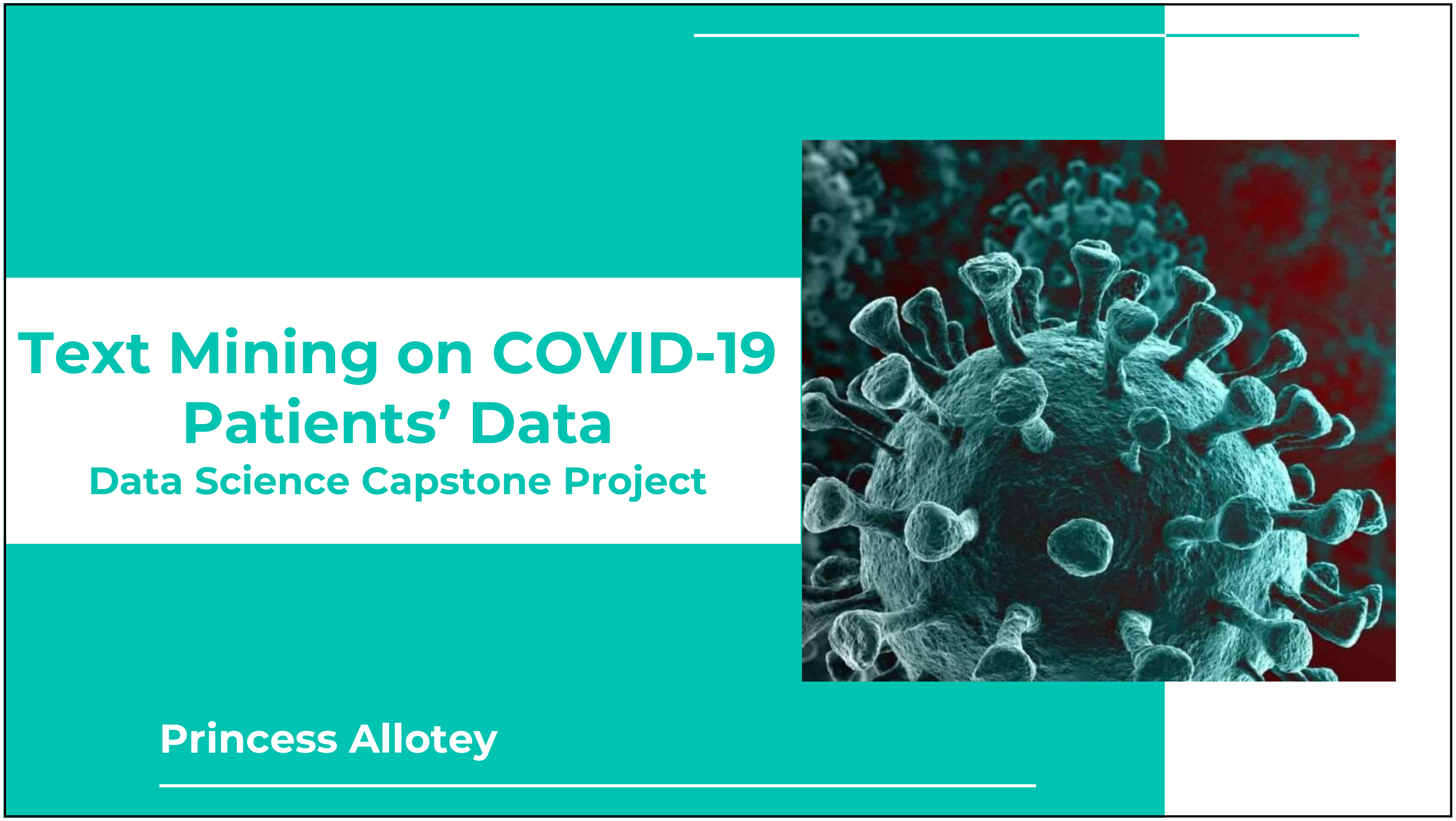
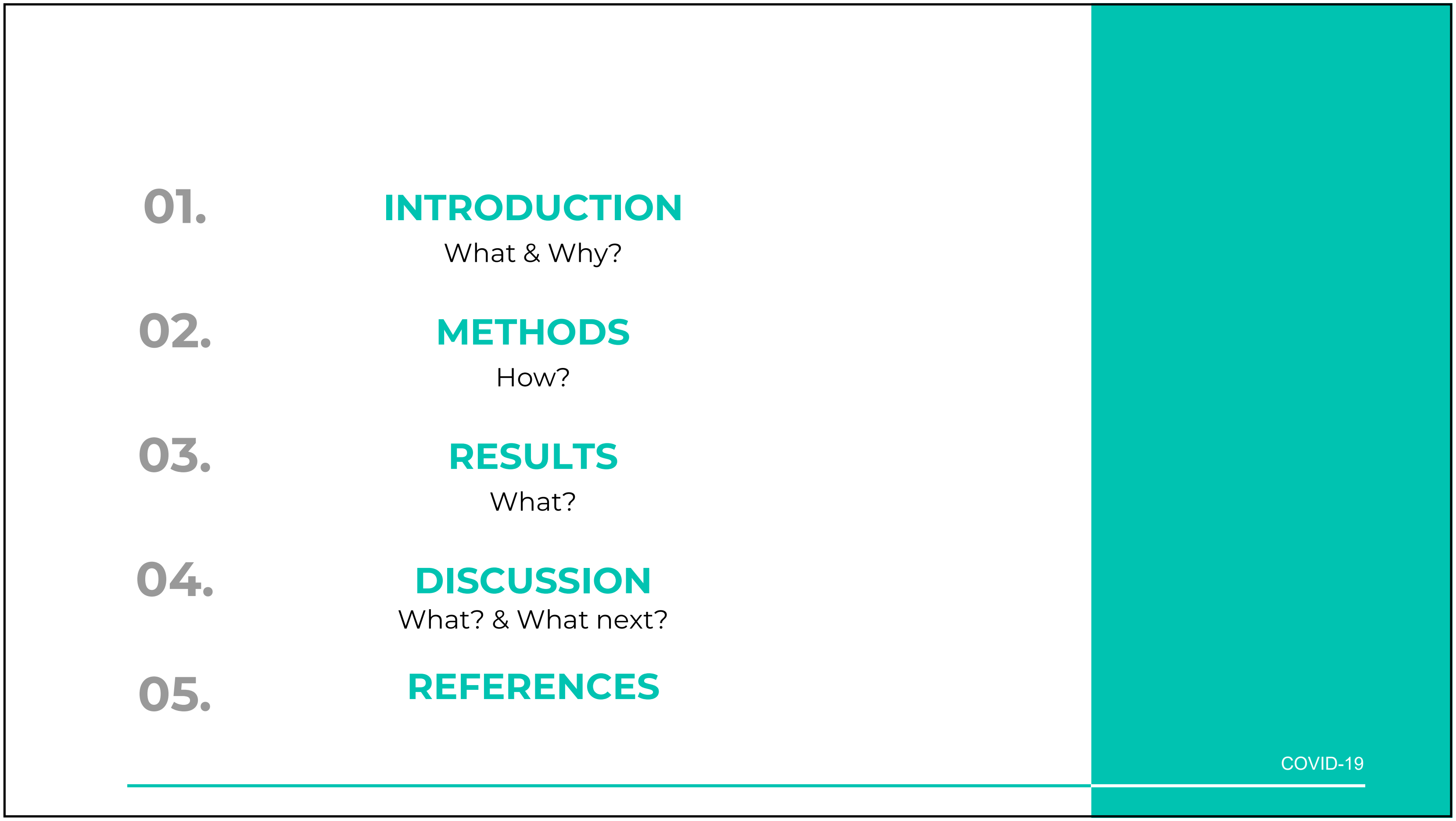


Text Mining on COVID-19 Patients’ Data

Data Science Capstone Project

Princess Allotey



01. **INTRODUCTION**
What & Why?

