Leveraging R and ChatGPT for Epidemiological Analysis

Prof Dr Jamalludin Ab Rahman MD MPH FPHMM
Dean, Kulliyyah of Medicine, IIUM
President, PPPKAM

Online WORKSHOP

LEVERAGING R AND CHATGPT FOR EPIDEMIOLOGICAL ANALYSIS

AUG 10TH

8:30 AM

4 CPD Point MMA & NSR

JOIN US

https://forms.gle/BTd7DYh8ktiz9yve9

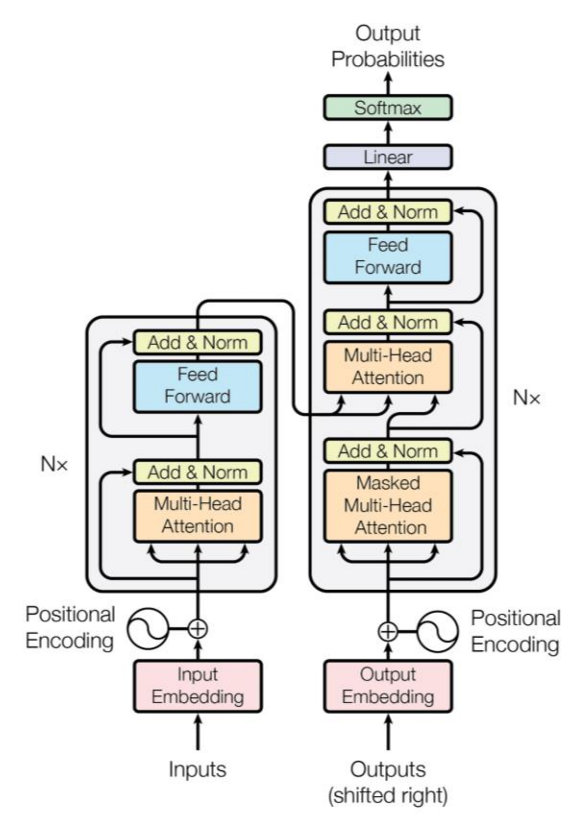
Fees

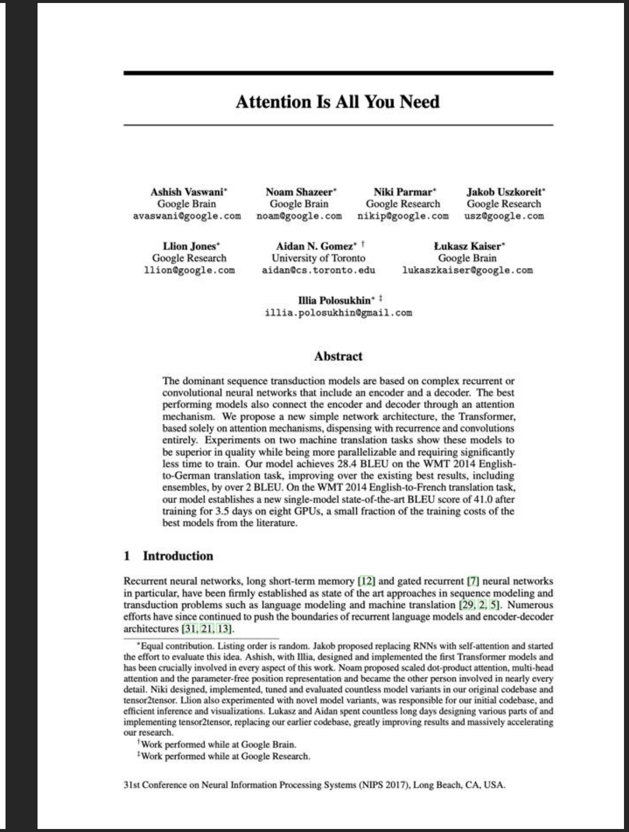
PPPKAM/PPHPA MEMBERS: RM10
NON-MEMBERS: RM30

Prof Dr Jamalludin Ab Rahman
Epidemiologist & Biostatistician
Dean, Kulliyyah of Medicine, IIUM
President, PPPKAM

Further information
✉ ppkapahang@gmail.com

What is ChatGPT





Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* [†] University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* [‡] illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†] Work performed while at Google Brain.

