The k-nearest neighbours algorithm

I/ Introduction: Taxicab geometry and the Manhattan distance

Taxicab geometry is a form of geometry, where the distance between two points A and B is not the length of the line segment AB as in the **Euclidean geometry**, but is calculated along a grid.
Taxicab geometry is very similar to Euclidean coordinate geometry. However it is meant to act as a better model of urban geography than Euclidean coordinate geometry. Nonetheless, taxicab geometry is an idealized model, so there are some basic assumptions that simplify working with this geometry:

1) the horizontal and vertical lines of the grid represents streets;

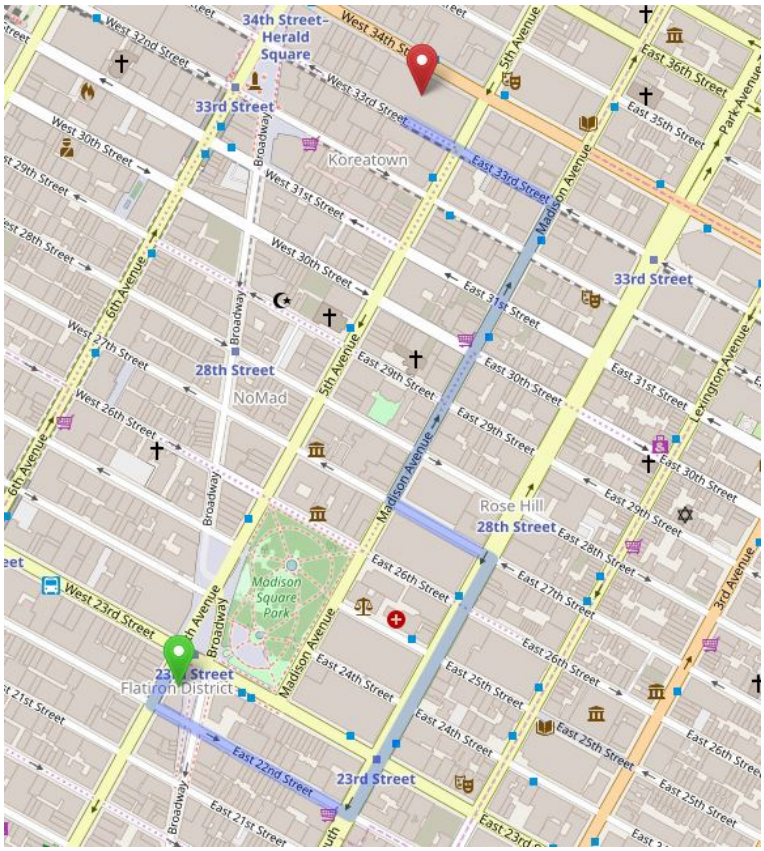
2) points can only be located at grid intersections;

3) numerical coordinates will always be integers;

4) the **taxicab distance** (or **Manhattan distance**) between two points is the smallest number of grid units that an imaginary taxi must travel to get from one point to the other.

Let $(x_A\,;\,y_A)$ and $(x_B\,;\,y_B)$ be the coordinates for two points A and B, then the Manhattan distance between those points is given by:

$$AB = |x_B - x_A| + |y_B - y_A|\,.$$

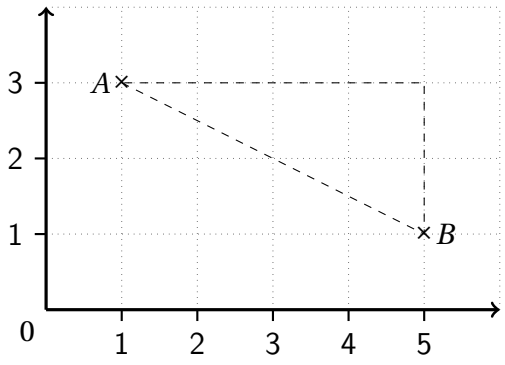


From OpenStreetMap, route from the Flatiron building to the Empire State Building for a car, in the streets of Manhattan

Example :

