



CRUX: INTRO TO MACHINE LEARNING WORKSHOP '22

WHAT WE ARE GOING TO COVER

1

What is machine learning?

2

Sample of the actual workshop:
Some basic ML concepts

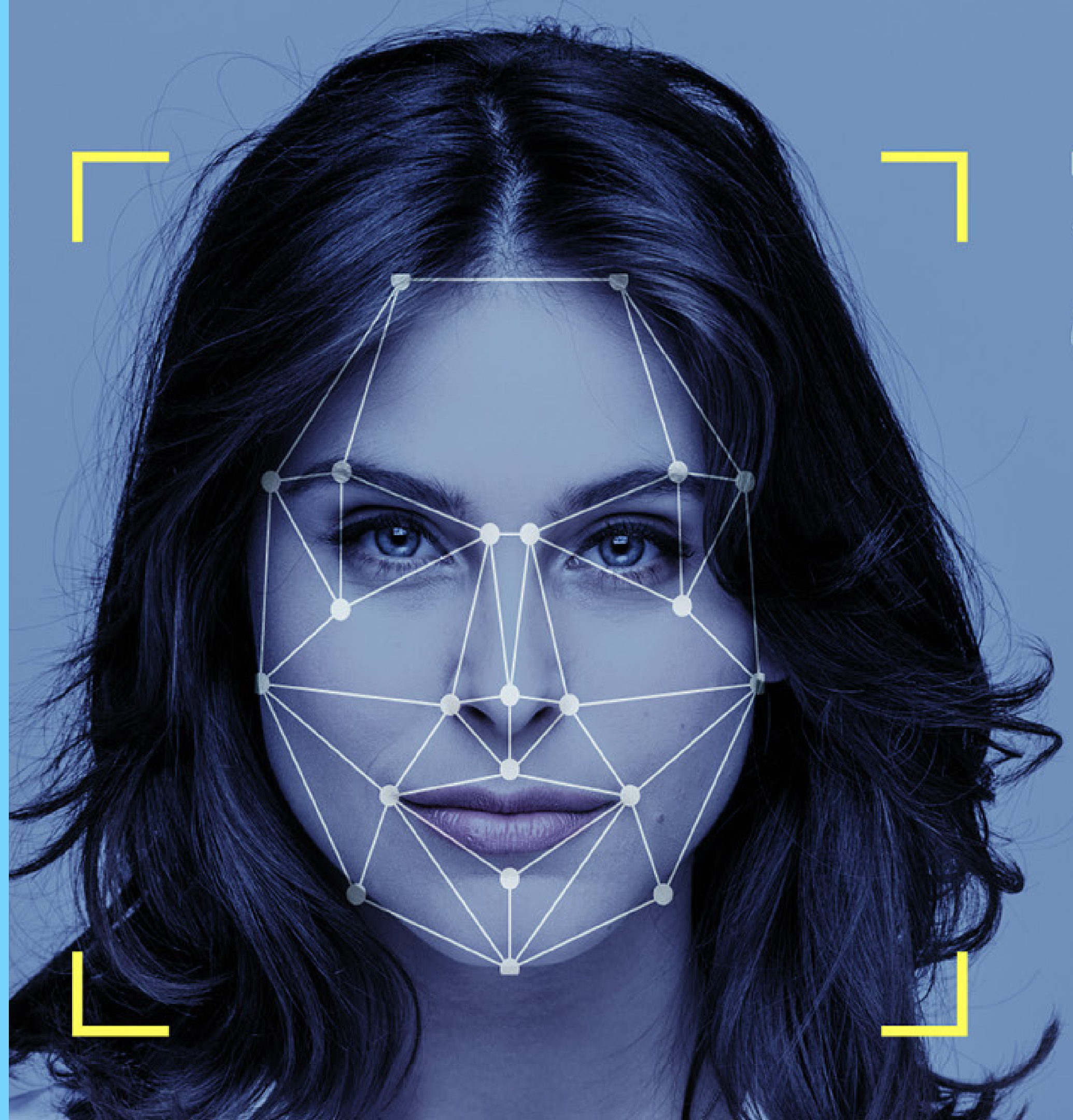
3

QnA Session

4

Outline of the main workshop

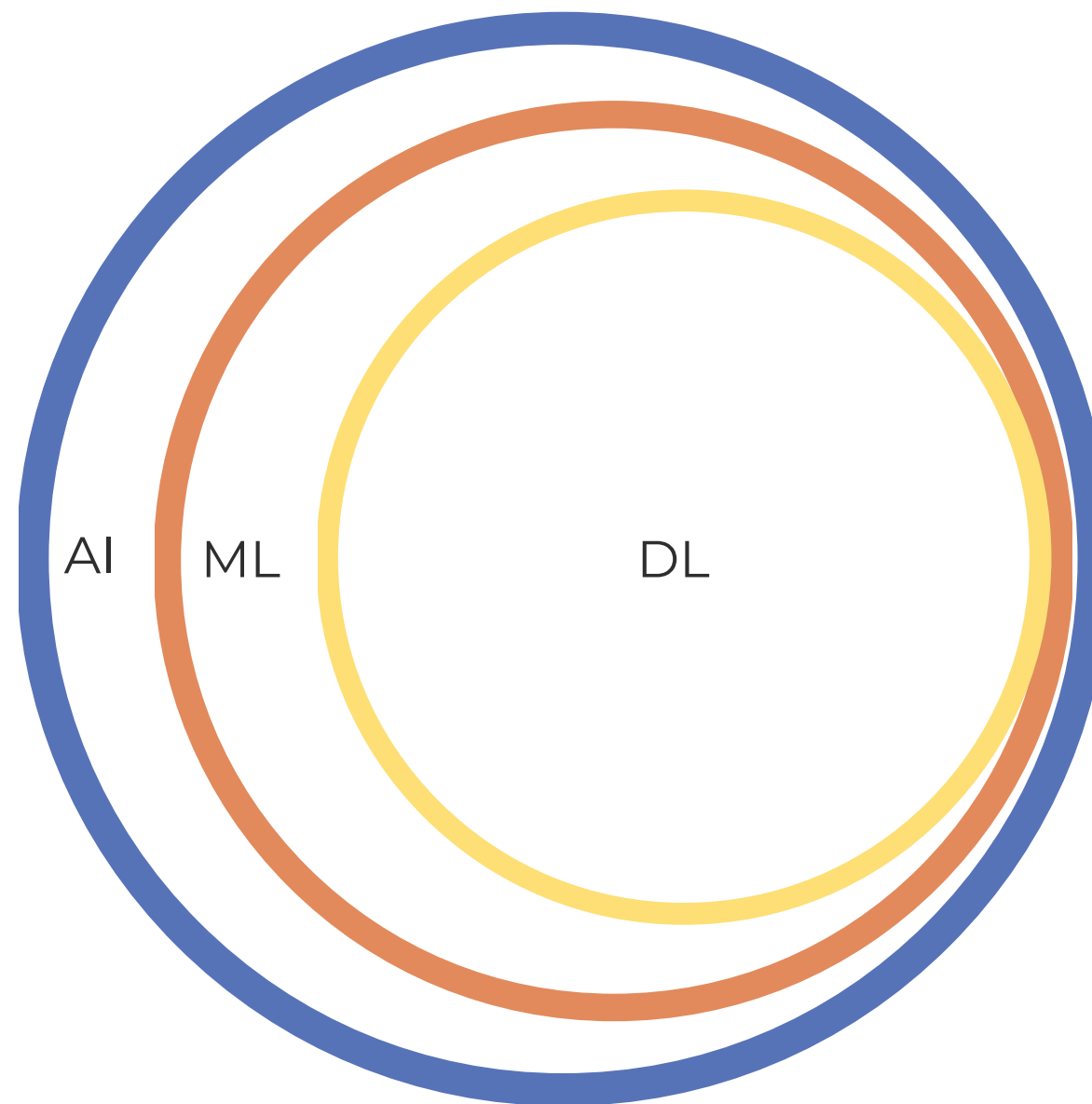
SO WHAT IS MACHINE LEARNING?