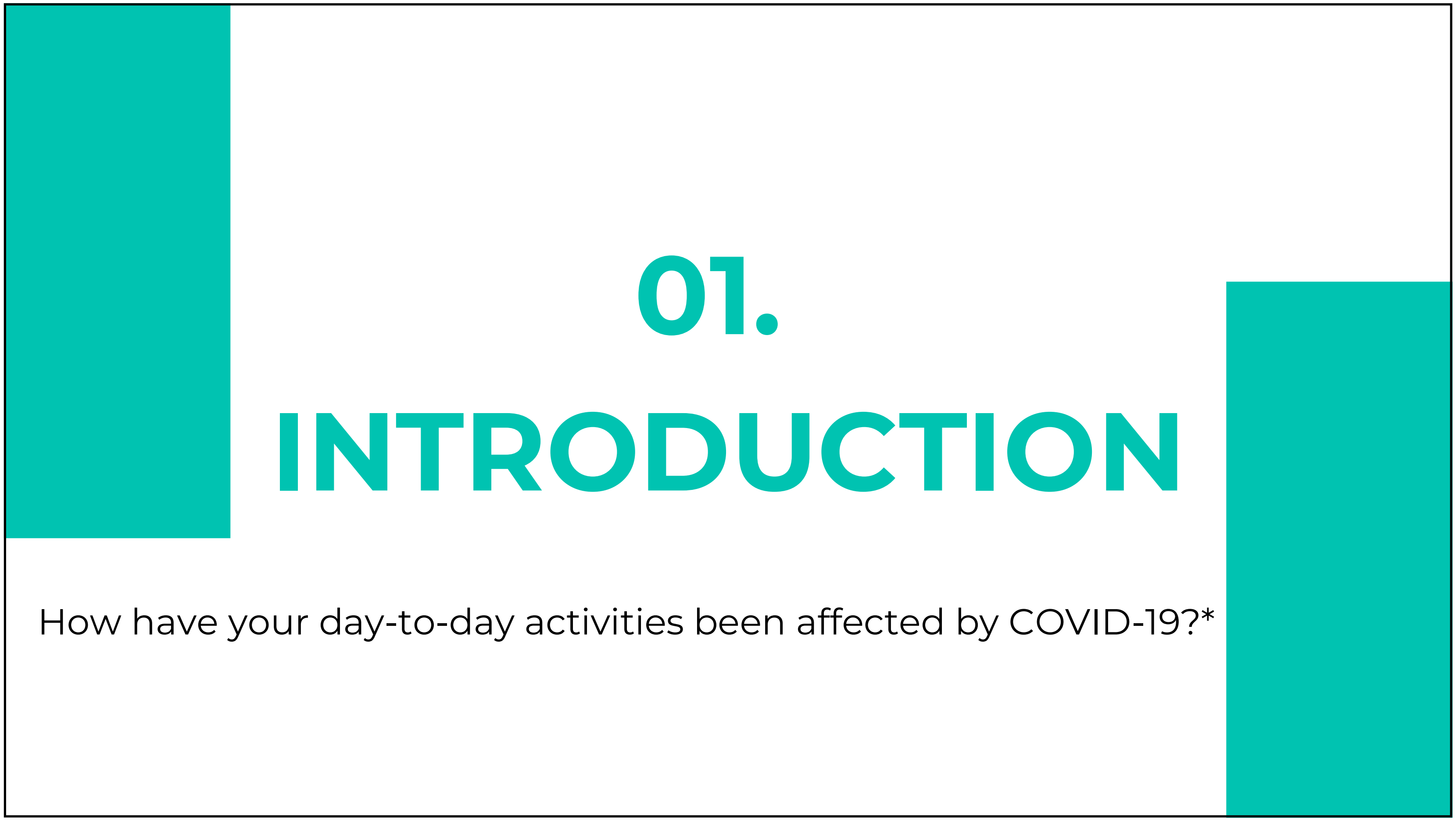
02. **METHODS**
How?

03. **RESULTS**
What?

04. **DISCUSSION**
What? & What next?

05. **REFERENCES**

COVID-19



01.
INTRODUCTION

How have your day-to-day activities been affected by COVID-19?*

The current state of the world

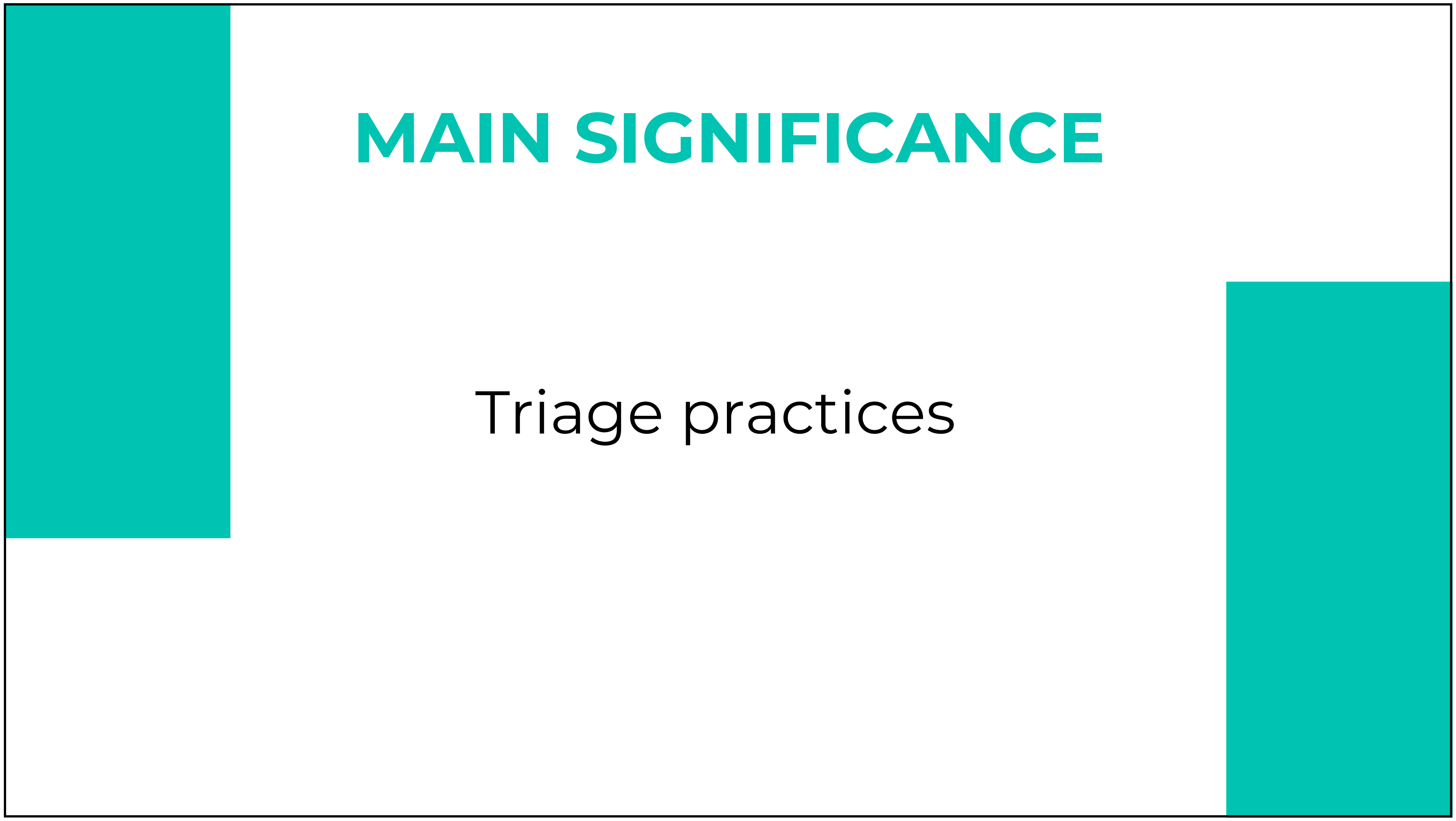
TIMELINE

- **December 2019** – discovered in Wuhan, China
- **March 11th, 2020** – declared a pandemic
- **September 14th, 2020** – 29 million confirmed cases worldwide, and close to 1 million deaths