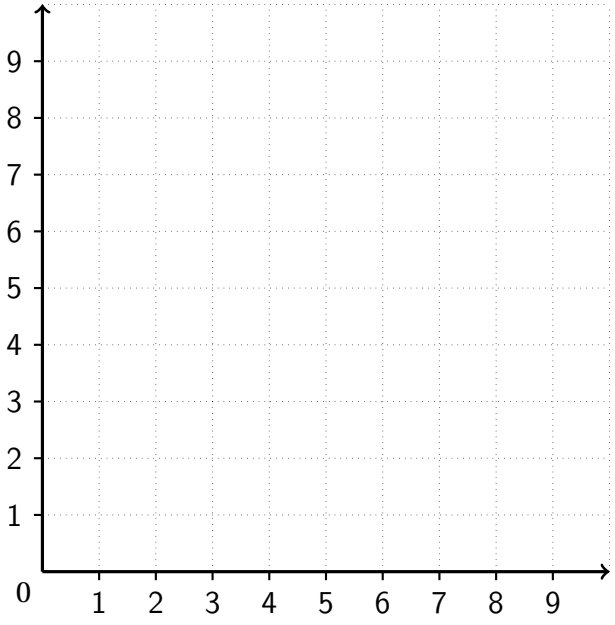
The taxicab distance AB is:

$$\begin{aligned} |x_B - x_A| + |y_B - y_A| &= |5-1| + |1-3| \\ &= 4+2 \\ &= 6 \end{aligned}$$



Questions :

1) Draw the following points on the coordinate system below: $A(1,1)$; $B(1,5)$; $C(5,1)$; $D(5,5)$; $E(7,9)$; $F(6,3)$.



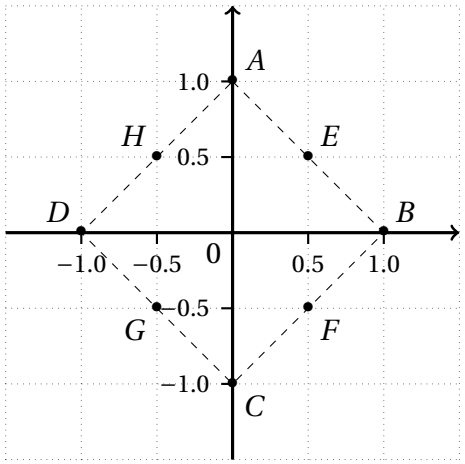
2) Fill in the following table:

Points	AB	AC	BC	DA	AE	EF
Taxicab distance						
Euclidean distance						

To calculate the Euclidean distance between A and B:

$$AB = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

3) When do you think the Taxicab distance is the same as the Euclidean distance?
4) State the definition of a circle.
5) Work out the distances OA, OB, OC, OD, OE, OF, OG and OH for the taxicab geometry.



6) What is the shape $ABCD$ for the Euclidian geometry?
7) What do you think it is for the Taxicab geometry?

II/ Machine learning basics

Watch the video, then try to complete its script.

Rewatch the video

The world is filled with ....., a lot of .....: pictures, music, words, spreadsheets, videos, and it doesn't look like it's going to slow down anytime soon. ..... brings the promise of deriving meaning from .....

Arthur C. Clarke famously once said: "Any sufficiently ..... is indistinguishable from ......"

I found ..... not to be ....., but rather ..... that you can utilize to .....

This is Cloud AI Adventures. My name is Yufeng Guo, and each episode, we will be exploring the art, science and tools of ..... Along the way, we'll see just how easy it is to create amazing experiences and yield valuable insights.

The value of ..... is only just beginning to show itself.

There is a ..... in the world today generated not only by ....., but also by ....., ... and ..... This will only continue ..... in the years to come.

Traditionally, humans have ..... and ..... to the ..... However, as the ..... surpasses the ability for humans to make sense of it and manually write those rules, we will turn increasingly to ..... that can ..... and importantly, ..., to ..... to a .....

We see machine learning all around us in the products we use today. However, it isn't always apparent that machine learning is behind it all. While things like ..... inside of ..... are clearly machine learning at play, it may not be immediately apparent that ..... is also powered by machine learning.

Of course, perhaps the biggest example of all is ..... Every time you use ....., you're using a system that has many machine learning systems at its core, from ..... of your ..... to ..... based on your ....., such as knowing which results ..... when searching for ..... depending on whether you're a ..... or a ..... – perhaps you're both.

Today, machine learning's immediate applications are already quite wide-ranging, including ....., ... and ....., as well as ..... too.

These powerful capabilities can be applied to a wide range of fields, from diabetic retinopathy and ..... to ..... and of course, ... in the form of .....

It wasn't that long ago that when a company or product had machine learning in its offerings, it was considered novel. Now, every company is pivoting to use machine learning in their products in some way. It's rapidly becoming, well, an ..... Just as we expect companies to have a ..... that works on your mobile device or perhaps an app, the day will soon come when it will be expected that our technology will be .....