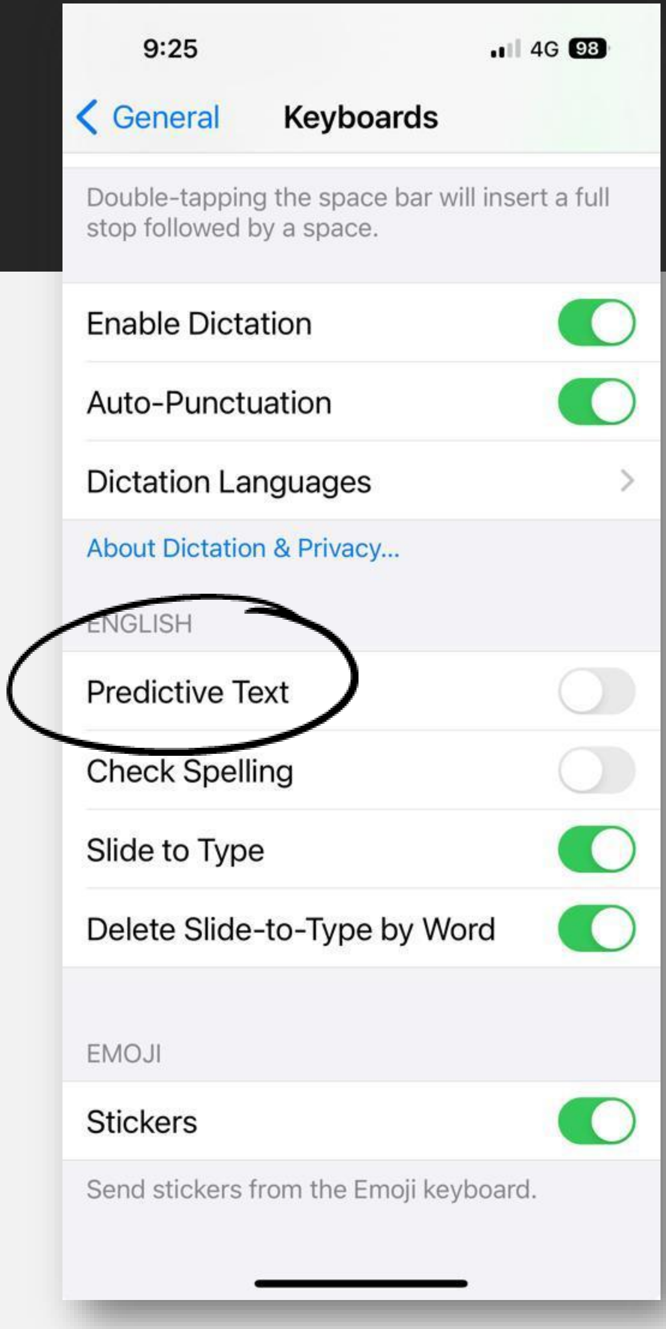
[‡] Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

- Developed by OpenAI
- Built on the GPT (Generative Pre-trained Transformer) – Vaswani et al. 2017
- Transformer – a new neural network architecture
- Attention is a mechanism that allows neural networks to focus on specific parts of their input.

All start with prediction

- *I'm going to the…..*
- Use the Language Model to predict the next word
- It is already in the smartphone – predictive text
- The language model works on a certain computational framework (neural network, many types, one of them is the Transformer model)



Double-tapping the space bar will insert a full stop followed by a space.

Enable Dictation ●

Auto-Punctuation ●

Dictation Languages >

About Dictation & Privacy...