PROJECT GOAL & VISION OF THE FUTURE

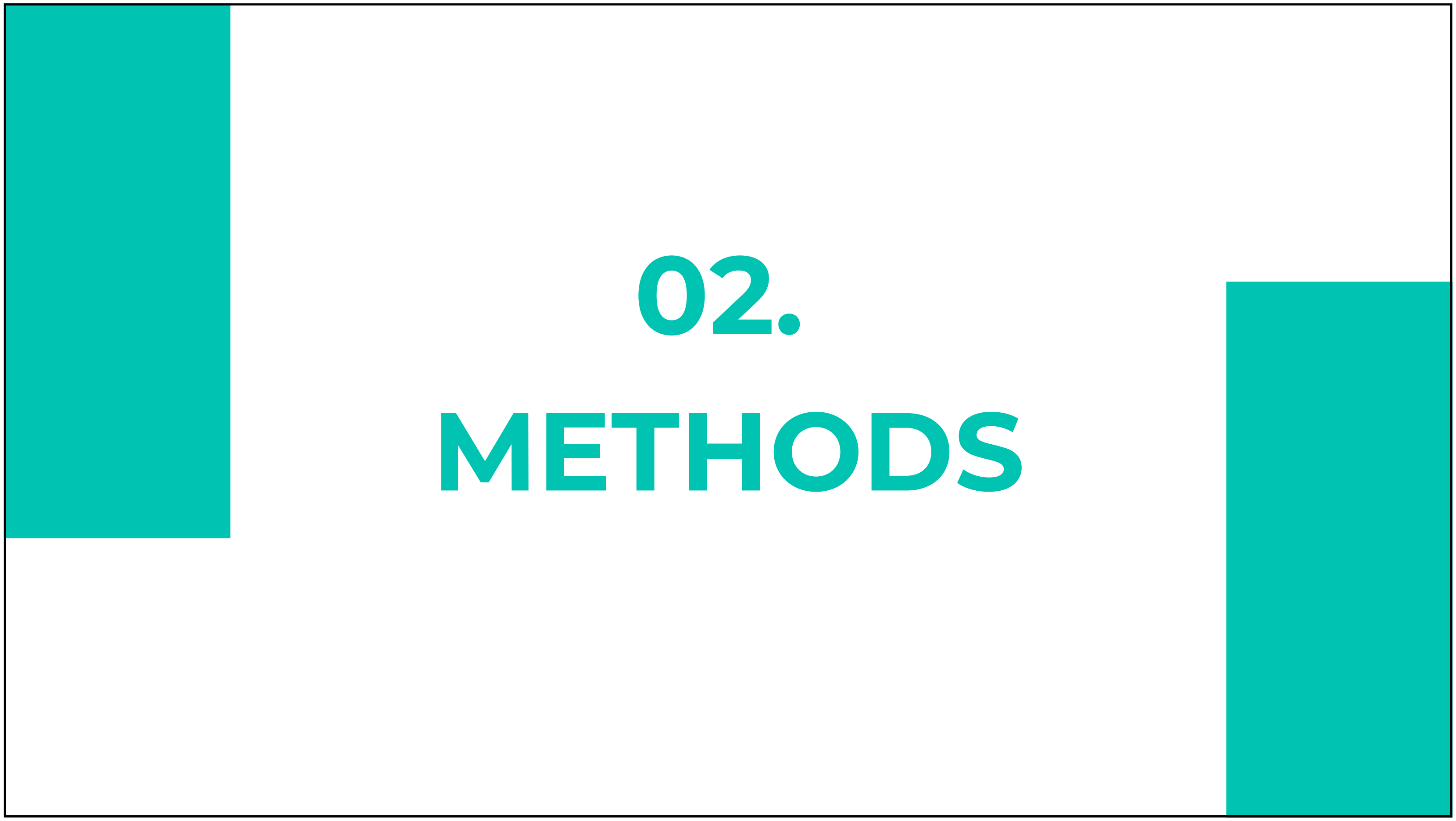
Predicting the outcome of future patients from electronic health records

COVID-19



MAIN SIGNIFICANCE

Triage practices



02.

METHODS

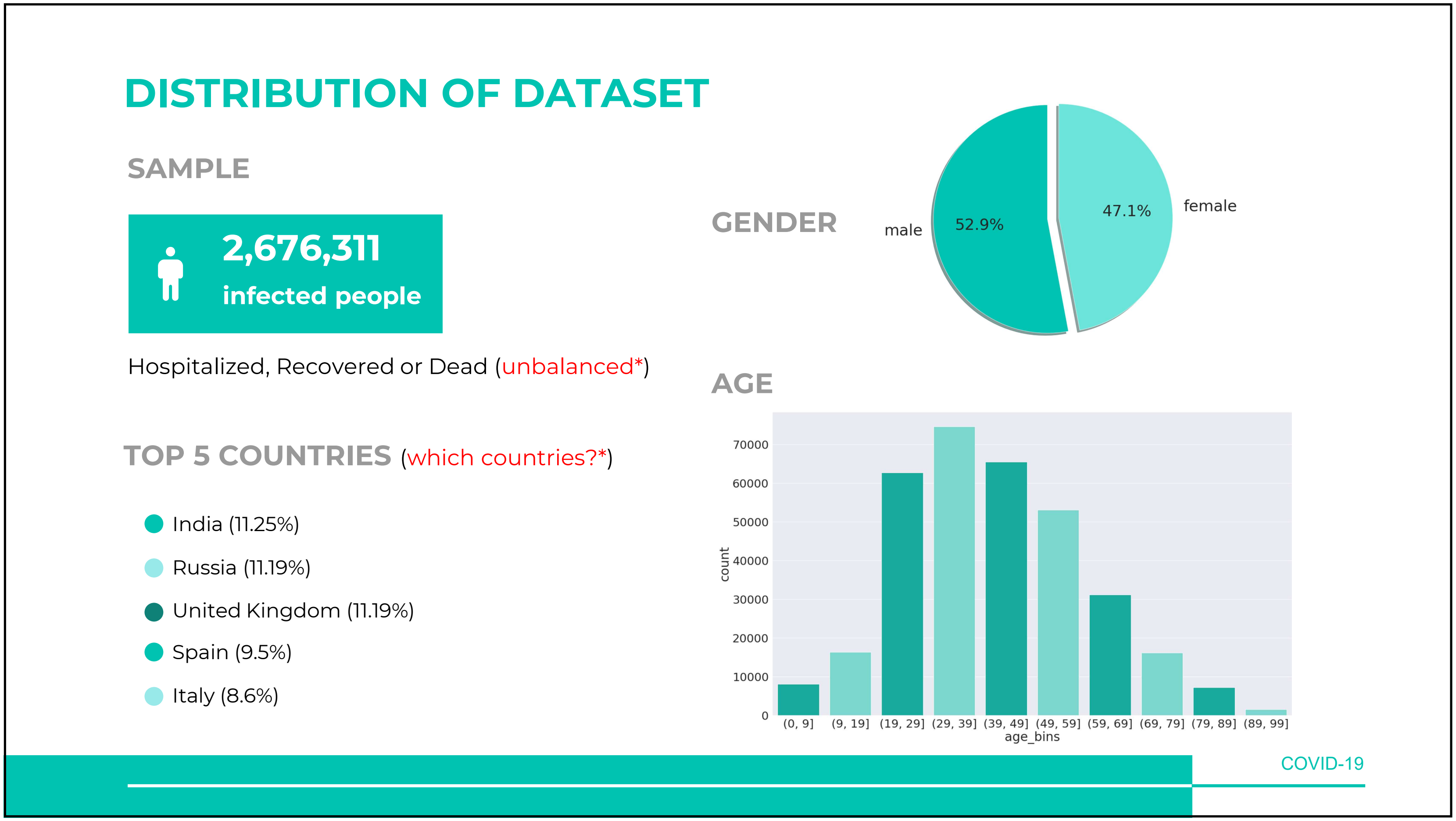
8

DATA SOURCE

Institute for
Health Metrics
and Evaluation

University of Washington

COVID-19



DISTRIBUTION OF DATASET

SAMPLE

2,676,311

infected people

Hospitalized, Recovered or Dead (unbalanced*)

TOP 5 COUNTRIES (which countries?*)