A HIGH LEVEL OVERVIEW

1

Relationship between AI, ML and DL



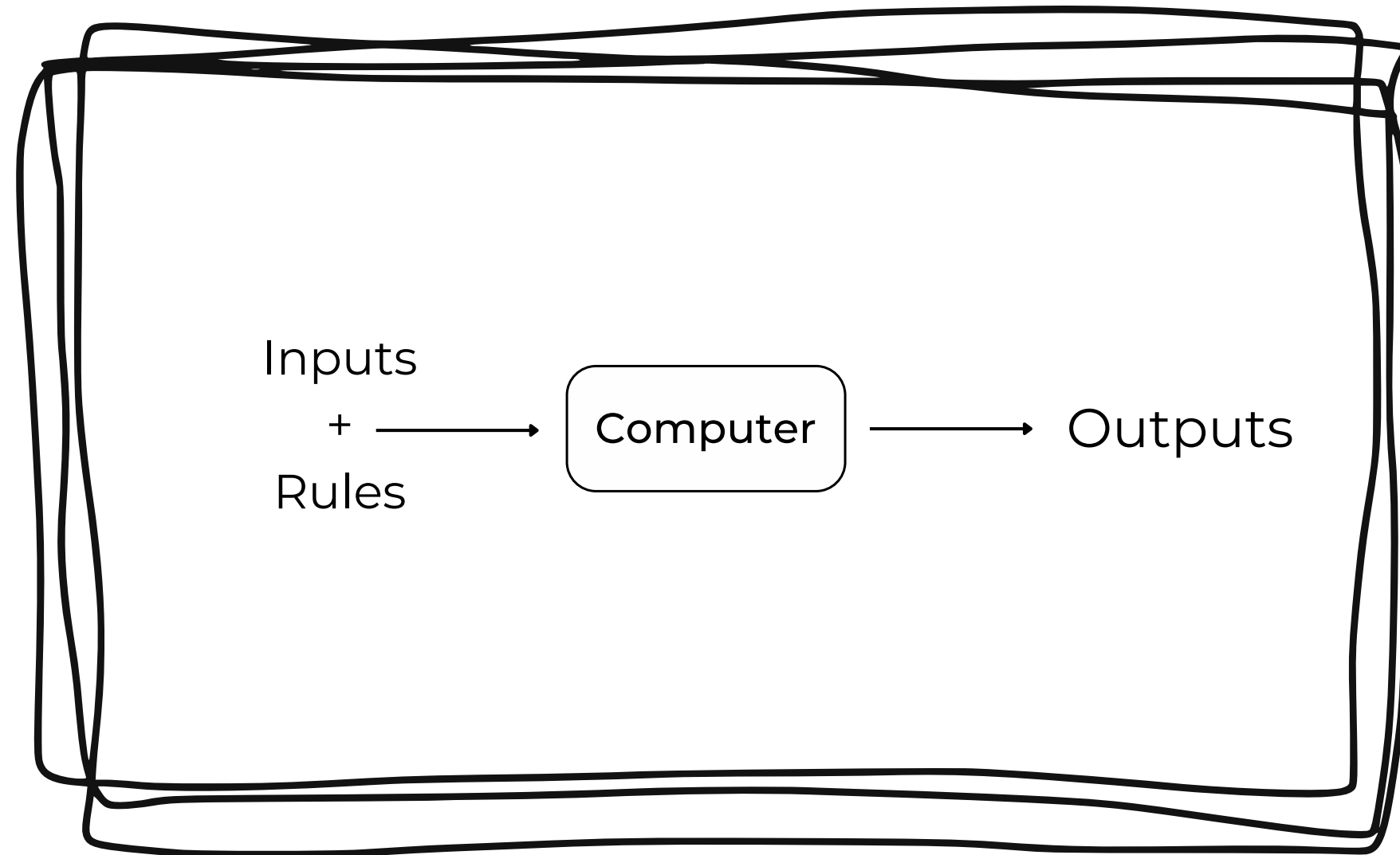
TYPES OF MACHINE LEARNING ALGORITHMS (AND/OR) PROBLEM STATEMENTS

1. Supervised Machine Learning:
 - a. Both features and labels/outputs are known
 - b. Example problem: Classifying between cats and dogs after learning from labelled images.
2. Unsupervised Machine Learning:
 - a. No knowledge of output variable
 - b. Example problem statement: Clustering in social media applications, search engines etc.
3. Reinforcement Learning:
 - a. Decision making in environments
 - b. Example problem statement: Self driving cars, AI in games etc.