ENGLISH

Predictive Text ○

Check Spelling ○

Slide to Type ●

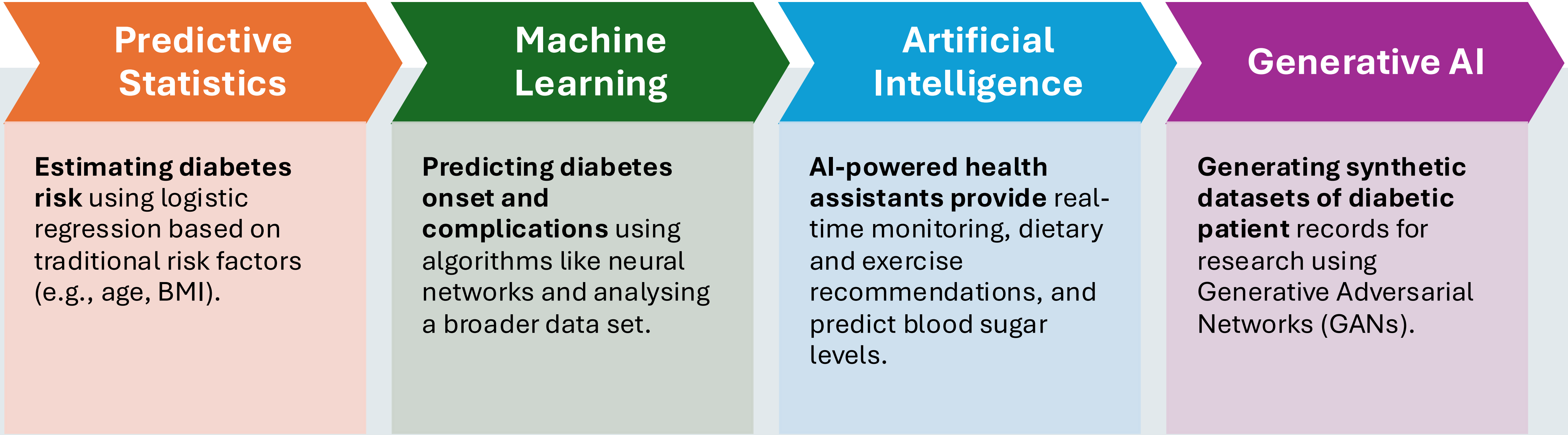
Delete Slide-to-Type by Word ●

EMOJI

Stickers ●

Send stickers from the Emoji keyboard.

Example of the ‘evolution’ in diabetes epidemiology



Predictive Statistics

Estimating diabetes risk using logistic regression based on traditional risk factors (e.g., age, BMI).

Machine Learning

Predicting diabetes onset and complications using algorithms like neural networks and analysing a broader data set.

Artificial Intelligence

AI-powered health assistants provide real-time monitoring, dietary and exercise recommendations, and predict blood sugar levels.

Generative AI

Generating synthetic datasets of diabetic patient records for research using Generative Adversarial Networks (GANs).

