



香港浸會大學  
HONG KONG BAPTIST UNIVERSITY

# JOUR7280

## Big Data Analytics for Media and Communication

---

Instructor: Dr. Xiaoyi Fu

<https://github.com/shary777/JOUR7280>

# Contact Information

- Dr. Xiaoyi Fu
  - Office: TBC
  - E-mail: [xiaoyifuu@gmail.com](mailto:xiaoyifuu@gmail.com)  
[xiaoyifu@hkbu.edu.hk](mailto:xiaoyifu@hkbu.edu.hk)
- Time
  - MON 09:30-12:20
- Venue
  - CVA703

# Course Objective

- To introduce the fundamental knowledge and hands-on skills of big data analytics in the field of media and communication
- This course works on different types of (online) data
  - How to fetch data.
  - How to analyze, interpret, and visualize data.
- This course (may) be hard...
- This course is interdisciplinary
- This course is informative and helpful!

# Course Overview

Unit	Content	Week
Unit 1	Data science fundamentals and preliminary Python programming	Week 1-4
Unit 2	Automated web data collection	Week 5-8
Unit 3	Data (pre-)processing and data management	Week 8-10
Unit 4	Data exploration	Week 10-12

# Assessment Components

Type of Assessment	Weighting	Description of Assessment Tasks
Class participation & tutorial tasks	15%	Students will be introduced in lectures and guided readings to the key concepts and methods on data acquisition and processing in the digital age.
Individual exercises	25%	Students develop and test customized algorithms individually to collect and process social media data.
Group project and presentation	60%	Students work in teams to collect, process, and analyze social media data and present their findings in data product and an oral presentation.

# Assessment

- **Individual Exercises (25%)**
  - 1 automated online data acquisition challenge
  - 1 automated data processing (data cleaning) challenge
  - For each assignment:
    - 500– 800 words in English;
    - defining the problems
    - identifying the (online) data sources
    - presenting the codes and analytical process step by step
    - briefly reporting the results

# Assessment

- **Group Project (60%):** Presentation (30%) and report (30%)
  - 3 - 4 students per group;
  - Develop a data-driven investigation and storytelling project based on first-hand automated web data collection and present the research questions, methods and approaches, and findings;
  - In-class presentation, 12 - 15 minutes
    1. Appropriateness of materials;
    2. Proper application of data science skills;
    3. Organization – clear and logical flow, coherent and cohesive;
    4. Delivery – clear and focused presentation;

# Assessment

- **Group Project (60%):** Presentation (30%) and report (30%)
  - Written report
    1. No more than 12 A4 pages in English (excluding references, codes, and appendix)
    2. Focusing on questions, and data analytical steps (the pipelines)
    3. Interpreting the results
  - Declare workload distribution
    - NO “free rider” please.
    - You can discuss, negotiate and allocate the workload on your own within your team.
    - A declaration form will be submitted with your final project report.
    - During the in-class presentation, a table clearly show the labor distribution is required.



# Class Rules

- On Class-Participation
  - No attendance requirement, but the participation is counted into the final assessment.
  - The regular participation
  - Please be punctual

# Class Rules

- In-class rules.
  - “Silence is golden” - as long as you keep quite...
  - Eating and drinking?
  - You may bring your own laptop



香港浸會大學  
HONG KONG BAPTIST UNIVERSITY

# Data Science and Media Data Analytics at a Glance

---

# Agenda

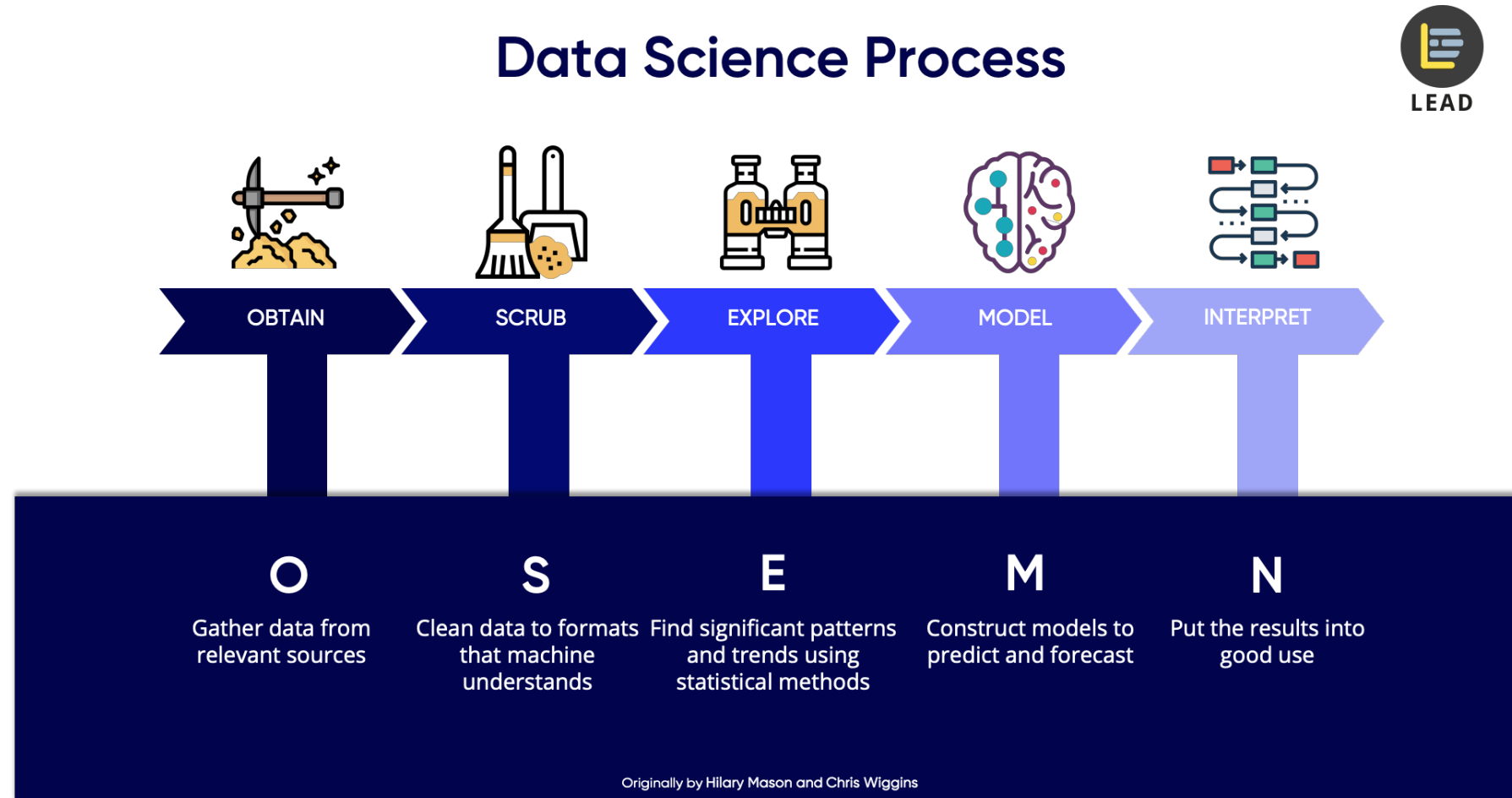
- **What**
  - What do data scientists do?
  - What is social media data acquisition and processing?
- **Why**
  - Why shall I learn this course?
  - Why web/social data?
- **How** (to start and succeed)
  - Tools installation
  - Social scientists' and humanities scholars' domain knowledge
  - Practice and getting your hands dirty!