- ● India (11.25%)
- ● Russia (11.19%)
- ● United Kingdom (11.19%)
- ● Spain (9.5%)
- ● Italy (8.6%)

GENDER

AGE

COVID-19

FOR EACH COVID-19 PATIENT:

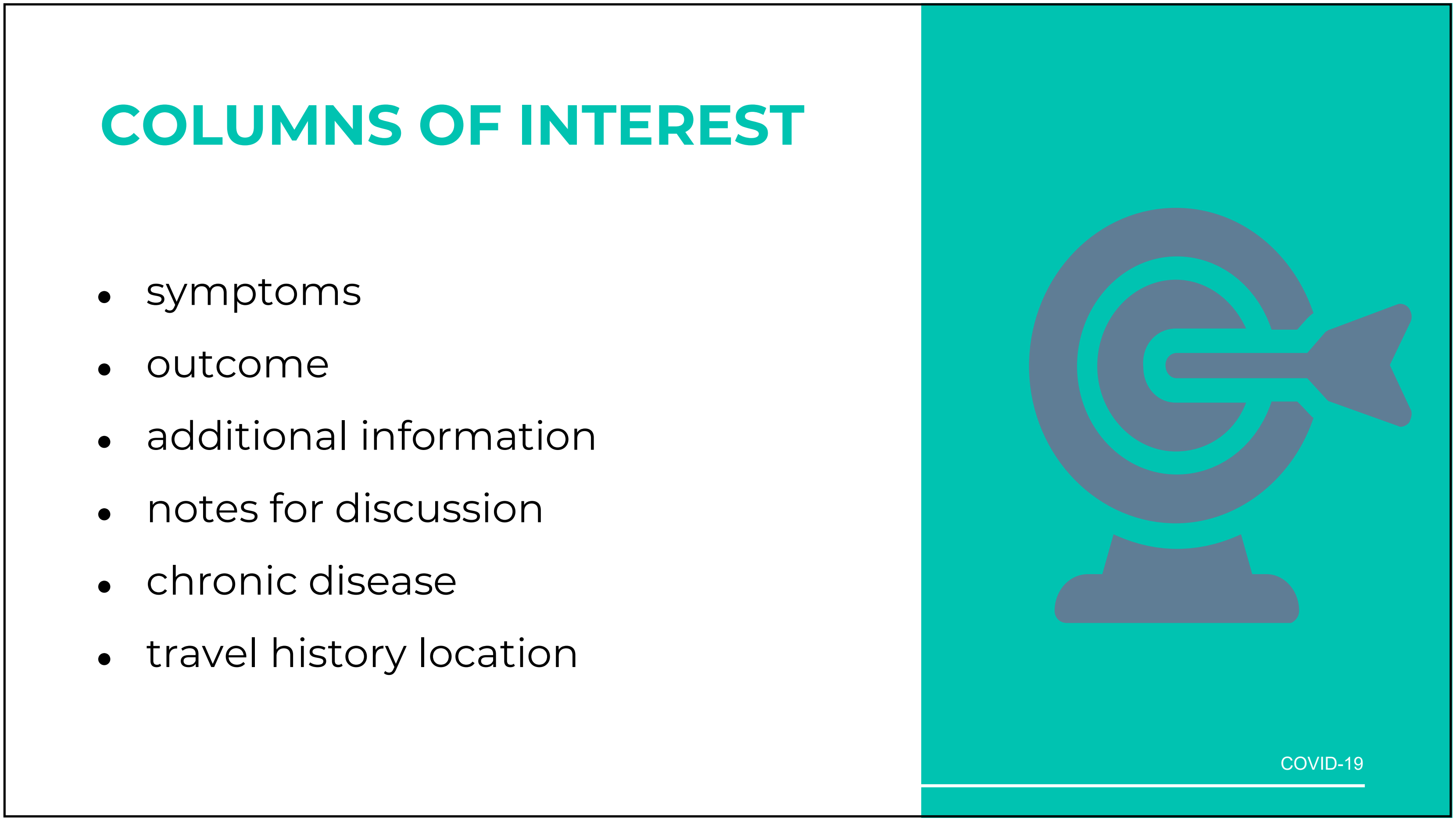
AGE	65
SEX	Male
COUNTRY	Vo Euganeo, Italy
CHRONIC DISEASE	Hypertension