A HIGH LEVEL OVERVIEW

2

Machine learning as a new approach to solving problems



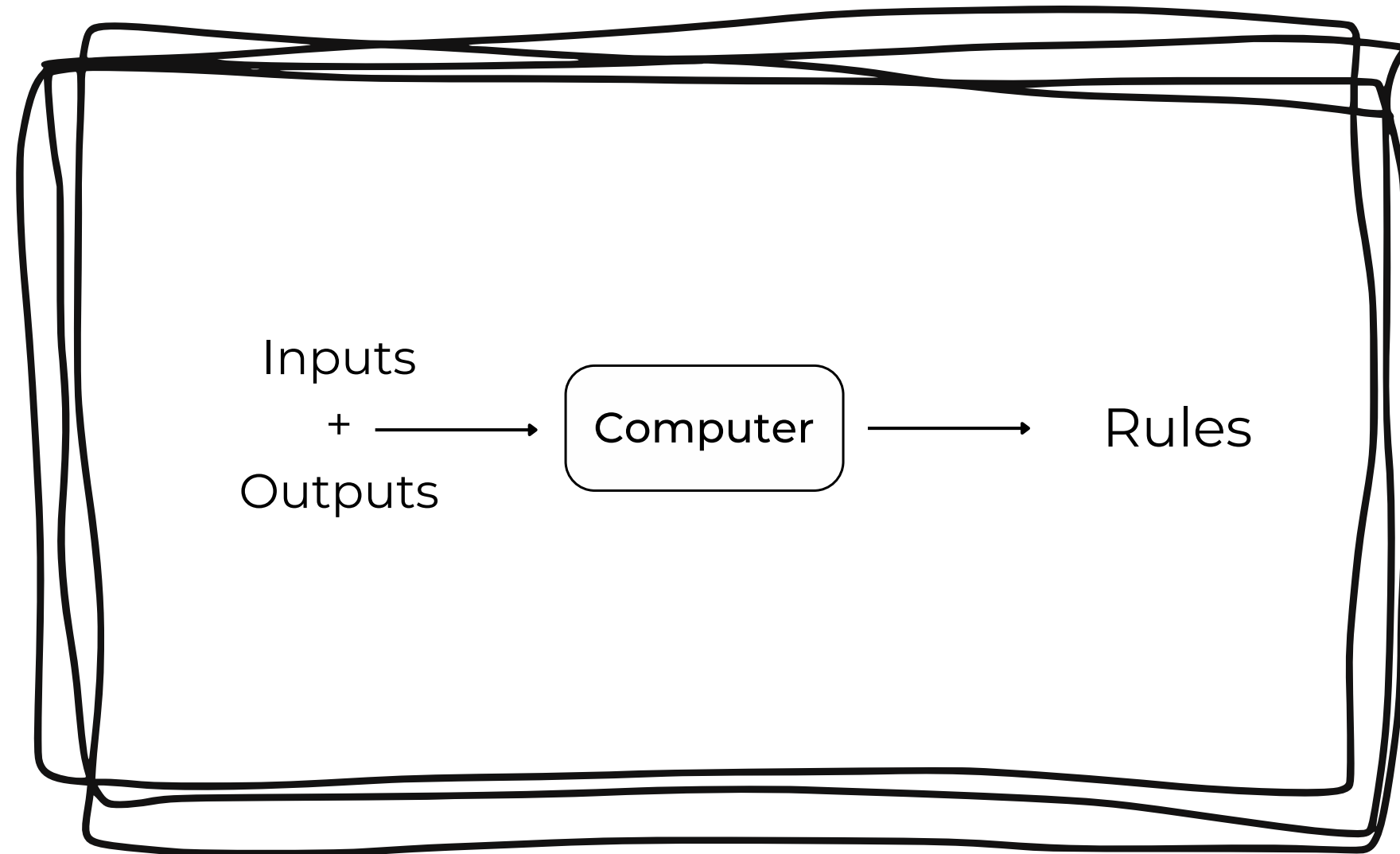
The traditional
approach of
programming

What do you think the
ML approach would
be?

A HIGH LEVEL OVERVIEW

2

Machine learning as a new approach to solving problems



The machine learning approach



Whats the point?

SCAM CALL DETECTION



Hello, this is
Alex from
Microsoft...


ML works
more
intuitively
than you
think!!

Let's go over it in a non-coding way



Let's say we have the following data about the caller
from some source:

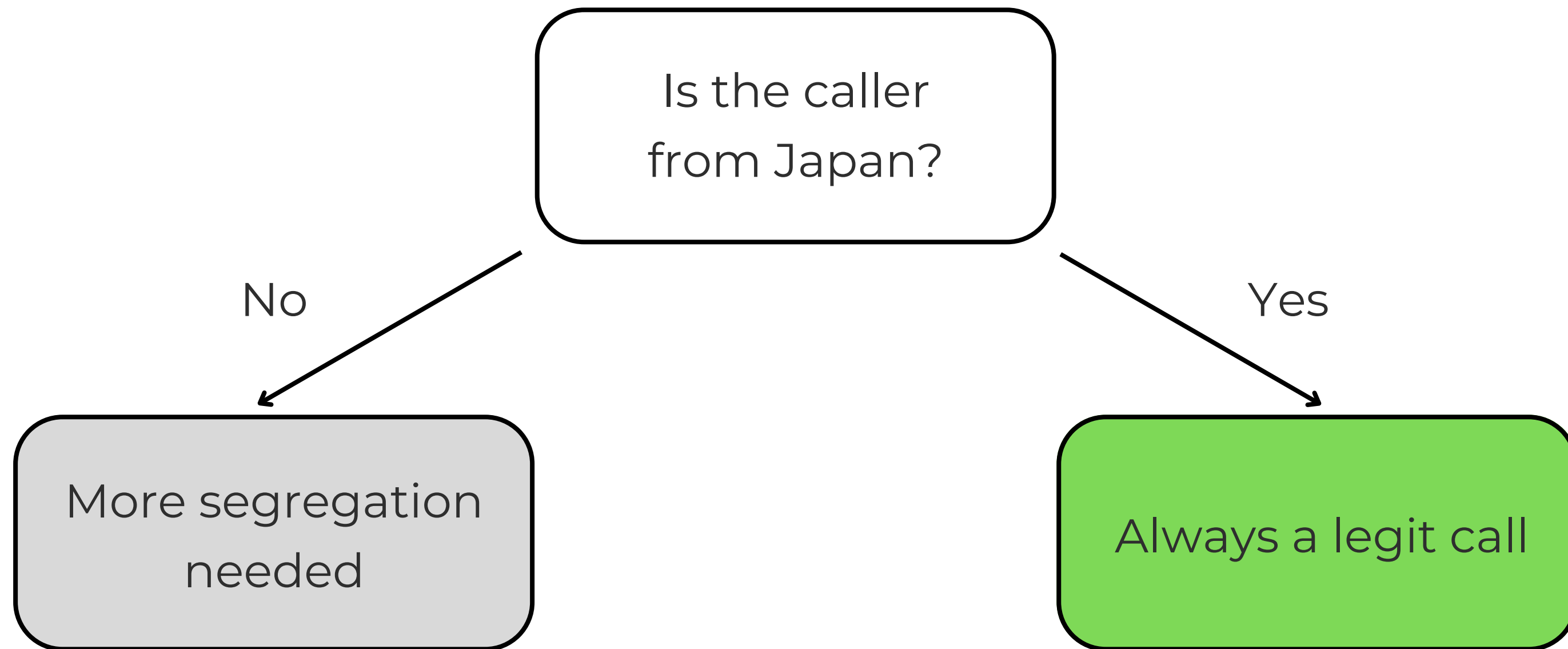
1. Their gender,
2. Age,
3. Origin
4. If the call was a phishing scam or not



