



PARAGON
INTERNATIONAL UNIVERSITY
C A M B O D I A

Khmer Word Segmentation

Instructor: Rina Buoy

Team Members: Porchhay Ngov

Sreyneang Cho

Monespiseth Ly

TABLE OF CONTENTS

01 INTRODUCTION

02 METHODOLOGY

03 IMPLEMENTATION

04 FUTURE WORK

05 DEMO

06 Q&A

WHO INVOLVED IN THIS PROJECT



Rina Buoy

Instructor



Porchhay Ngov

CS Department
Senior



Sreyneang Cho

CS Department
Senior



Monespiseth Ly

CS Department
Junior

Abstract geometric shapes, including triangles and polygons, in various shades of blue and white, located in the top-left corner of the slide.

01

INTRODUCTION

Abstract geometric shapes, including triangles and polygons, in various shades of blue and white, located in the bottom-right corner of the slide.

OVERVIEW

Word segmentation is a problem that involves breaking down a string of written language into its corresponding words.

For example:

Original sentence	ယွှန်းရွာတဲ့ ပြောရတဲ့လူ
Word separation	ယွှန်း ရွာတဲ့ ပြောရတဲ့ လူ



OBJECTIVE

- Effectively segments words and determine word boundaries from a given phrase/article
- Able to identify multiple possible meanings as Khmer sentences can have multiple valid word boundaries that still produce perfectly valid meaning

PROBLEM

Since Khmer words does not have a standard for separating each word like how spacing works in English, we need a way to analyze a sentence or article and able to break it down to its individual words.

For example:

English: I go to school.

Khmer: ខ្ញុំទៅសាលារៀន។



02 METHODOLOGY

TECHNOLOGY



HTML/CSS/JS



VUE.js



TAILWIND



PYTHON



PYTORCH



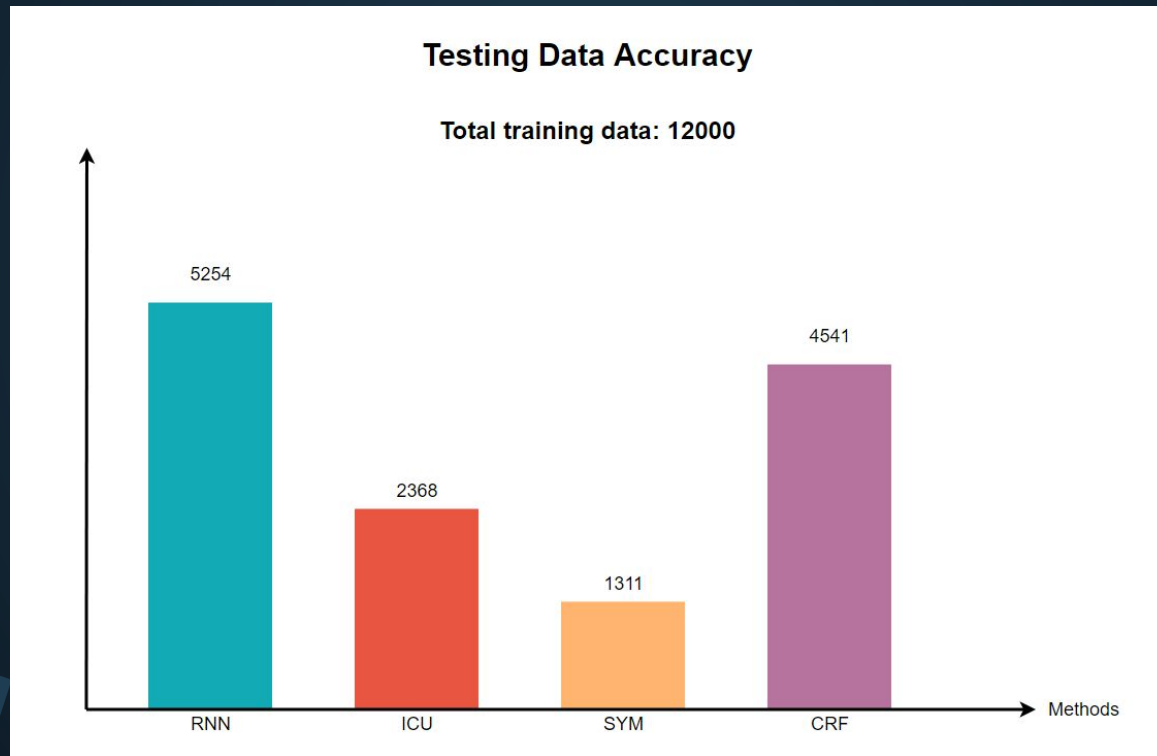
FAST API

The slide features a dark blue background with several light blue geometric shapes, including triangles and polygons, scattered around the edges. The title 'OUR SEGMENTATION METHODS' is centered at the top in a bold, white, sans-serif font.