TRAVEL HISTORY DATES

DATE OF ONSET OF SYMPTOMS

DATE OF HOSPITAL ADMISSION

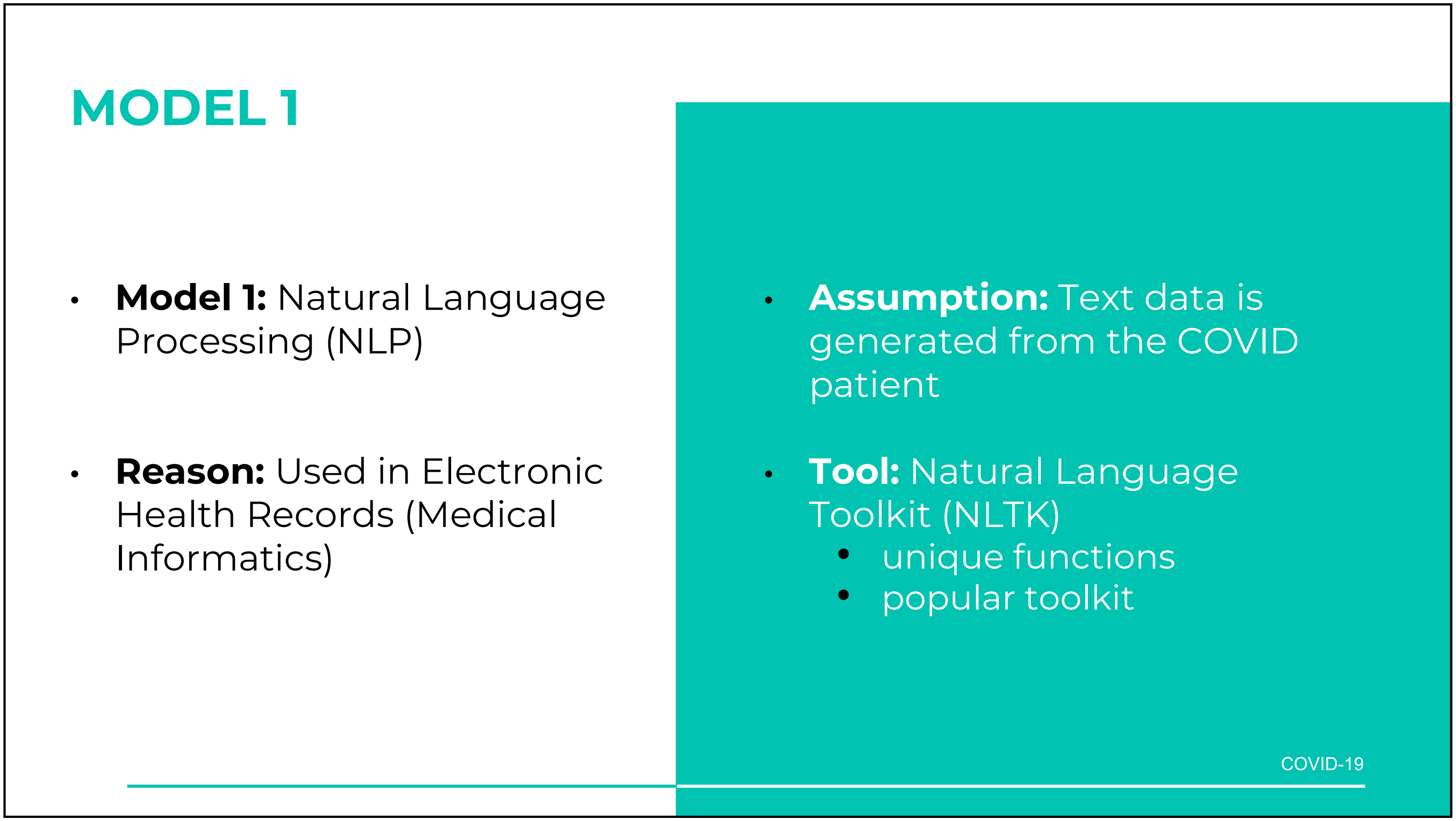
DATE OF DEATH OR DISCHARGE



COLUMNS OF INTEREST

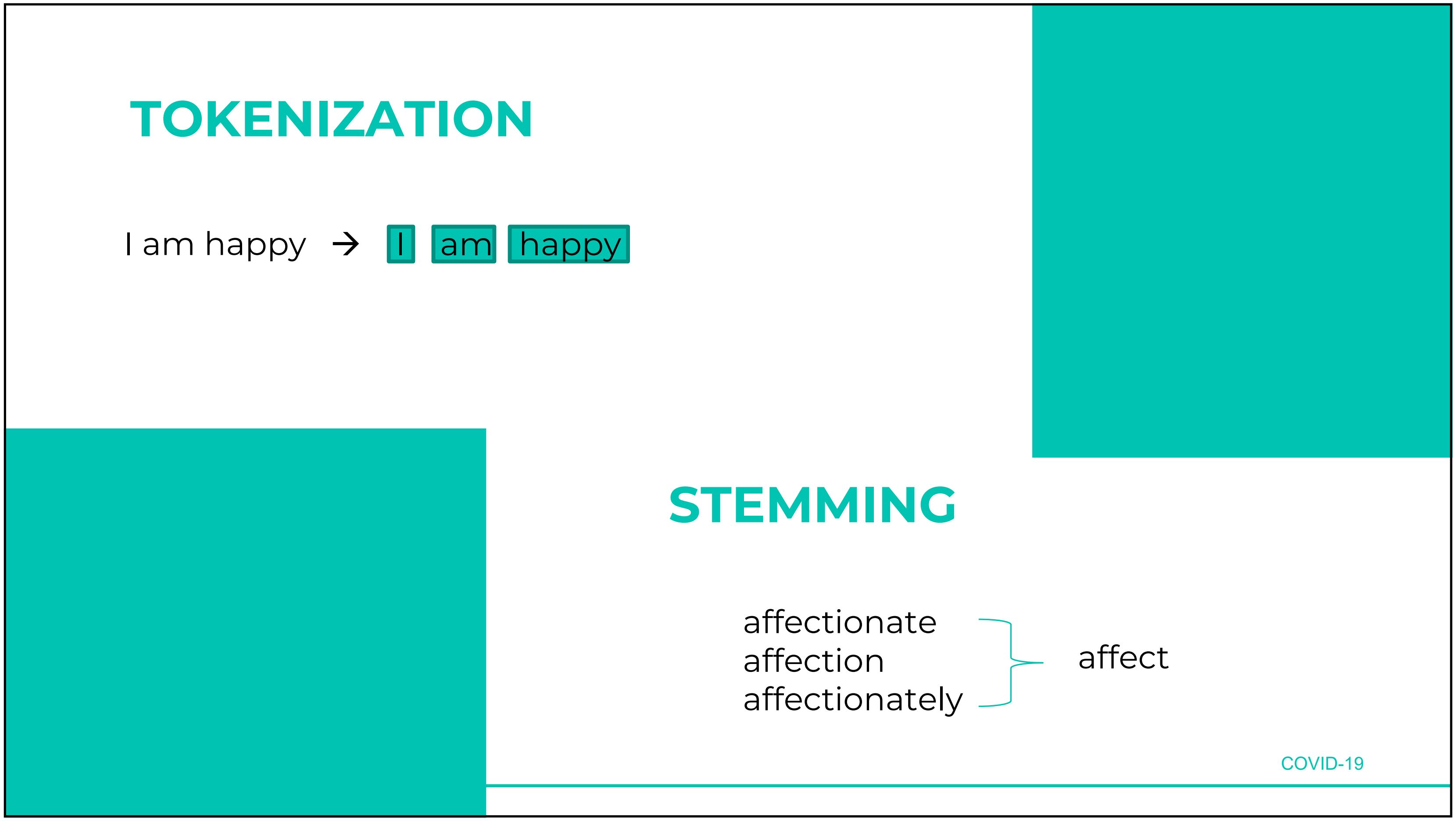
- symptoms
- outcome
- additional information
- notes for discussion
- chronic disease
- travel history location

COVID-19



MODEL 1

- **Model 1:** Natural Language Processing (NLP)
- **Reason:** Used in Electronic Health Records (Medical Informatics)
- **Assumption:** Text data is generated from the COVID patient
- **Tool:** Natural Language Toolkit (NLTK)
  - unique functions
  - popular toolkit