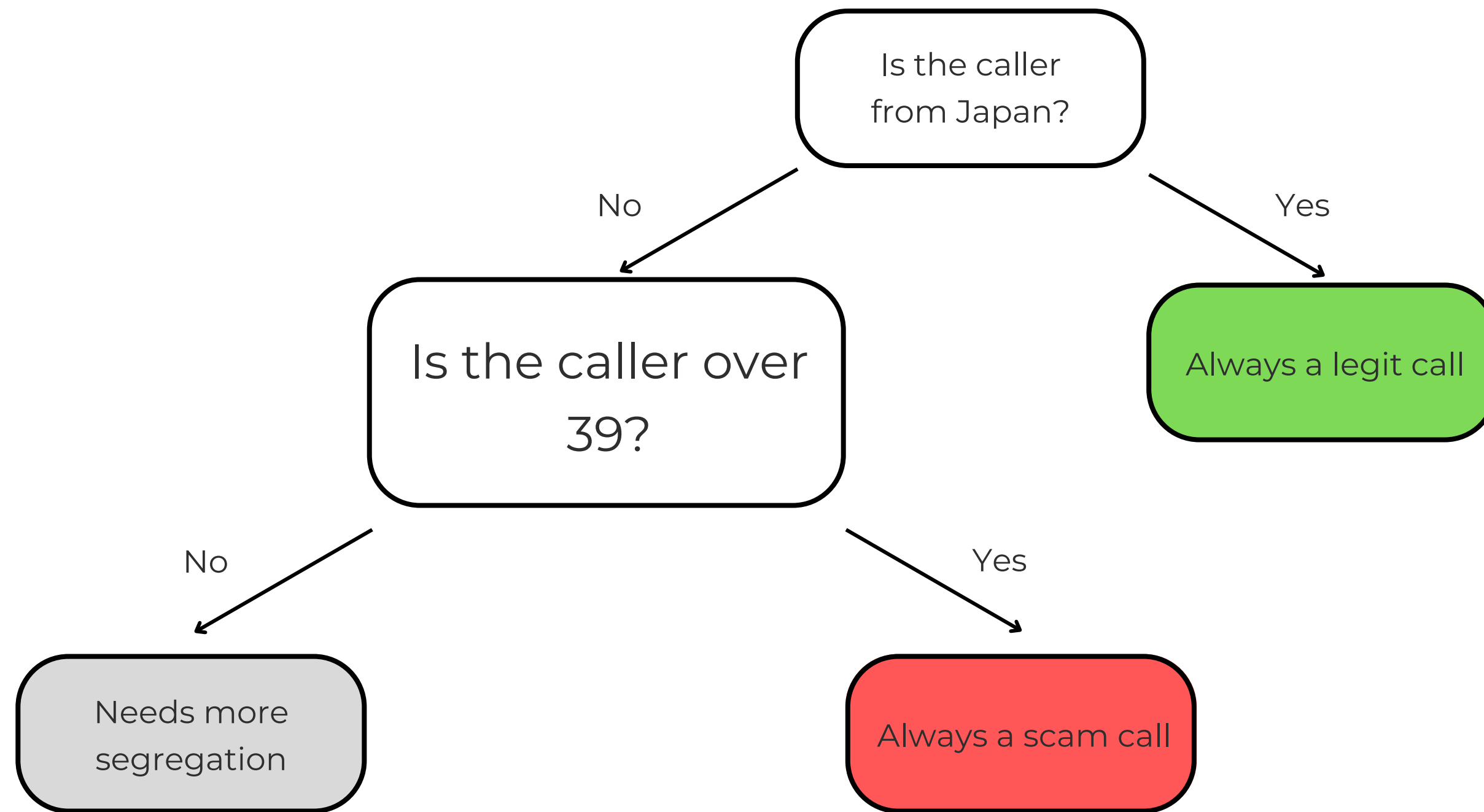
What would
you do if you
had to
program it the
traditional
way?

| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |

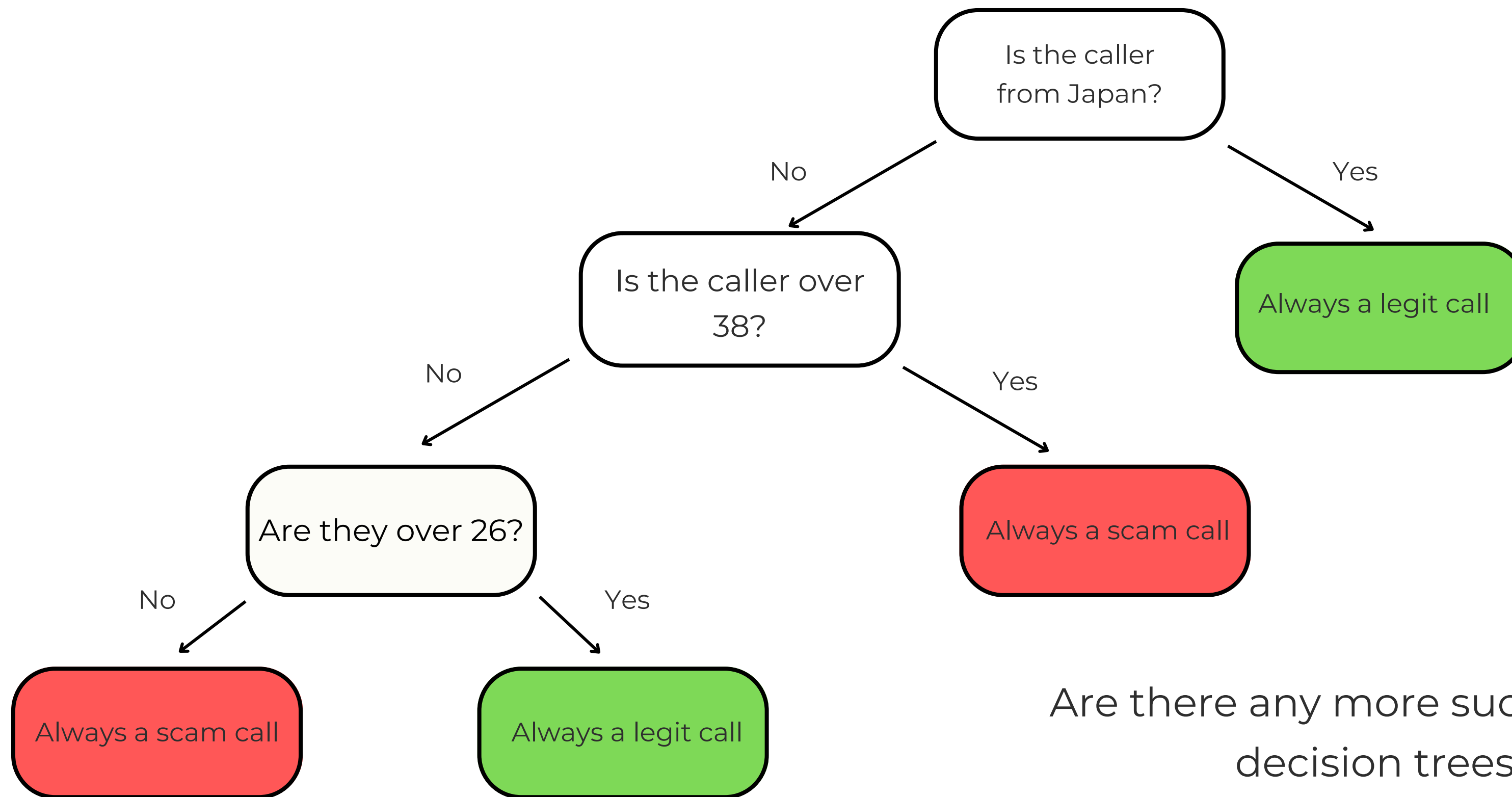
| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |



| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |



| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |



Are there any more such possible decision trees?