OUR SEGMENTATION METHODS

- Symspell
- International Components for Unicode (ICU)
- Conditional Random Field (CRF)
- Recurrent Neural Network (RNN)

SEGMENTATION METHODS TESTING FROM A KHMER ARTICLE





03 IMPLEMENTATION

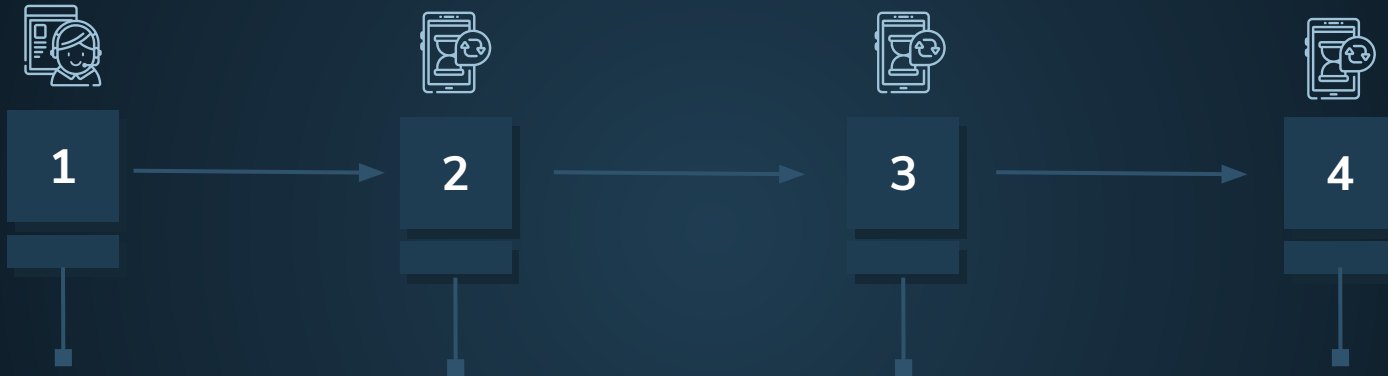


WHAT IS RECURRENT NEURAL NETWORK (RNN)?

Recurrent Neural Network is a type of neural network that uses their reasoning from previous experiences to inform the upcoming events

RNN is very common in Natural Language Processing (NLP) problems such as Khmer word segmentation.

IMPLEMENTATION PROCESS USING RNN



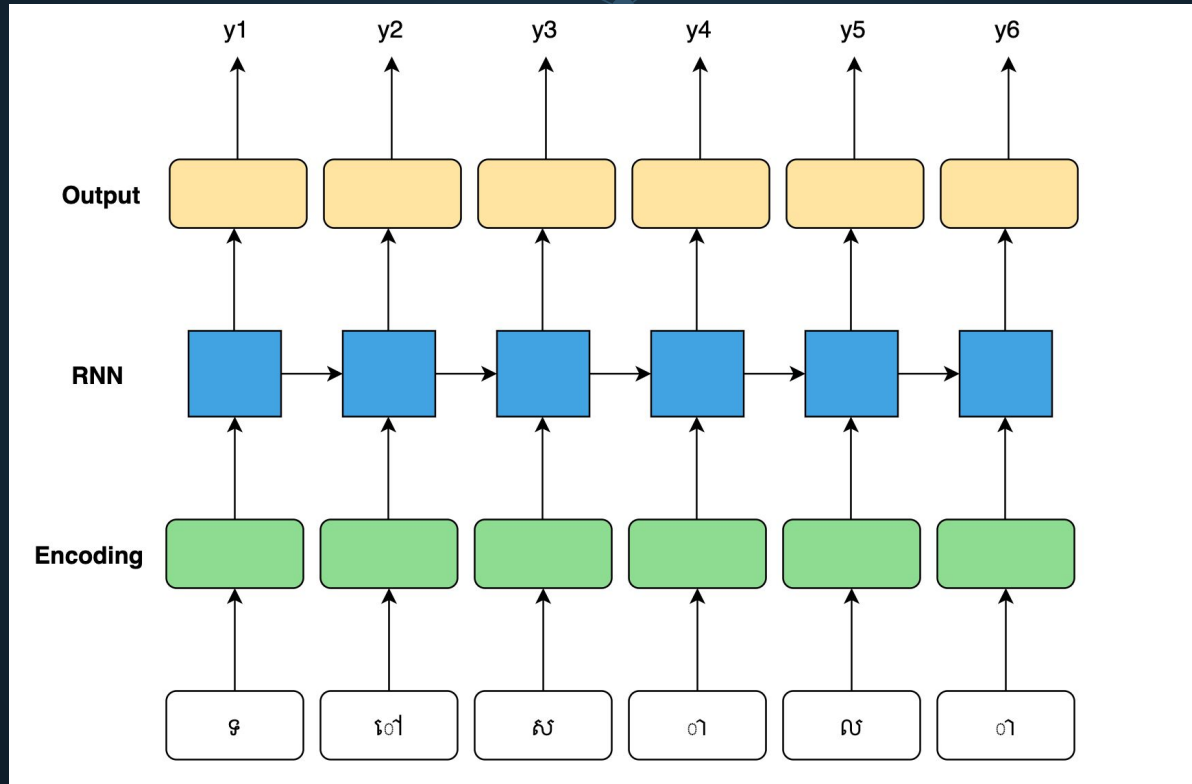
Scrape data from Khmer news articles and manually segment the articles