TOKENIZATION

I am happy → I am happy

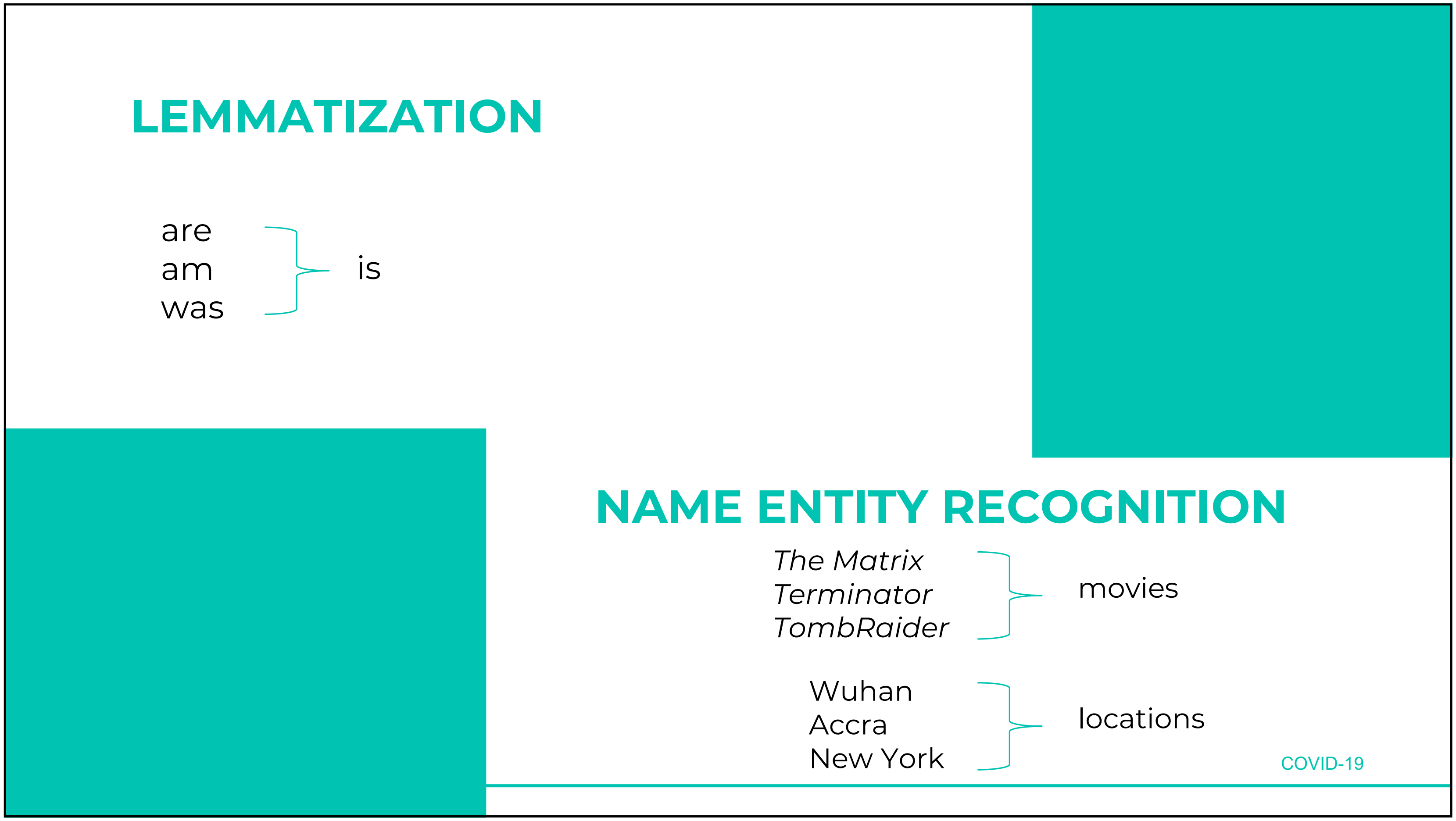
STEMMING

affectionate
affection
affectionately

affect

COVID-19

14



LEMMATIZATION

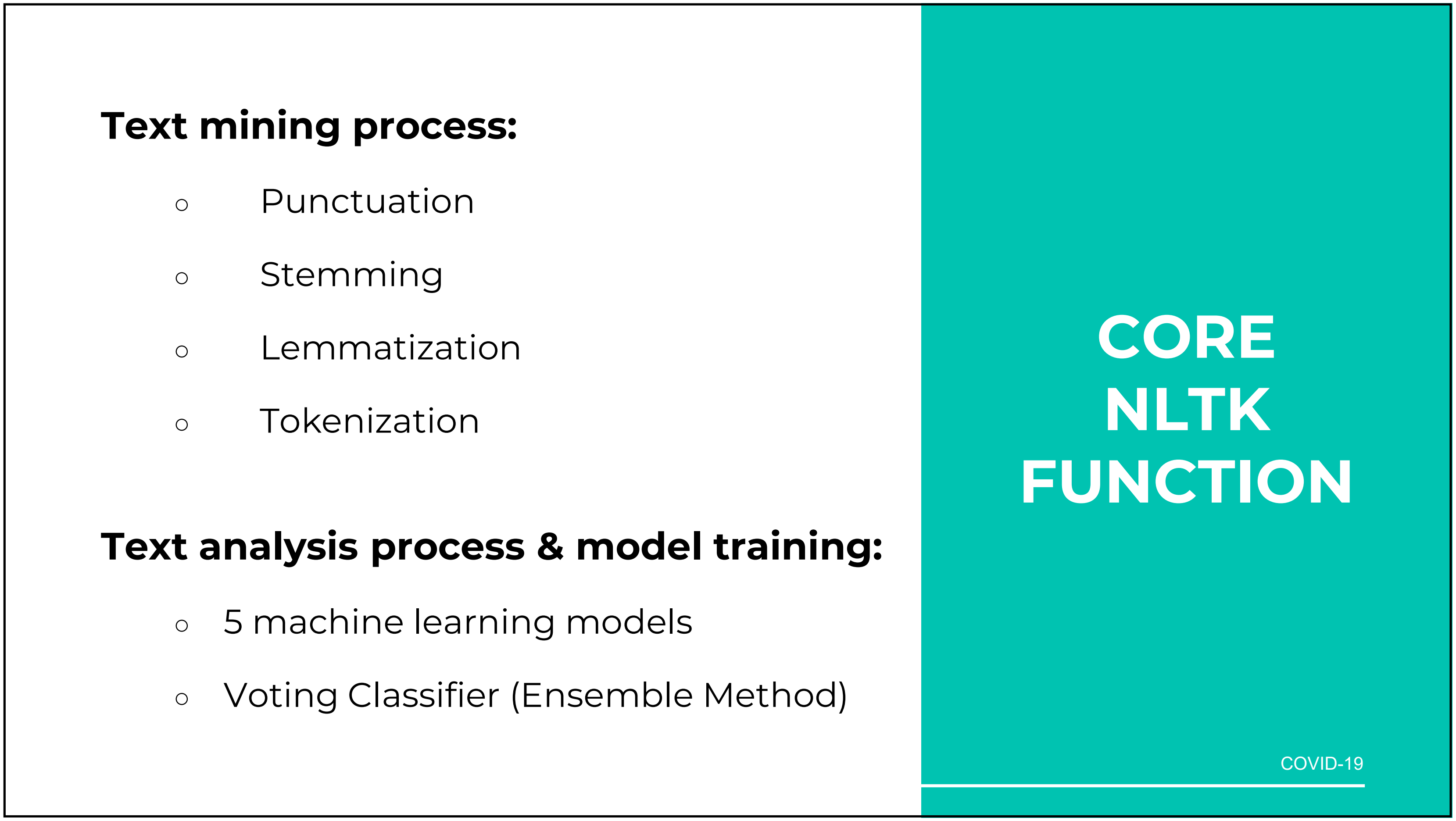
are
am
was
is

NAME ENTITY RECOGNITION

The Matrix
Terminator
TombRaider
movies

Wuhan
Accra
New York
locations

COVID-19



CORE NLTK FUNCTION

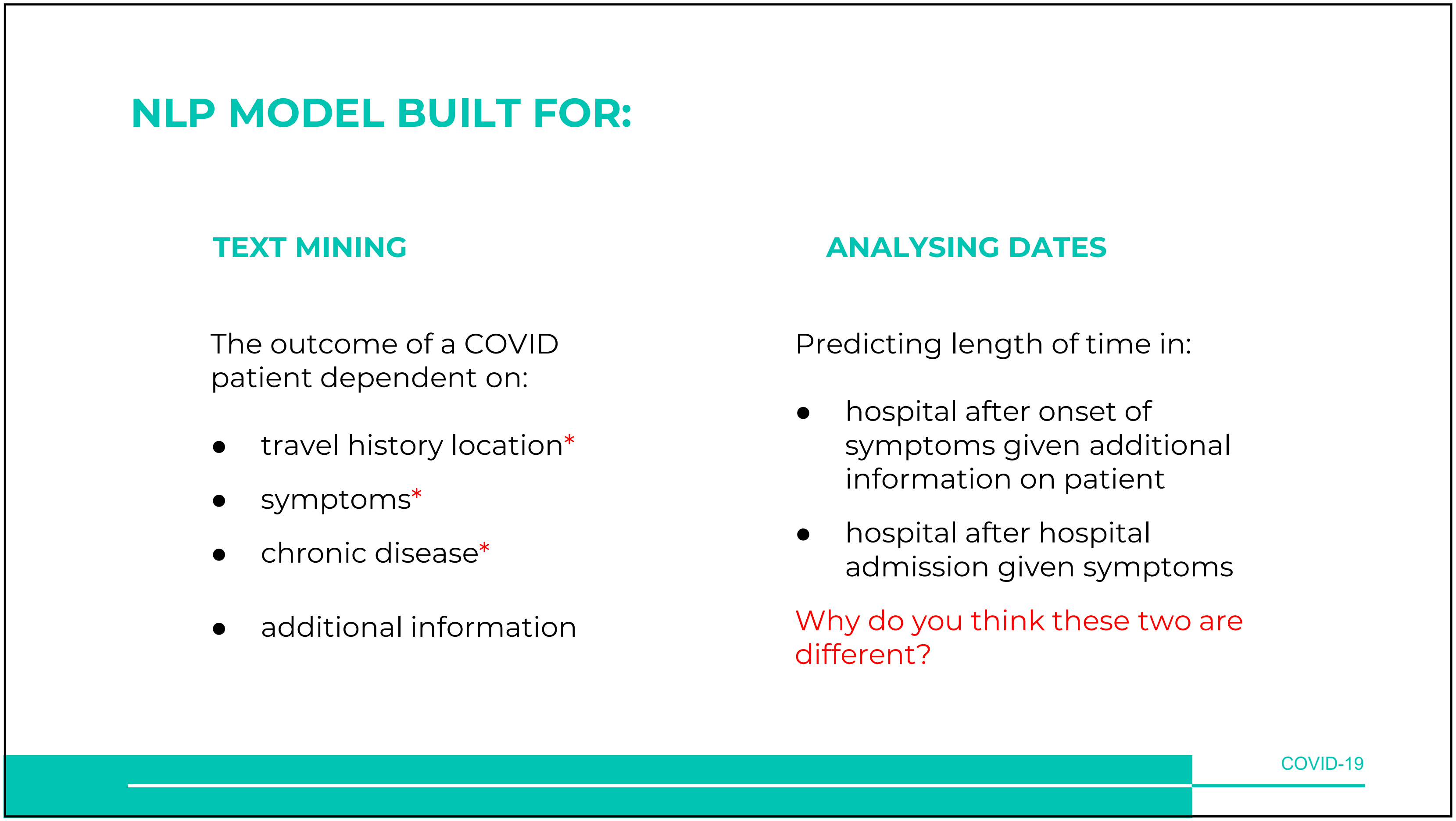
Text mining process:

- Punctuation
- Stemming
- Lemmatization
- Tokenization

Text analysis process & model training:

- 5 machine learning models
- Voting Classifier (Ensemble Method)

COVID-19



NLP MODEL BUILT FOR:

TEXT MINING

The outcome of a COVID patient dependent on:

- travel history location*
- symptoms*
- chronic disease*
- additional information

ANALYSING DATES

Predicting length of time in:

- hospital after onset of symptoms given additional information on patient
- hospital after hospital admission given symptoms

Why do you think these two are different?