ISN'T THIS TOO EASY? WELL...NOT REALLY

In this particular example, everything was handed to you on a silver platter. You'll come to see there are many problems you'll face if you try to do the same thing we did earlier manually:

1. As we saw, there are multiple decision trees possible. How do we choose one from them? Are we sure it's the best one? How do I verify it, or feel confident about my model?
2. Where do I get the data from? Can the source be trusted? What am I allowed to use the data for?
3. How do I know what features are essential?
4. How tf am I supposed to explain all of this to the computer in the form of code???

Q1. RELATED TO MODELLING

- There aren't just multiple possible "decision trees" to solve that problem. There are tons of other popular algorithms that could've been used to classify the scam callers, and each one of them has various ways it can be implemented.
- This is where libraries come into play. Scikitlearn is one of the most popular machine-learning libraries that come with most of these models already coded out for you. So all that's left for you at the end is a few lines of code.
- What this workshop will aim to do is help you develop an interest and an intuition on how to quickly learn the fundamentals of whichever models you require in your projects.
- So while it won't be theory-heavy, there will be some basic math and theory you'll have to bear with, in this workshop 😞.

A LITTLE THEORY (BEAR WITH ME PLEASE)

- Gini impurity: $1 - (P(\text{scam}))^2 - (P(\text{not a scam}))^2$
- We add the impurities of both the leaves generated to get the impurity of the segregator/classifier at each stage
- We find the impurity for every possible segregator in the same fashion at each stage.
- Finally we choose the one which generates the lower impurity at each stage
- Let's quickly go back to our example, and see how to choose from various possible decision trees

| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |

- Country based segregator (India?):

- Gender based segregator (F?):

- Age based segregator (>38?):

| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |

- Country based segregator (India?):

$$(1 - 0.44 - 0.11) + (1 - 0.25 - 0.25) = 0.944$$

Note: Color coding for this and the next two slides is based on "yes" and "no" from each segregator and **not** whether its a scam or not

| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |

- Country based segregator (India?):

$$(1 - 0.44 - 0.11) + (1 - 0.25 - 0.25) = 0.944$$

- Gender based segregator (F?):

$$(1 - 0.25 - 0.25) + (1 - 0.44 - 0.11) = 0.944$$

| Age | Gender | Origin | Scam? |
|-----|--------|--------|-------|
| 23 | M | India | Yes |
| 33 | F | Japan | No |
| 37 | M | Japan | No |
| 29 | F | India | No |
| 39 | F | China | Yes |
| 43 | M | India | Yes |
| 27 | M | China | No |

- Country based segregator (India?):

$$(1 - 0.44 - 0.11) + (1 - 0.25 - 0.25) = 0.944$$

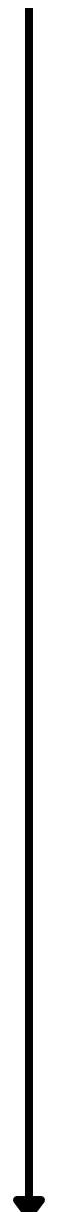
- Gender based segregator (F?):

$$(1 - 0.25 - 0.25) + (1 - 0.44 - 0.11) = 0.944$$

- Age based segregator (>38?):

$$(1 - 1 - 0) + (1 - 0.11 - 0.44) = 0.444$$

Q. How did we choose the age based on which we segregated?



| Age | Gender | Origin | Scam? |
|----------------|--------|--------|-------|
| 23 25 27 | M | India | Yes |
| 28 29 34 | M | China | No |
| 39 41 43 | F | India | No |
| | F | China | Yes |
| | M | India | Yes |

- Age based segregator (>25?):

$$(1 - 0.25 - 0.25) + (1 - 1 - 0) = 0.5$$

- Age based segregator (>28?):

$$(1 - 0.11 - 0.44) + (1 - 0.25 - 0.25) = 0.94$$

- Age based segregator (>34?):

$$(1 - 1 - 0) + (1 - 0.11 - 0.44) = 0.44$$

- Age based segregator (>41?):

$$(1 - 1 - 0) + (1 - 0.25 - 0.25) = 0.5$$

Q2. RELATED TO DATA (FEATURE EXTRACTION, ENGINEERING, ETC.)

- This is probably one of the hardest parts of any ML/Data Science project. And while it doesn't go completely away, a concept we spoke briefly about earlier (remember deep learning?) makes things become considerably easier in many cases.
- Feature extraction, engineering, and selection.
- Popular sources of data: Kaggle, Google Datasets, and even GitHub (sometimes).
- Some problem statements just can't be done with simple machine learning models with the data at hand (at least if you want good results).
- Less amount of data at hand: transfer learning?