Feed the segmented data in to our Recurrent Neural Network

Train data by splitting into 80% for training, 20% for testing

Given a sentence, the neural net model will determine where is the start of each words in a sentence.

RNN Architecture





04 FUTURE WORK

FUTURE WORK

Implement our project into:

- SPELL-CHECK
- PART-OF-SPEECH
- VOICE-RECOGNITION



05 DEMO



06 Q&A

THANKS!

Do you have any questions?

IMPLEMENTATION PROCESS



1



2



3



4

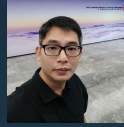
Scrape data from
Khmer news article
and manually
segment the articles

Feed the
segmented data in
to our Recurrent
Neural Network

NEPTUNE

Neptune is the
farthest planet

WHO INVOLVED IN THIS PROJECT



Rina Buoy

Instructor



JAMES SCOT

You can tell more
about this person



LINDA SMITH

You can tell more
about this person



SUSAN DOE

You can tell more
about this person

WHO

25%



MARS

Despite being
red, it's cold

50%



JUPITER

Jupiter is a very
big planet

75%



SATURN

Saturn is the
ringed planet

95%



MERCURY

Mercury is the
smallest planet

OUR VALUES



MISSION

Saturn is the ringed planet
and a gas giant



VISION

Neptune is the farthest
planet from the Sun

SYSTEM ADMIN PROFILE

MARS

Despite being red, it's cold



SATURN

Saturn is the ringed planet



JUPITER

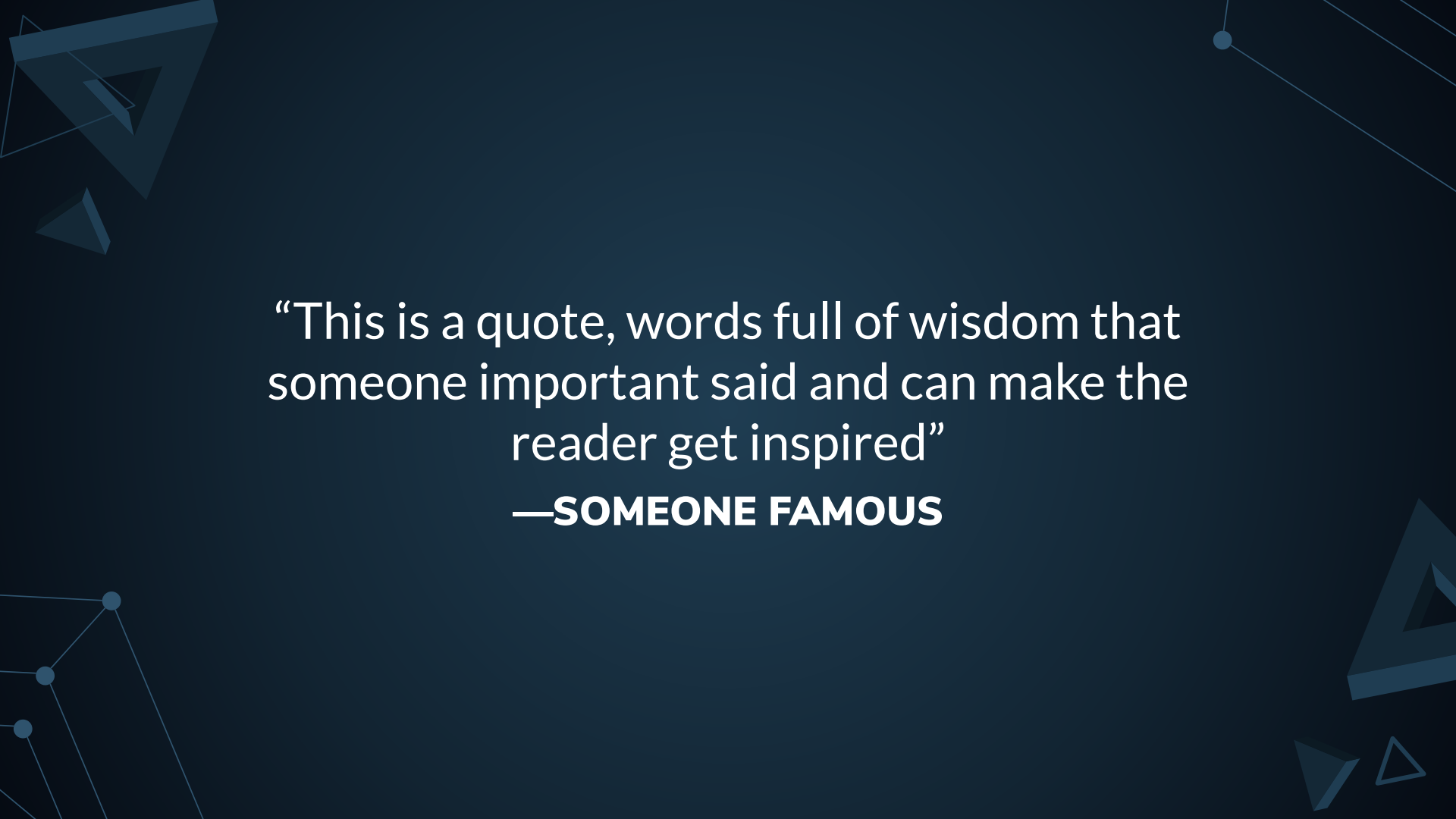
Jupiter is a very big planet

MERCURY

Mercury is the smallest planet

The background is a dark blue gradient. It features several abstract geometric elements: in the top left, a series of nested, slightly offset triangles; in the top right, a 3D pyramid-like structure; in the bottom left, a network of thin lines connecting small circular nodes; and in the bottom right, a series of concentric, slightly offset square frames.

AWESOME WORDS

The background is a dark blue gradient. It features several abstract geometric elements: a large, dark blue triangle in the top left corner, a network of thin white lines with small blue dots at the intersections in the bottom left, and a large, dark blue triangle in the bottom right corner. The text is centered in the middle of the image.

“This is a quote, words full of wisdom that
someone important said and can make the
reader get inspired”

—**SOMEONE FAMOUS**



A PICTURE ALWAYS REINFORCES THE CONCEPT

Images reveal large amounts of data, so remember: use an image instead of a long text

The background is a dark blue gradient. In the top-left corner, there is a network of thin white lines connecting several small blue dots. On the right side, there are several overlapping, semi-transparent geometric shapes, including a large triangle and a square, creating a sense of depth. In the bottom-left and bottom-right corners, there are more geometric shapes, including triangles and squares, some of which are also semi-transparent.