As we use machine learning to make human tasks ..... than before, we can also look further into the future when machine learning can help us do tasks that we never

could have ..... . Thankfully, it's not hard to take advantage of machine learning today.

The tooling has gotten quite good. All you need is ....., ... and a willingness to take the plunge. For our purposes, I've shortened the definition of machine learning down to just five words: ..... .

While I wouldn't use such a short answer for an essay prompt on exam, it serves a useful purpose for us here. In particular, we can split the definition into two parts: ..... and ..... . These two pieces broadly outline the two sides in machine learning, both of them equally important.

..... is what we refer to as ....., while ..... is referred to as ..... .

Now let's drill into those two sides briefly for a little bit. ..... refers to ..... to ..... the creation and fine tuning of a ..... . This ..... can then be used to serve up ..... on previously unseen data and answer those questions. As ..... is gathered, the model can be ..... over time and ..... deployed.

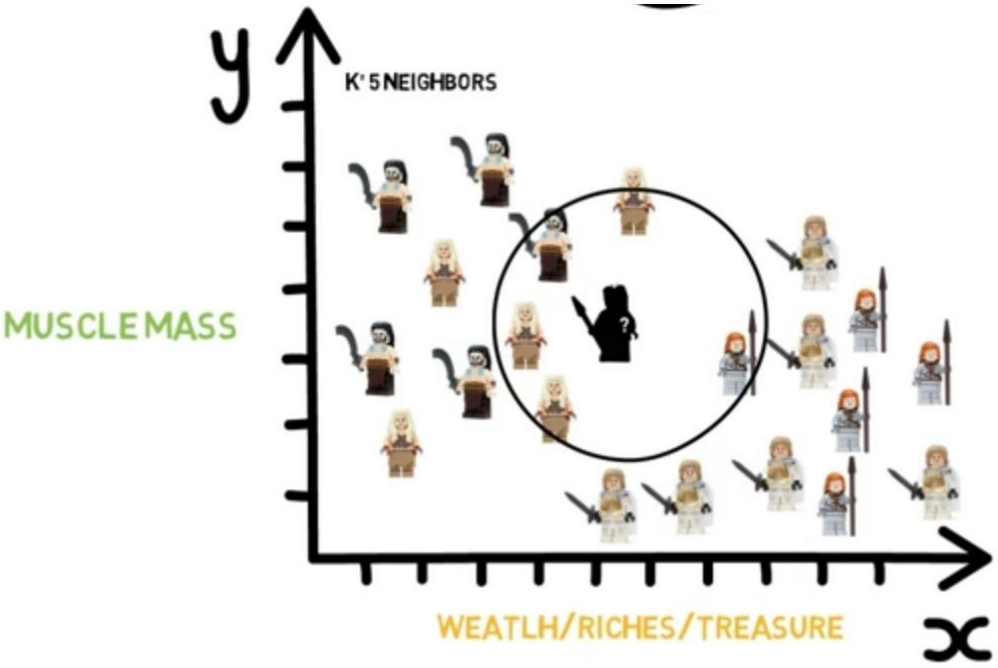
As you may have noticed, the key component of this entire process is data. Everything hinges on data. Data is the key to unlocking machine learning, just as much as machine learning is the key to unlocking that hidden insight in data.

This was just a high level overview of machine learning – why it's useful and some of its applications. Machine learning is a broad field, spanning an entire family of techniques when ..... from data. So in future episodes, we'll aim to give you a better sense of what approaches to use for a given data set and question you want to answer, as well as provide the tools for how to accomplish it.

III/ Discovering the k-NN algorithm with *Game of Thrones*

The **k-nearest neighbours algorithm** is one of the algorithm that is used in the field of artificial intelligence. It takes part in various fields of machine learning.
It is an example of what is called **supervised learning**: taking in data for which we already know what the result should be and using this to 'train' a chosen algorithm. We can then use a 'test set' of this known data to determine how well the algorithm has learnt the proper way to give us the correct output. The 'accuracy' of the algorithm is judged by how many answers it got right on its test.
The k-nearest neighbours algorithm (k-NN) is a method used for **classification**. It can be used in scenarios where there is little or no prior knowledge about distribution data. As the name suggests, the k-NN algorithm tries to find a predefined number (that number being k) of data points surrounding the point being investigated, so that it can predict the label for the investigated point based on the surrounding ones, using the most common label in the chosen neighbourhood.
Let's take a simple example from *Game of Thrones* to understand how it is working. Suppose we have to design a classifier to determine whether an unknown person is a Dothraki or a Westerosian. We can use two features to classify or predict which clan the person belongs to. So for this example, we can use muscle mass, wealth and riches as our independent variables (in machine learning, those variables are called **features**).
For the Dothraki, lets assume that they have a greater muscle mass and, for the Westerosians, that they are high in wealth/riches, but their muscle mass is significantly lower than that of a Dothraki.
Let's have a try with $k = 5$, meaning that we're going to look to the 5 nearest neighbours of our unknown person, given the muscle mass and the wealth of that person. The chosen distance is the Euclidean distance.