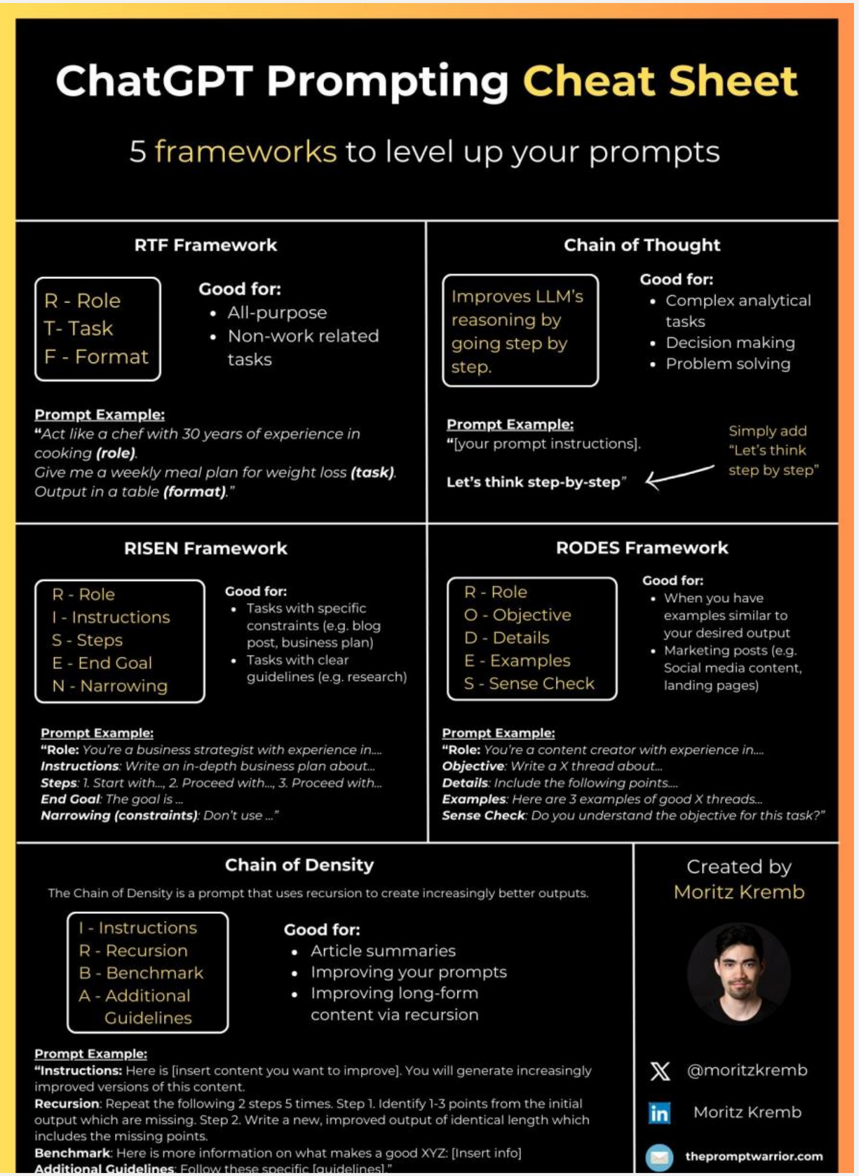
How ChatGPT works in statistics

- Trained from all information available (easier & faster literature search)
- Propose the best* solution, best* method
- Describe concepts faster and easier
- Propose coding/algorithm

* Depending on the trained data

Prompting tips

1. Persona/role – a lecturer, a professor, a student, a PhD candidate
2. The task/instruction – to improve, to describe, to analyse, to compare etc
3. Expectation/end goals – simple, complex, layman's terms
4. Filter – narrow the output
5. Format the output – table, diagram



ChatGPT Prompting Cheat Sheet

5 frameworks to level up your prompts

RTF Framework

R - Role
T- Task
F - Format

Good for:
- All-purpose
- Non-work related tasks

Prompt Example:
“*Act like a chef with 30 years of experience in cooking* ***(role)***.
Give me a weekly meal plan for weight loss ***(task)***.
Output in a table ***(format)***.”

Chain of Thought

Improves LLM’s reasoning by going step by step.

Good for:
- Complex analytical tasks
- Decision making
- Problem solving

Prompt Example:
“[your prompt instructions].