500,000

Big numbers catch your audience's attention

OUR SERVICES



MERCURY

Mercury is a small planet



MARS

Mars is red and cold



VENUS

Venus is the hottest planet



NEPTUNE

Neptune is far away from us



9h 55m 23s

Is Jupiter's rotation period

333,000.000

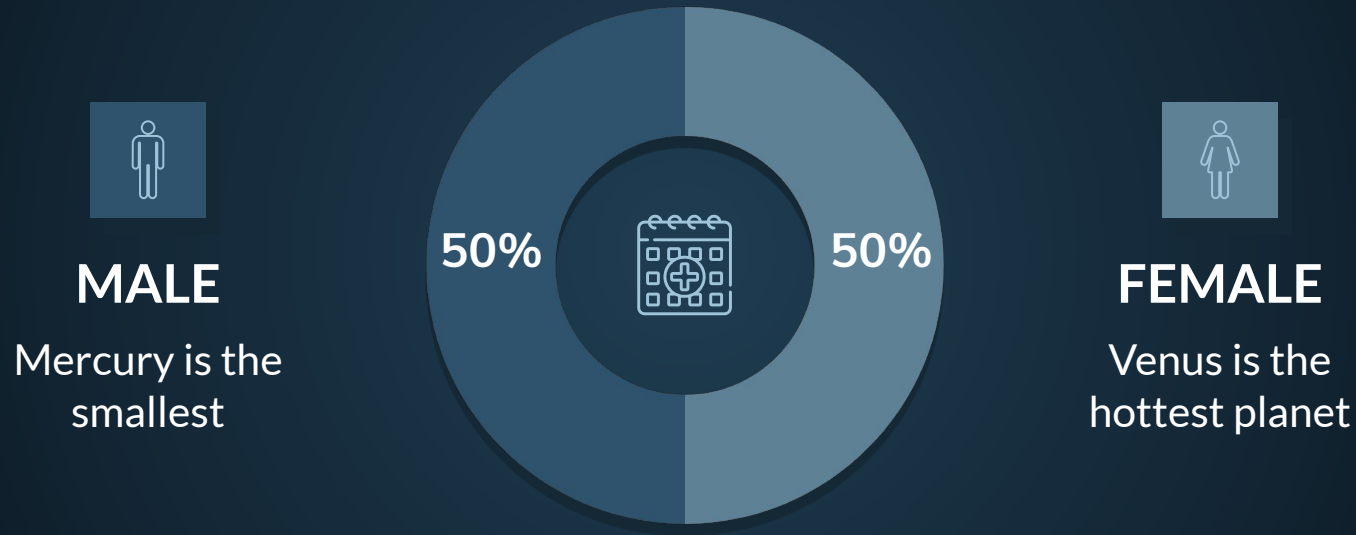
Earths is the Sun's mass

386,000 km

Is the distance between Earth and the Moon



OUR USERS: GENDER



Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

OUR CLIENTS



■ SATURN

Saturn is a gas giant
and has rings

■ JUPITER

Jupiter is the biggest
planet of them all

THE TIMELINE

1



MERCURY

Mercury is the smallest planet

2



SATURN

Saturn is the ringed planet

3



MARS

Mars is full of iron oxide dust

4



JUPITER

Jupiter is the biggest planet

THIS IS AN INFOGRAPHIC

JUPITER
Jupiter is the
biggest planet

1



3

MERCURY
Mercury is the
smallest planet

2

SATURN
Saturn is the
ringed planet

CALENDAR OF ACTIVITIES



SATURN

Saturn is ringed



JUPITER

Jupiter is big

MON	TUE	WED	THU	FRI	SAT	SUN
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

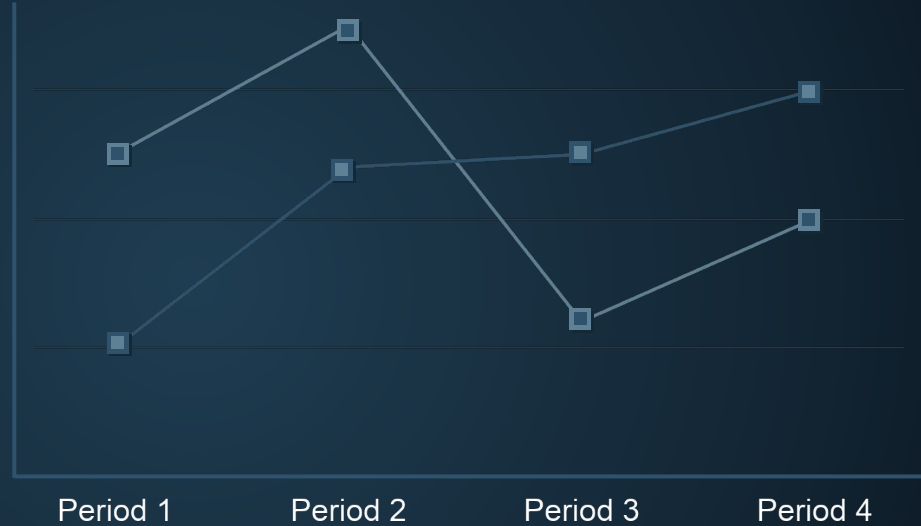
OUR BUDGET

- **JUPITER**

Jupiter is the biggest planet of them all

- **VENUS**

Venus has a pretty name, but it's hot



Follow the link in the graph to modify its data and then paste the new one here. **For more info, click here**