MODEL 2

- **Model 2:** Logistic Regression
- **Reason:** Integrates previous explorations
- Predicts the likelihood that a patient will recover

COVID-19

TARGET AND FEATURES

Y:

- Outcome

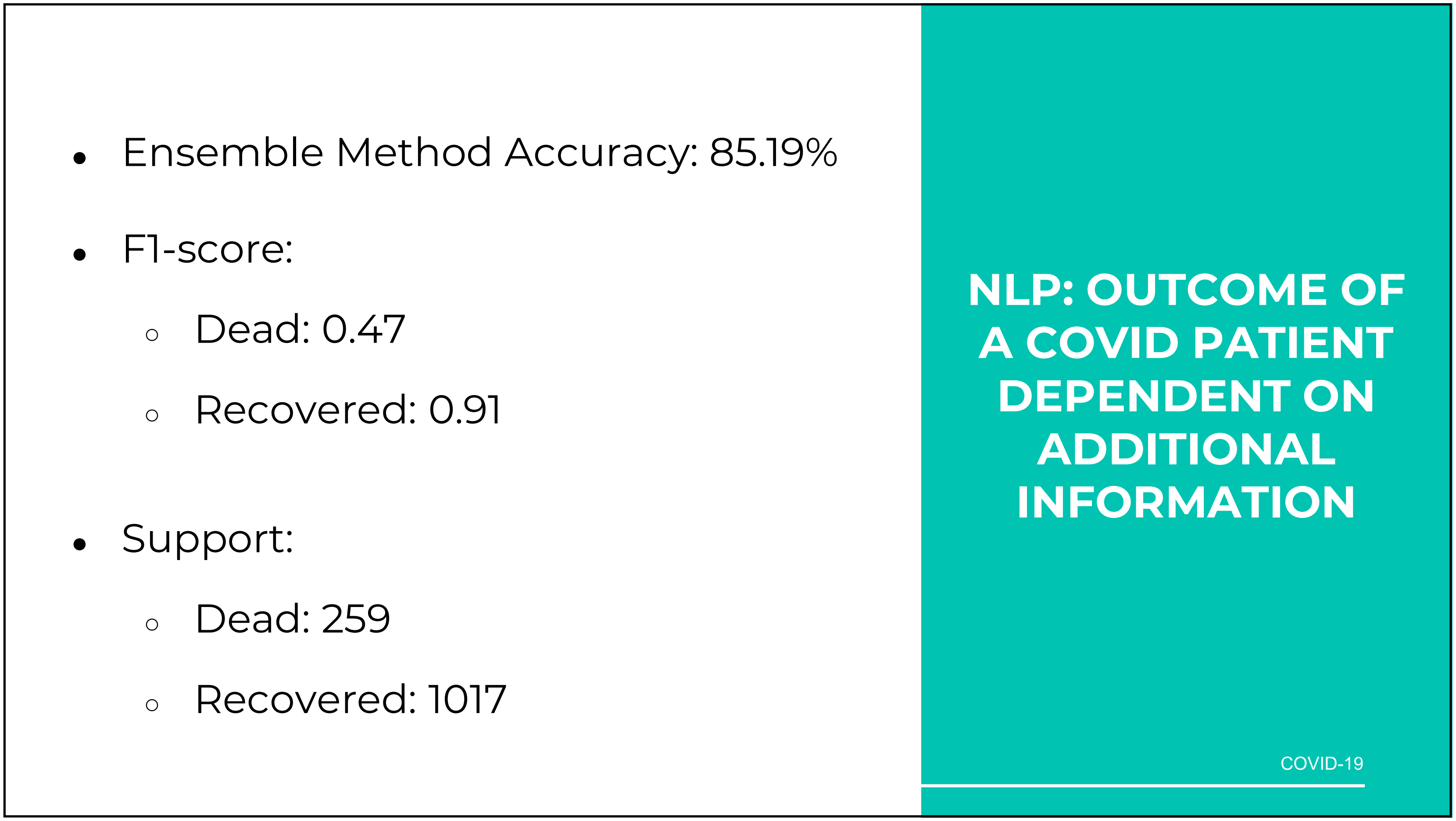
X:

- Age
- Sex
- Chronic disease binary
- 2 date differences
- Symptoms

COVID-19

03.

RESULTS



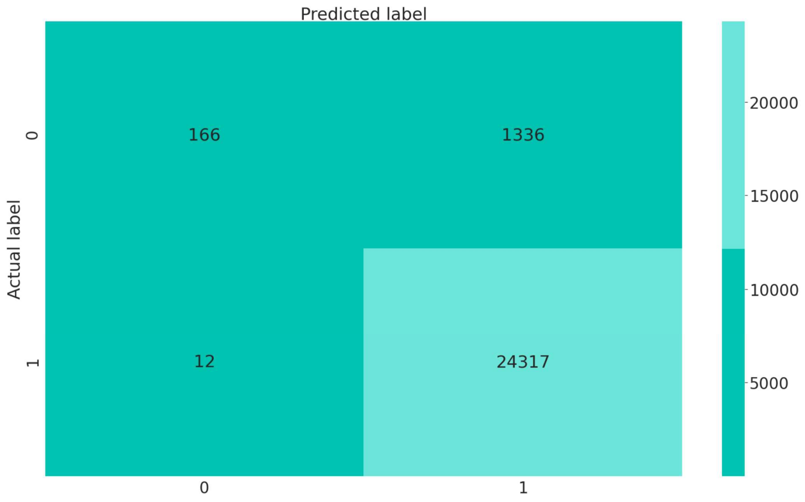
NLP: OUTCOME OF A COVID PATIENT DEPENDENT ON ADDITIONAL INFORMATION

- Ensemble Method Accuracy: 85.19%
- F1-score:
  - Dead: 0.47
  - Recovered: 0.91
- Support:
  - Dead: 259
  - Recovered: 1017