Let’s think step-by-step”

Simply add “Let’s think step by step”

RISEN Framework

R - Role
I - Instructions
S - Steps
E - End Goal
N - Narrowing

Good for:
- Tasks with specific constraints (e.g. blog post, business plan)
- Tasks with clear guidelines (e.g. research)

Prompt Example:
“**Role:** *You’re a business strategist with experience in....*
Instructions: *Write an in-depth business plan about...*
Steps: *1. Start with..., 2. Proceed with..., 3. Proceed with...*
End Goal: *The goal is ...*
Narrowing (constraints): *Don’t use ...*”

RODES Framework

R - Role
O - Objective
D - Details
E - Examples
S - Sense Check

Good for:
- When you have examples similar to your desired output
- Marketing posts (e.g. Social media content, landing pages)

Prompt Example:
“**Role:** *You’re a content creator with experience in....*
Objective: *Write a X thread about...*
Details: *Include the following points....*
Examples: *Here are 3 examples of good X threads...*
Sense Check: *Do you understand the objective for this task?*”

Chain of Density

The Chain of Density is a prompt that uses recursion to create increasingly better outputs.

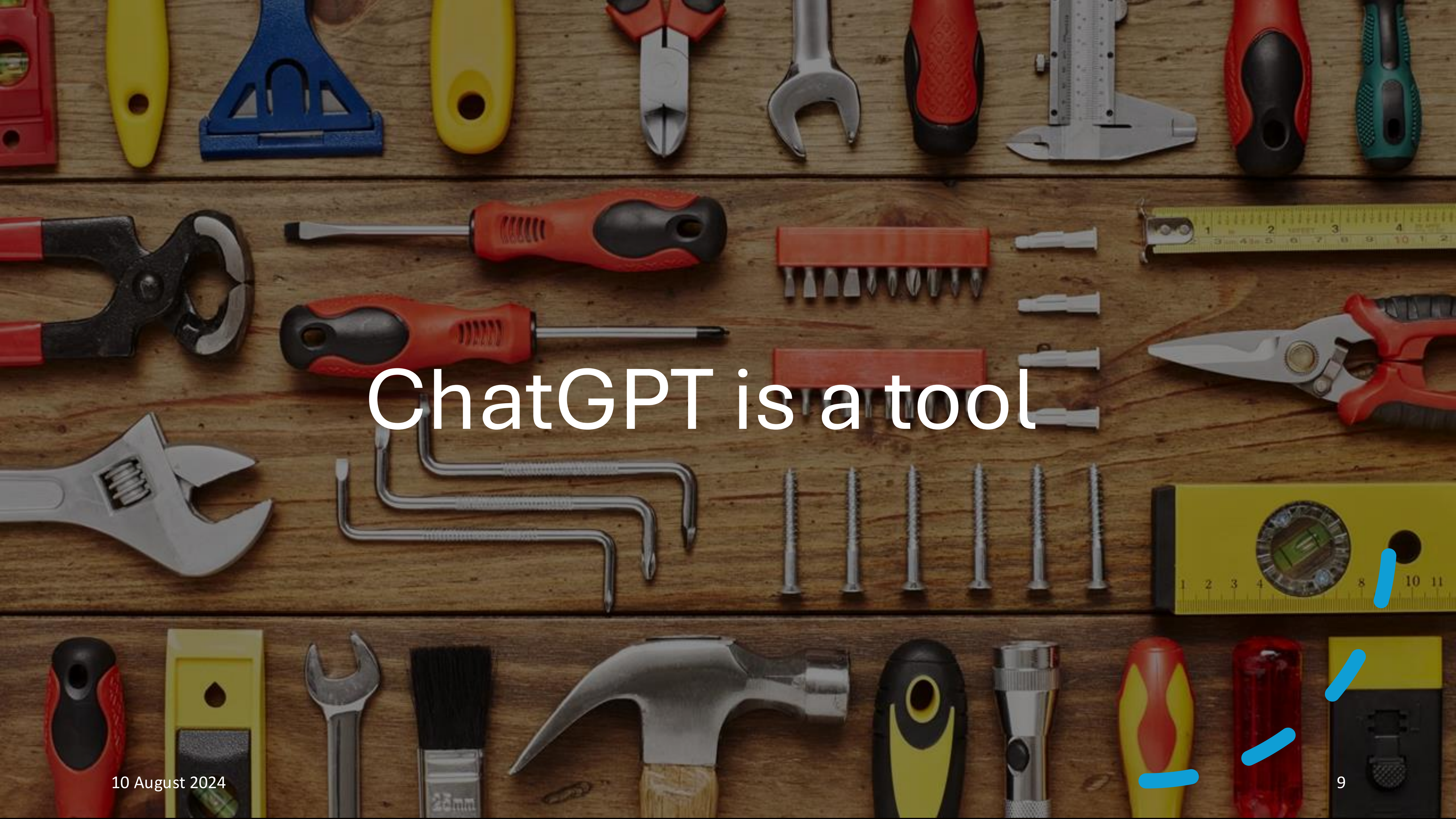
I - Instructions
R - Recursion
B - Benchmark
A - Additional Guidelines