1) Using the chart below, can you guess if our unknown person is a Dothraki or a Westerosian?



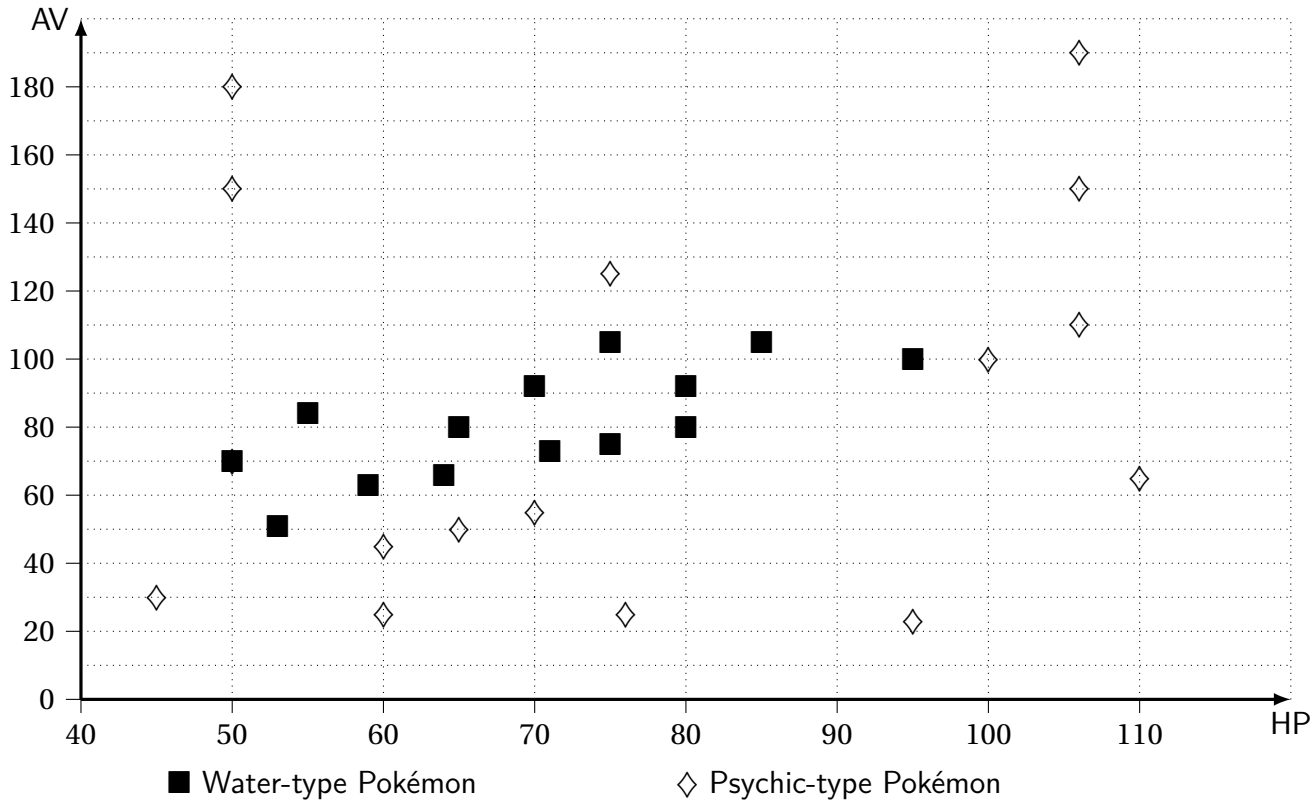
2) Assuming that each graduation represents 1 unit on both axes, place an unknown person with 7 in wealth and 4 in muscle mass. Using the 5 nearest neighbours to this person, predict the clan of this person.