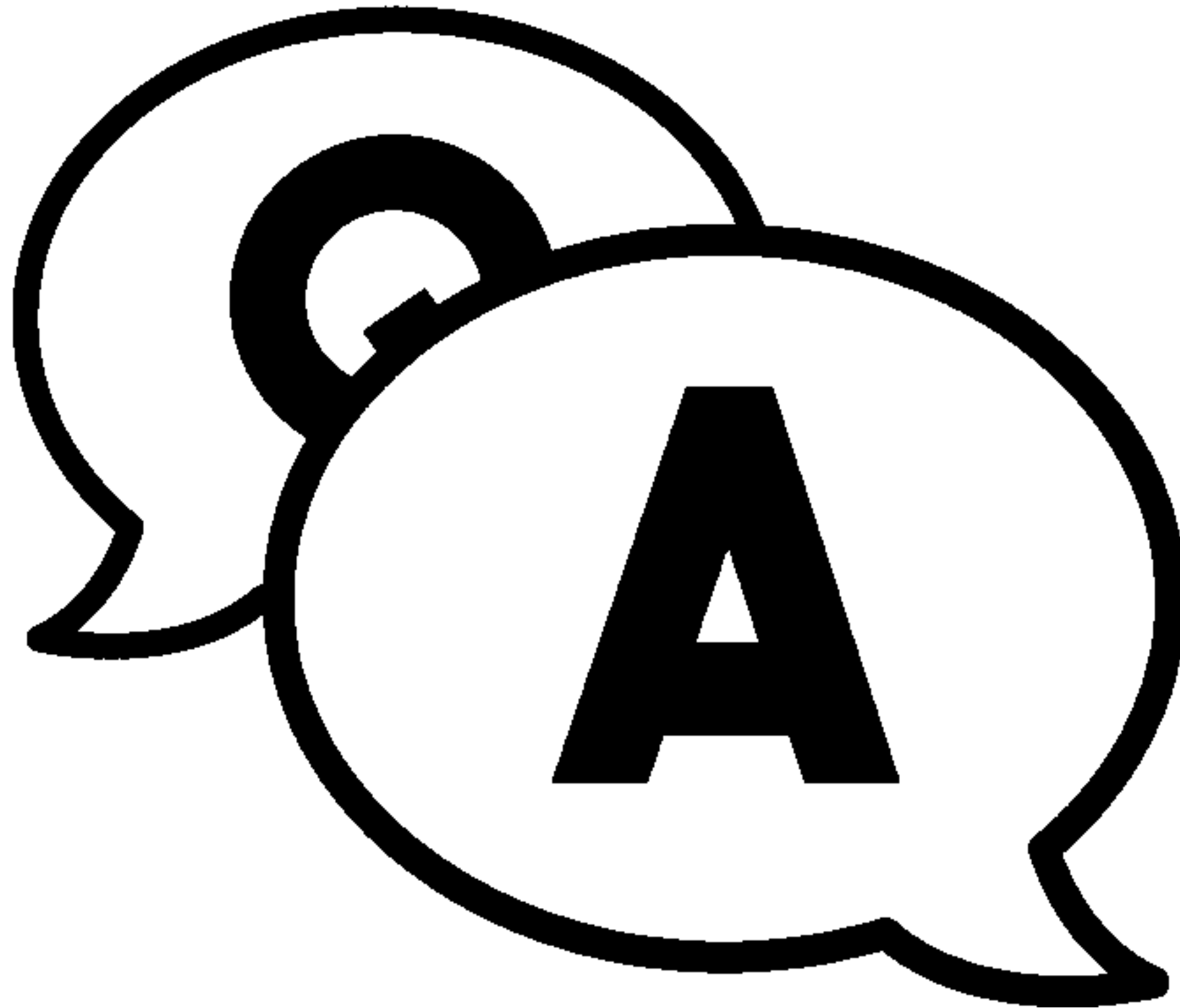
FEELING TIRED?



464762

Q3. RELATED TO TESTING

- Train-test-validation split
- Overfitting
- Underfitting
- Loss functions
- Metrics
- How math helps validating the final result





i.am.ai

AI Roadmap

AI Roadmap

Follow these roadmaps to become an Artificial Intelligence expert.

AMAI AI Roadmap

JANUARY WORKSHOP OUTLINE:

PREREQUISITES (Resources will be mailed)

1. Python
2. Virtual environments
3. Numpy, Pandas, Plotly/Matplotlib

Remember the goal is to learn the fundamentals and start finding the approach more intuitive!

Weekend 1



ML Basics +
Classification

Weekend 2



Regression +
Clustering

Weekend 3



Intro to Neural
Nets

Weekend 4



Specialization/ML
Contest

FEEDBACK



THANK YOU

Arunachala Amuda Murugan
Phone Number: +91 96635 65683

CRUx contact details:

Prez: Aayushi Sah (+91 83691 16037)
Sec: Srikanth Tangirala (+91 98212 95377)
Sec: Varun Vaddiraju (+91 99493 34384)