Good for:
- Article summaries
- Improving your prompts
- Improving long-form content via recursion

Prompt Example:
“**Instructions:** Here is [insert content you want to improve]. You will generate increasingly improved versions of this content.
Recursion: Repeat the following 2 steps 5 times. Step 1. Identify 1-3 points from the initial output which are missing. Step 2. Write a new, improved output of identical length which includes the missing points.
Benchmark: Here is more information on what makes a good XYZ: [Insert info]
Additional Guidelines: Follow these specific [guidelines].”

Created by
Moritz Kremb

@moritzkremb
Moritz Kremb
thepromptwarrior.com



ChatGPT is a tool

What is R

- Open-source programming language environment
- Used especially for **statistical** computing and **graphics**
- Free
- *Steep* learning curve *– this is where ChatGPT is useful*
- Progressive **–** constant update *– again, this is where ChatGPT is useful*

slido

Please download and install the Slido app on all computers you use

How do you rate your skill using R?

ⓘ Start presenting to display the poll results on this slide.

For this workshop

1. Good to have two monitors
2. Install R, download from https://cran.r-project.org
3. Install RStudio from https://rstudio.com
4. Learning materials at https://github.com/profjamal/chatgpt
5. Data from https://github.com/MoH-Malaysia/covid19-public

The steps

1. Set your objective very clear
2. Understand the data (if they are not yours)
3. Prepare the data – download (link), clean (for missing values, outliers, etc), and visualise the data (in table)
4. Generate the coding from ChatGPT by using proper **prompting**
5. Run the coding in R
6. Verify the results – need your understanding of epid & stat

