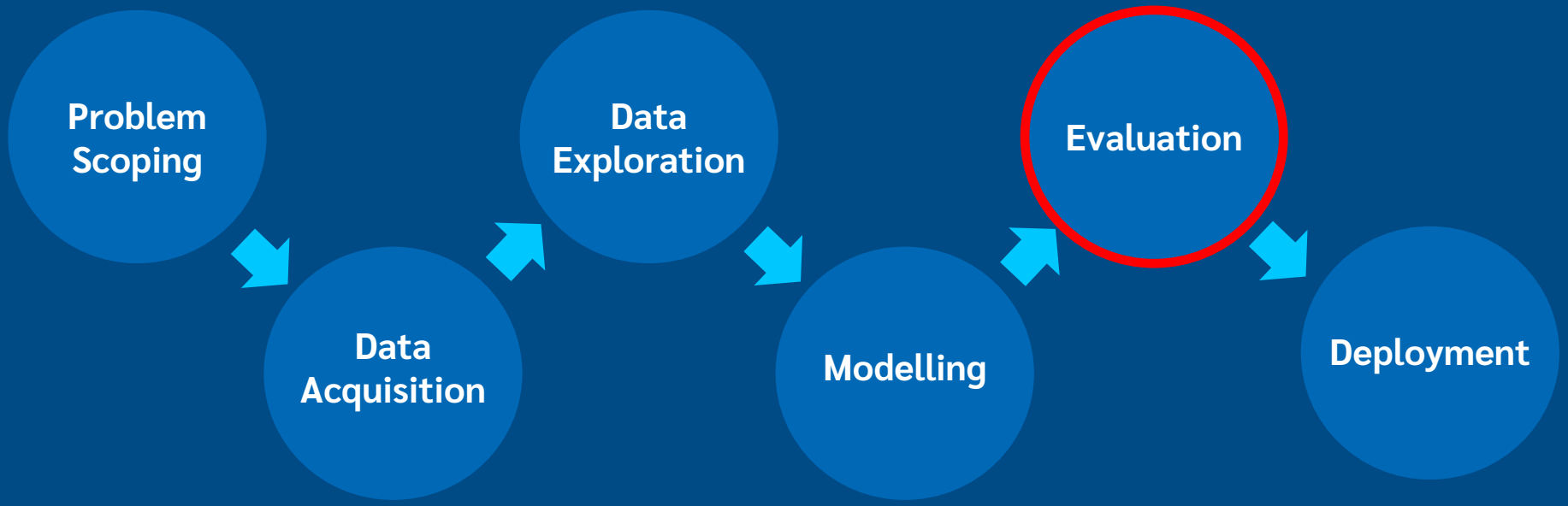
- **Speech Recognition**

  (e.g. Telling our smartphones to set an alarm)
- **Character Recognition**

  (e.g. Verification of the signature of a person)
- **Face Recognition**

  (e.g. faces being detected on our smartphones)