THE TITLE GOES HERE

Mercury is the closest planet to the Sun and the smallest one in the Solar System



JUPITER

Jupiter is the biggest planet



NEPTUNE

Neptune is the farthest planet



THE TITLE GOES HERE



JUPITER

Jupiter is the
biggest planet



MERCURY

Mercury is the
smallest planet



NEPTUNE

Neptune is the
farthest planet

OUR GOALS

MARS

Mars is a very cold planet



JUPITER

Jupiter is the biggest planet



MERCURY

Mercury is a small planet



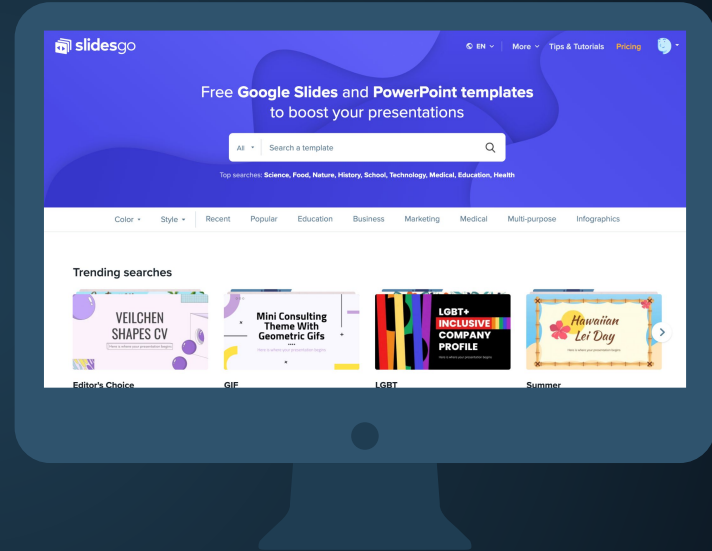
NEPTUNE

Neptune is very far away



DESKTOP SOFTWARE

You can replace the image on the screen with your own work. Right-click on it and then choose "Replace image" so you can add yours



A 4x10 grid of 40 white line-art icons on a dark blue background. The icons represent various concepts in web development, digital marketing, and technology. The first row includes icons for a desktop monitor, a web browser with 'WWW', a laptop with a flame, a document with a hand cursor, a laptop with a double arrow, a laptop with a padlock, a laptop with a play button and 'IMG', a network diagram with a globe, and a smartphone. The second row features crossed pencils, a person with a headset, a rocket, a monitor with a clock, crossed pencils and a ruler, a document with a person icon, a folder with a picture, a magnifying glass over a target, a bar chart with a play button, and a '404' error page. The third row shows a smartphone with a star, a keyboard with a wrench, a bug, a monitor with a gear, a cloud with a server, a sink with a faucet, a document with an envelope and pencil, a cloud with a gear, a monitor with a clock, and three interlocking gears. The fourth row contains a person's head, a skull with an exclamation mark, a 'CSS' monitor, a document with a person and play button, a magnifying glass over a document, a person with a headset, a globe with 'WWW', a smartphone with a plus sign, a document with a bar chart, and a document with a speech bubble. The fifth row includes a document with a gear, a browser window, a folder with a gear, two overlapping documents, a folder with a document, a browser window with a play button, a clipboard with a checkmark, a browser window with a play button, a smartphone with a play button, and a laptop with 'HTML'.

Fonts & colors used

This presentation has been made using the following fonts:

Nunito Sans

(<https://fonts.google.com/specimen/Nunito+Sans>)

Lato

(<https://fonts.google.com/specimen/Lato>)

#050a12

#ffffff

#1f3d52

#2f536d

#142836

#0c1a23

#142836

#5f8195

#08151e

#5f8195

Use our editable graphic resources...

You can easily [resize](#) these resources without losing quality. To [change the color](#), just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want.

Group the resource again when you're done. You can also look for more [infographics](#) on [Slidesgo](#).



...and our sets of editable icons

You can resize these icons without losing quality.

You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.

In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.



Educational Icons



Medical Icons



Business Icons



Teamwork Icons



Help & Support Icons



Avatar Icons



Creative Process Icons



Performing Arts Icons



SEO & Marketing Icons