IV/ Finding the type of a mystery Pokémon

Let's assume that Pokémon have only two characteritics: their Health Points (HP) and their Attack Value (AV). We also assume that there are only two types of Pokémon: Water-type and Psychic-type. As an example, we could train our k-NN algorithm with a table starting that way:

Name	Finneon	Deoxys	Chimecho	Grumpig	Politoed
HP	49	50	75	80	90
AV	49	95	50	45	75
Type	Water	Psychic	Psychic	Psychic	Water

1) Place the five points corresponding to the five Pokémon above on the following graph, on which 29 points are already placed, according to a database.

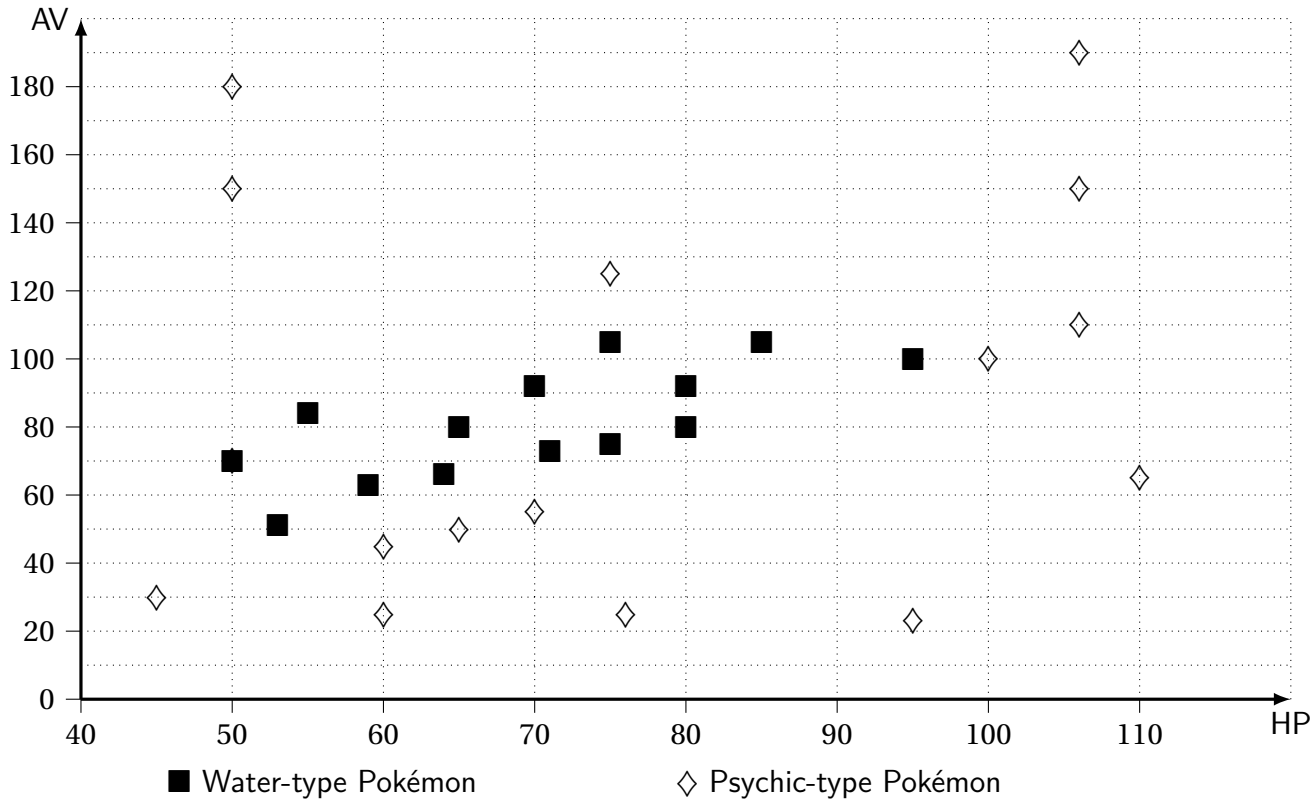


2) Using the 5 nearest neighbours and the Euclidean distance on the previous graph, predict the type of a mystery Pokémon, knowing that it has 65 HP and 40 AV.

3) What if we choose $k = 11$? $k = 34$?

4) And what is the prediction if we use the Manhattan distance?

5) Answer the same questions for a mystery Pokémon with 90 HP and 150 AV.



V/ Helping the Sorting Hat to choose a new student's House

The Sorting Hat is a magical hat at Hogwarts that determines which of the four school Houses each new student belongs most to. These four Houses are Gryffindor, Hufflepuff, Ravenclaw, and Slytherin. Its decision is based on the courage, loyalty, wisdom and mischief of the new student.
The Sorting Hat have to its disposal a table (*on the next page*) in which are recorded the data for a sample of 50 Hogwarts students.