COVID-19

LOGISTIC REGRESSION

PREDICTING RECOVERY



1 = Recovered, 0 = Dead

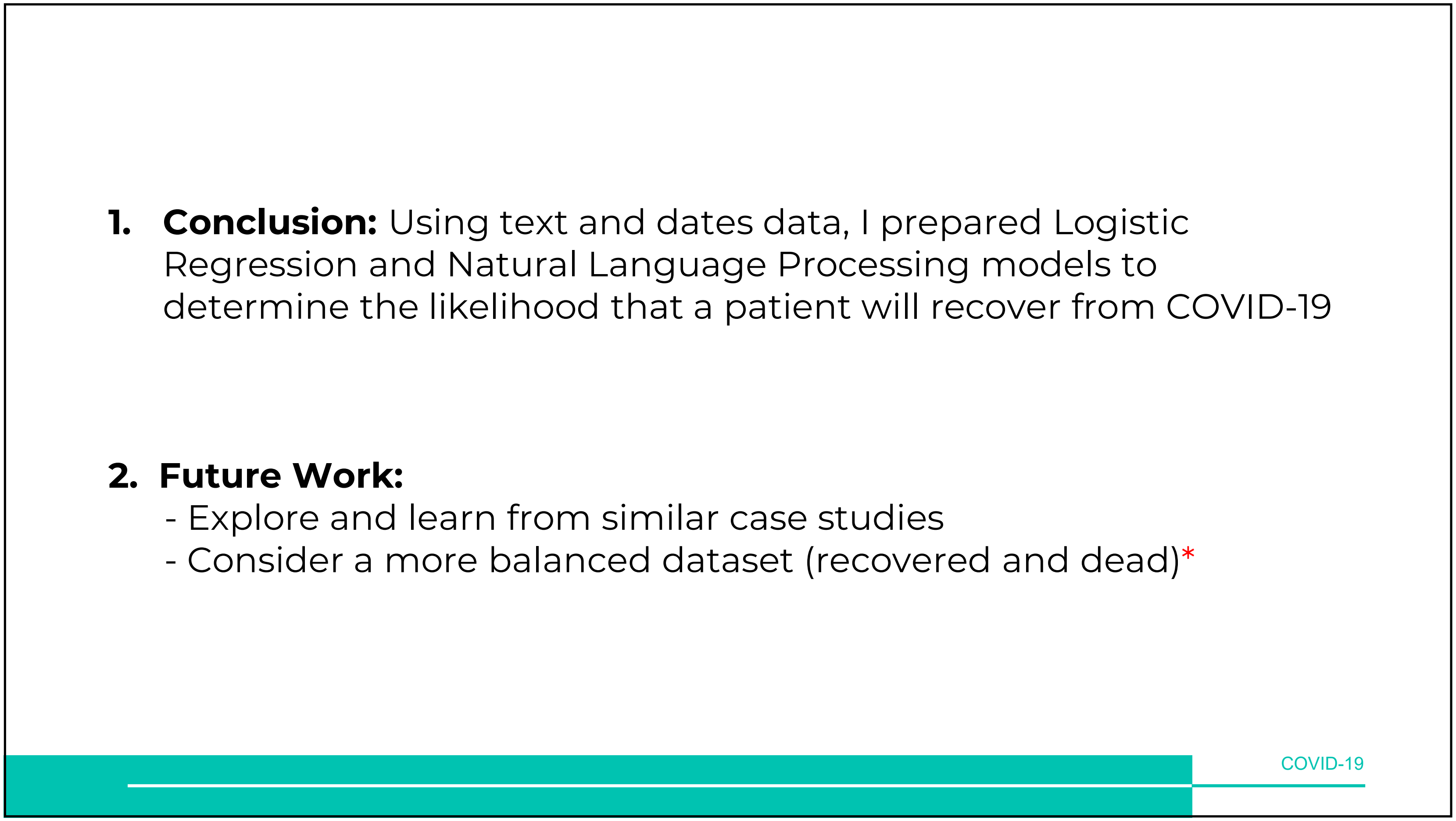
- Accuracy score: 0.95
- F1-score:
  - Dead: 0.20
  - Recovered: 0.97
- Support:
  - Dead: 1502
  - Recovered: 24329

04.

DISCUSSION

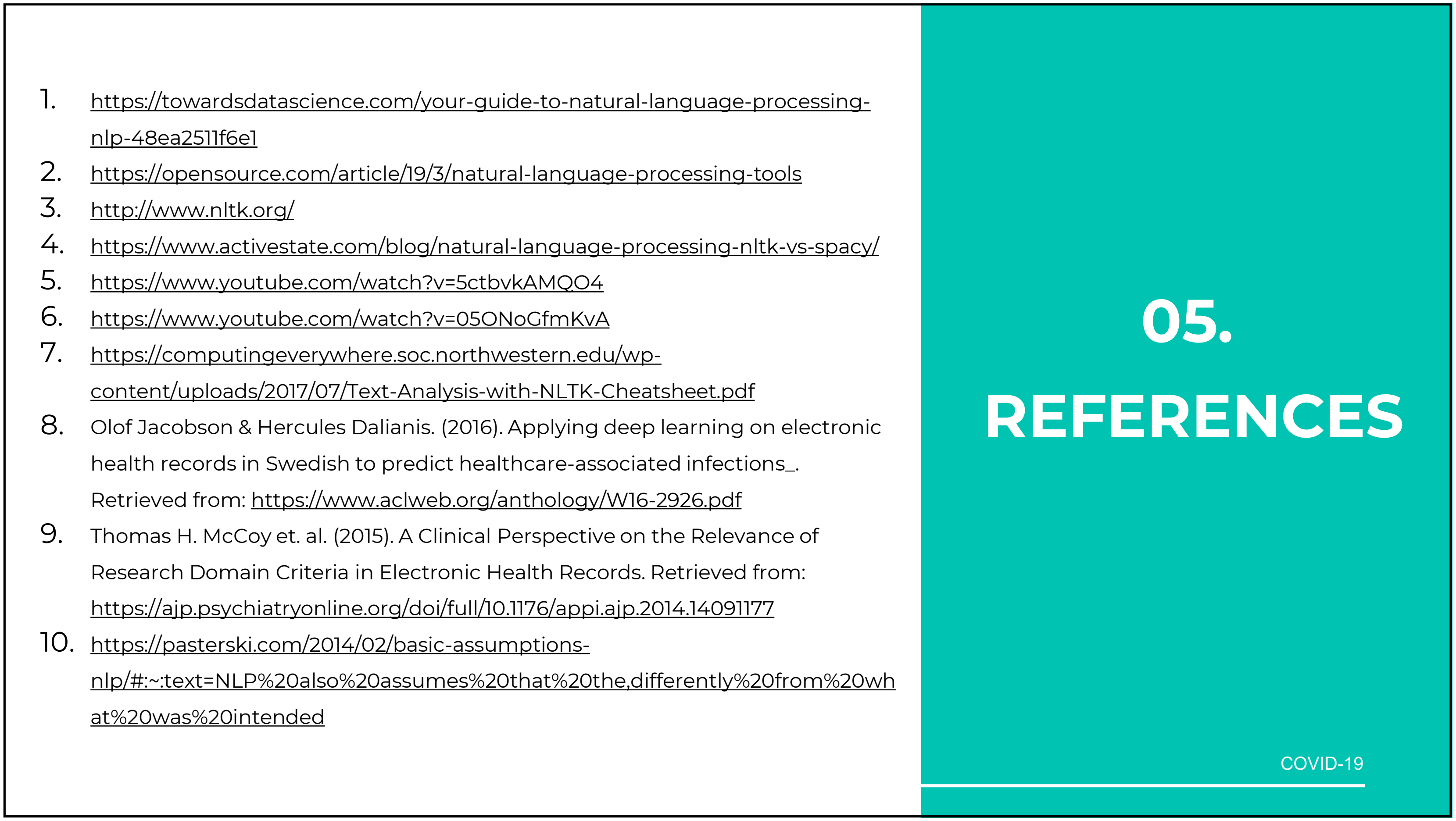
What did I find?

What are my next steps?



1. **Conclusion:** Using text and dates data, I prepared Logistic Regression and Natural Language Processing models to determine the likelihood that a patient will recover from COVID-19

2. **Future Work:**
   - Explore and learn from similar case studies
   - Consider a more balanced dataset (recovered and dead)*



1. https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1
2. https://opensource.com/article/19/3/natural-language-processing-tools
3. http://www.nltk.org/
4. https://www.activestate.com/blog/natural-language-processing-nltk-vs-spacy/
5. https://www.youtube.com/watch?v=5ctbvkAMQO4
6. https://www.youtube.com/watch?v=05ONoGfmKvA
7. https://computingeverywhere.soc.northwestern.edu/wp-content/uploads/2017/07/Text-Analysis-with-NLTK-Cheatsheet.pdf
8. Olof Jacobson & Hercules Dalianis. (2016). Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections_. Retrieved from: https://www.aclweb.org/anthology/W16-2926.pdf
9. Thomas H. McCoy et. al. (2015). A Clinical Perspective on the Relevance of Research Domain Criteria in Electronic Health Records. Retrieved from: https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2014.14091177
10. https://pasterski.com/2014/02/basic-assumptions-nlp/#:~:text=NLP%20also%20assumes%20that%20the,differently%20from%20what%20was%20intended

05.

REFERENCES

COVID-19

slidesgo

26

Let’s tackle COVID-19 quickly and efficiently!

Do you have any questions?
princess.allotey@centre.edu
+1 859 319 0168

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.