Set your analysis objectives

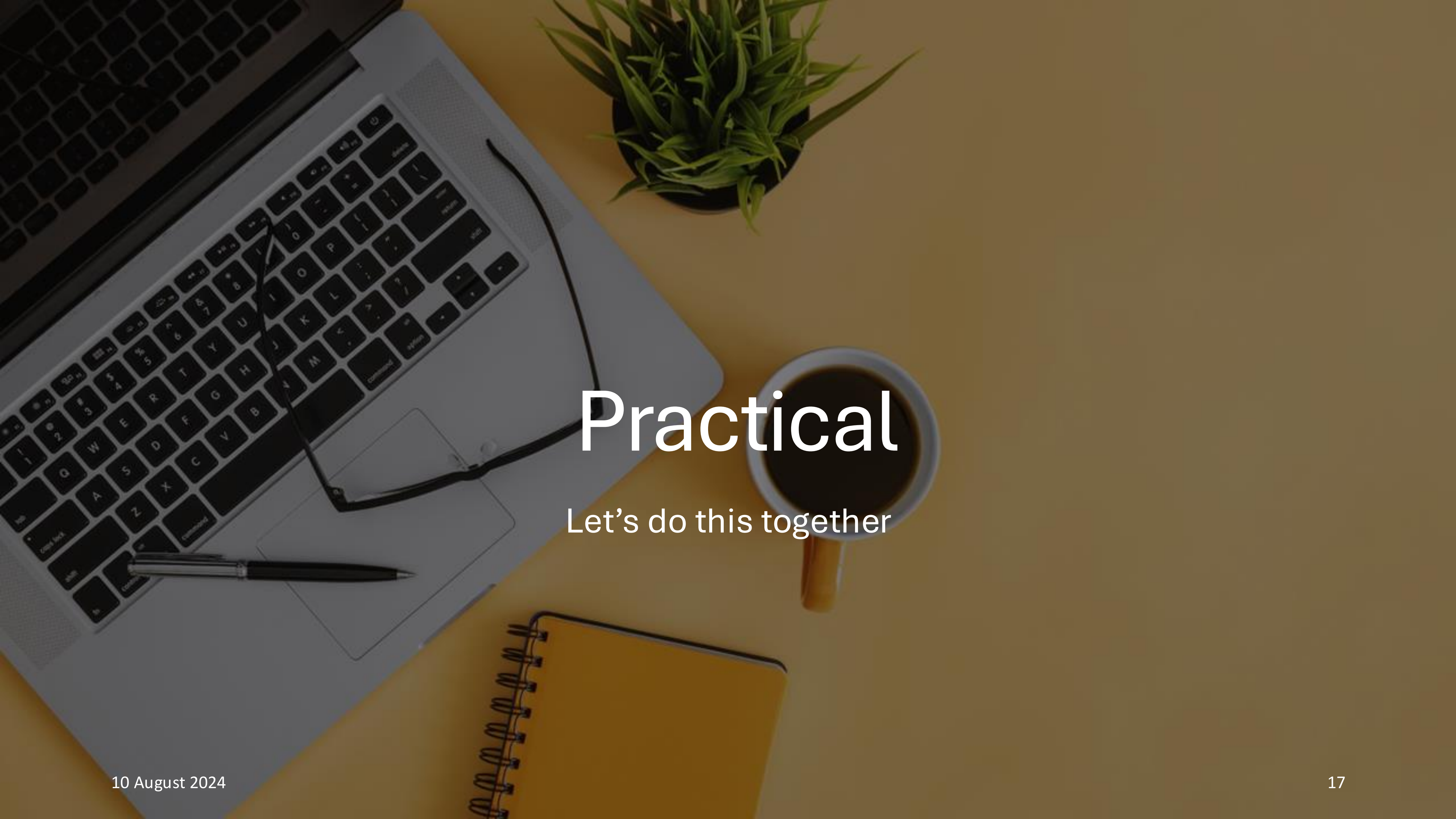
- Begin with the end in mind
- Set clear objectives
- For this workshop, our analysis is on Malaysia's COVID-19 death (using the line listing):
  - https://github.com/MoH-Malaysia/covid19-public/blob/main/epidemic/linelist/linelist_deaths.csv

The prompt, example

Role	Who are you?	Epidemiologist that is responsible for managing outbreaks in a population of 36,000,000 people
Task	What do you want?	Generally the aim is to evaluate the deaths from the COVID-19 outbreak in Malaysia. Specifically, 1) the overall incidence by year 2) its distribution based on age, sex, vaccination status, and type of vaccines.
Format	What output format do you want?	In R script and/or the visualisations

Tips

- Faster if we upload the data to ChatGPT
- Good to proceed in stages
    1. Preview data
    2. Clean data
    3. Analyse base on objective
    4. Request for visualisation
    5. Can even request the narrative
- At each stage, verify the codes and the result of the analysis
- Let the ChatGPT run the whole analysis first (in Phyton)
- Once you are satisfied, then you can ask the code in R, verify again



Practical

Let’s do this together

Let ChatGPT do all the analysis first

- https://chatgpt.com
- GPT-4, GPT-4o mini, GPT-4o
- Let’s start with your first prompt

Objectives

1. Calculate the overall mortality rate
2. Compare mortality rate by sex, vaccination status, vaccine doses

slido

Please download and install the
Slido app on all computers you use

Audience Q&A

ⓘ Start presenting to display the audience questions on this slide.