In the following table are the characteristics of the new students the Sorting Hat wants to place in the right Houses. We're going to help the magical hat to take its decision, using the k-NN algorithm.

Name	**Courage**	**Loyalty**	**Wisdom**	**Mischief**
Hermione	8	6	6	6
Drago	6	6	5	8
Cho	7	6	9	6
Cedric	7	10	5	6

1) We will use the Manhattan distance to calculate the distance between two students:

$$\text{distance(student1, student2)} = |c_1 - c_2| + |l_1 - l_2| + |w_1 - w_2| + |m_1 - m_2|\,.$$

a) According to that formula, check that the distance between Hermione and Adrian is equal to 8.

b) What is the distance between Arthur and Drago?

2) Write an algorithm that calculates the distance between Hermione and any other student.

3) Split the class into four groups. Each group will help the Sorting Hat for one of the four new students.

4) **a)** For the student you've chosen in your group, create on your calculator (or on a computer, using Python) a program to calculate the distance between your student and any other student.

b) Using your program, work out the distance between your student and the 50 students of the data table. You may add one column to the table.

5) Using the 7 nearest neighbours, what is the House you can tell the Sorting Hat to choose for your student?

6) Is the result consistent with the books/movies? You may check on the *Harry Potter Wiki* if you don't know the House of your student: https://frama.link/HPwiki.

Sources

- *The shortest way*, Arnaud Moragues, Emilangues.
- *Learning Machine Learning (at Hogwarts)*, Tom McKenzie.
- *An Introduction to KNN (K Nearest Neighbours). The Game of Thrones way*, Pratik Kotian.
- *Prépabac NSI 1re*, Céline Adobet, Guillaume Connan, Gérard Rozsavolgyi et Laurent Signac, Hatier, 2019

Name	Courage	Loyalty	Wisdom	Mischief	House
Adrian	9	4	7	10	Slytherin
Andrew	9	3	4	7	Gryffindor
Angelina	10	6	5	9	Gryffindor
Anthony	2	8	8	3	Ravenclaw
Arthur	10	4	2	5	Gryffindor
Bellatrix	10	4	9	9	Slytherin
Bole	7	4	6	10	Slytherin
Colin	10	7	4	7	Gryffindor
Cormac	9	6	5	4	Gryffindor
Dean	9	8	4	7	Gryffindor
Demelza	10	6	5	3	Gryffindor
Derrick	5	4	6	5	Slytherin
Eddie	5	7	10	3	Ravenclaw
Ernie	4	8	7	4	Hufflepuff
Euan	9	2	7	4	Gryffindor
Gilderoy	7	9	9	9	Ravenclaw
Gregory	6	9	7	8	Slytherin
Hannah	8	10	2	4	Hufflepuff
Harper	6	3	5	10	Slytherin
Jimmy	9	9	9	10	Gryffindor
Justin	5	10	7	10	Hufflepuff
Katie	10	2	3	9	Gryffindor
Lavande	10	8	8	6	Gryffindor
Lee	10	2	2	8	Gryffindor
Luna	2	9	9	2	Ravenclaw
Marcus	6	5	8	10	Slytherin
Marietta	10	8	10	9	Ravenclaw
Michael	4	2	6	5	Ravenclaw
Milicent	9	3	5	6	Slytherin
Mimi	4	4	9	10	Ravenclaw
Montague	5	7	2	10	Slytherin
Neville	10	5	6	4	Gryffindor
Norbert	3	10	7	6	Hufflepuff
Nymphadora	2	5	3	8	Hufflepuff
Padma	6	6	6	9	Ravenclaw
Paenny	2	8	9	8	Hufflepuff
Pansy	4	4	10	8	Slytherin
Parvati	10	5	2	6	Gryffindor
Pomona	5	10	7	8	Hufflepuff
Quirinus	7	10	10	2	Ravenclaw
Roger	9	10	10	8	Ravenclaw
Romilda	10	6	2	9	Gryffindor
Saemus	7	4	8	3	Gryffindor
Sirius	10	8	10	7	Gryffindor
Susan	5	6	5	5	Hufflepuff
Susan	4	10	10	5	Hufflepuff
Ted	5	9	8	4	Hufflepuff
Terence	6	9	2	8	Slytherin
Terry	8	4	10	5	Ravenclaw
Vincent	4	9	2	10	Slytherin