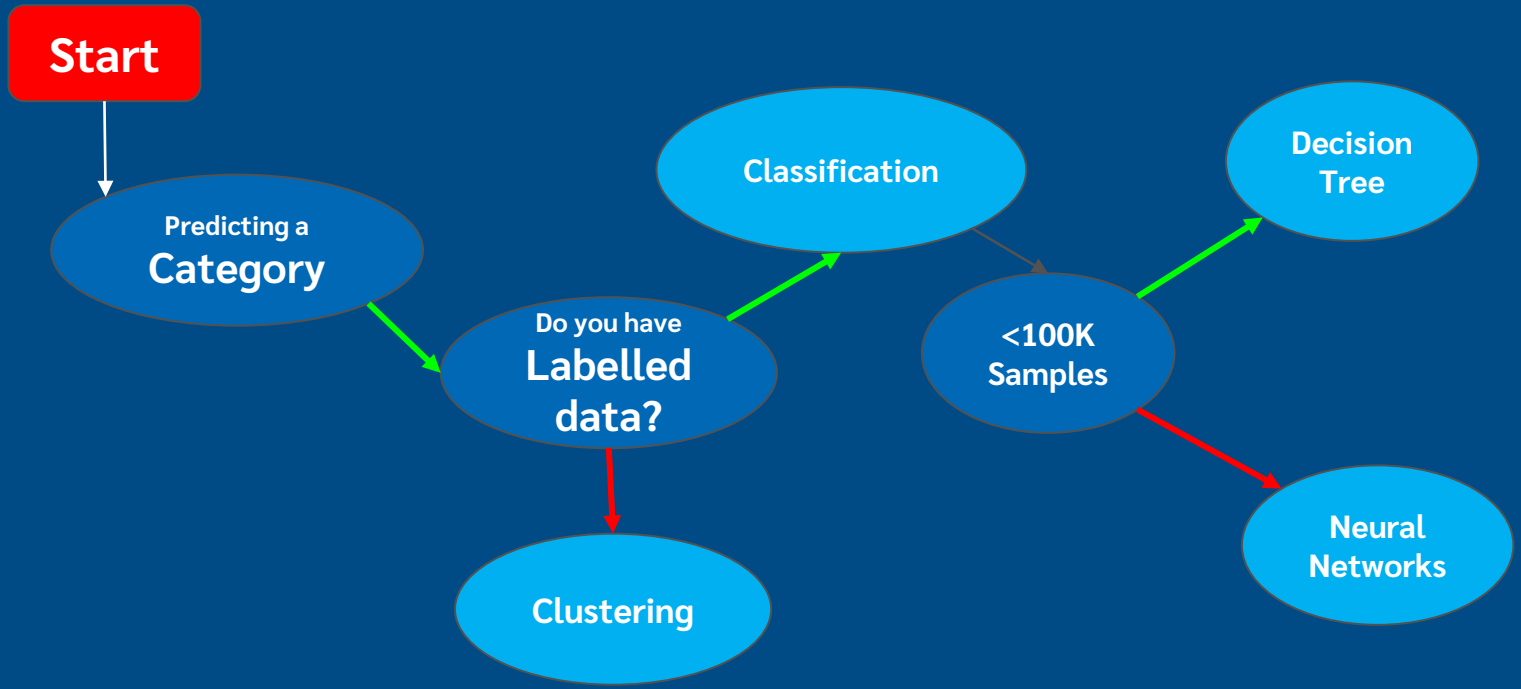
AI Project Cycle

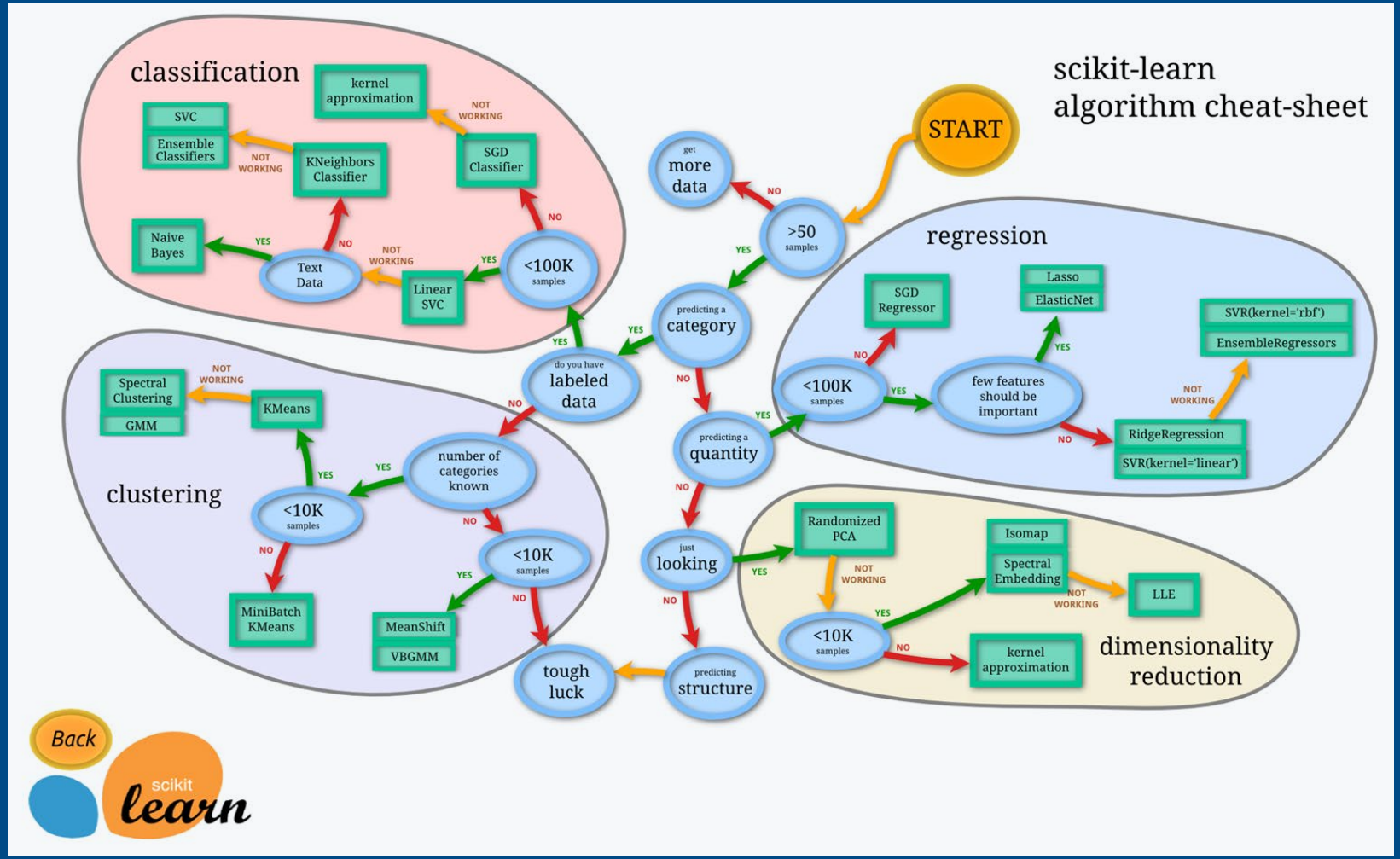


5. Model Evaluation

How do we select appropriate models?

Choosing Suitable Model




scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
NOT WORKING
YES
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
few features should be important
RidgeRegression
SVR(kernel='linear')
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K samples
kernel approximation
Back
scikit learn


Is there forest fire?
Prediction:
Yes / No
Reality:
Yes / No


Is there forest fire?
Prediction: Yes
Reality: Yes
True Positive


Is there forest fire?
Prediction: No
Reality: No
True Negative


Is there forest fire?
Prediction: No
Reality: Yes
False Negative


Is there forest fire?
Prediction: Yes
Reality: No
False Positive

Confusion Matrix

		Reality	
		Yes	**No**
Predicted	**Yes**	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Accuracy

- Accuracy is the percentage of correct prediction out of all the observations
- Accuracy = (TP + TN) / (TP + FP + TN + FN)
- Accuracy = (Correct Prediction) / (All Cases)
- Is high accuracy equivalent to good performance?


Is there forest fire?
Prediction: Always No
Reality: 2 % Probability
98% accurate
But it is usable?

Precision

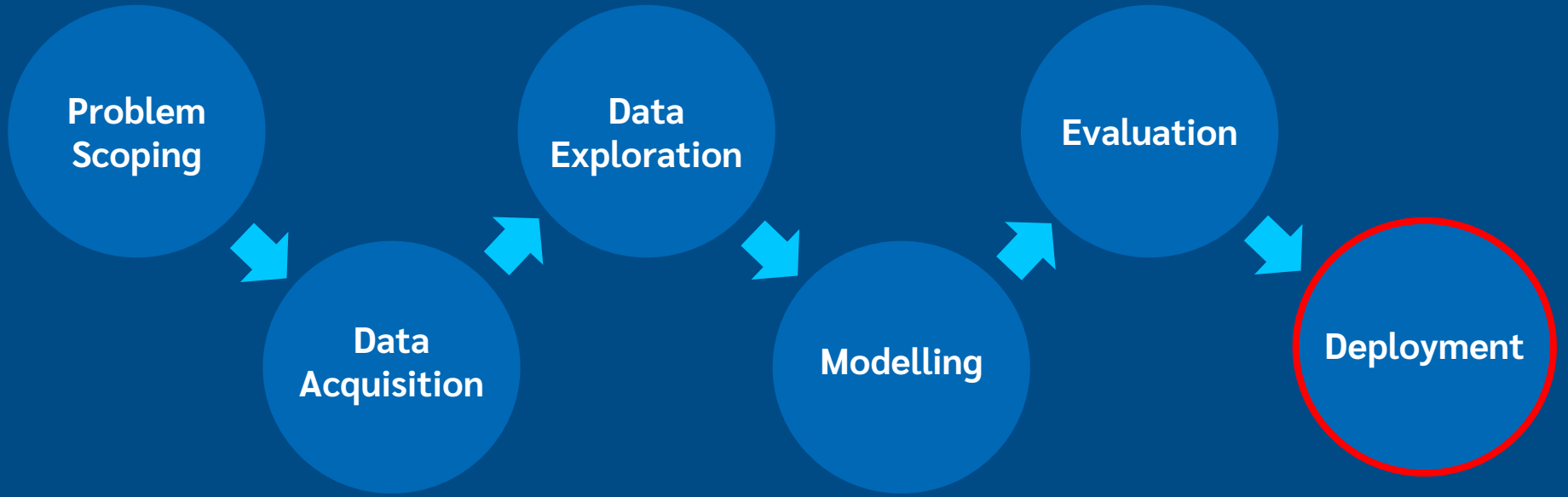
- Precision is the percentage of cases that are **predicted as positive (TP+FP)** that are **actually positive (TP)**
- Precision = (TP) / (TP + FP)
- Precision = (True Positive) / (Total Predicted Positive)

Recall

- Recall measures the fraction of **positive cases (TP+FN)** that are **correctly identified (TP)**
- Recall = (TP) / (TP + FN)
- Recall = (True Positive) / (Total actual positive)

In the case of forest fire, which one will you favour?
False Negative or False Positive?

AI Project Cycle



Hands-On: Data Collection

Hands-On: Model Training

Hands-On: Evaluate and Deploy the Trained Model

Review, Q&A, and Closing Remarks