# Data science in action

- Define the problem
- Scouting the data sources
- Accessing to and collecting the data
- (Pre-)processing and cleaning the data
- Exploring the data
- Analyzing the data
- Interpreting the results
- Offering insights and solutions
- ...

# Data science pipeline - a verbal explanation

- The “OSEMN Pipeline”



# Data science pipeline

## 1. Obtain Data

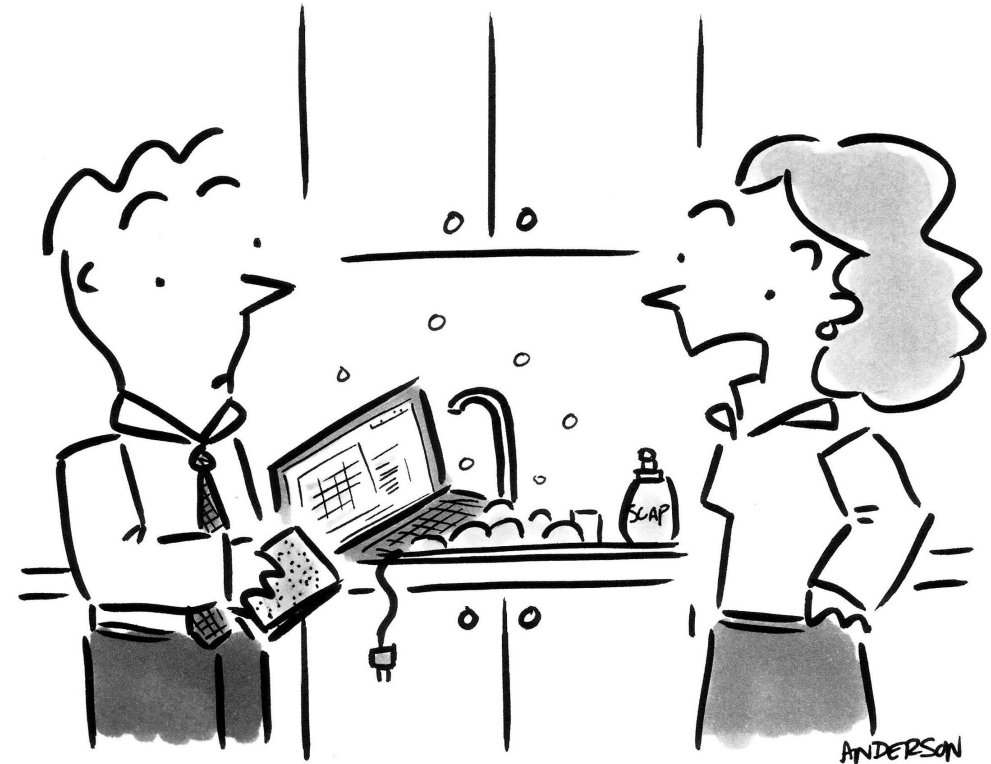
- You cannot do anything as a data scientist without even having any data
- **Identify** all of your available datasets (which can be from the internet or external/internal databases)
- Extract the data into a **usable format** (.csv, json, xml, etc..)



# Data science pipeline

## 2. Scrub / Clean Data

- **Examine the data:** understand every feature you're working with, identify errors, missing values, and corrupt records
- **Clean the data:** throw away, replace, and/or fill missing values/errors
- Garbage in garbage out



"This is not what I meant when I said 'we need better data cleansing!'"



# Data science pipeline

## 3. Explore Data

- Find patterns in your data through visualizations and charts
- Derive hidden meanings behind our data through various graphs and analysis



# Data science pipeline

## 4. Model Data

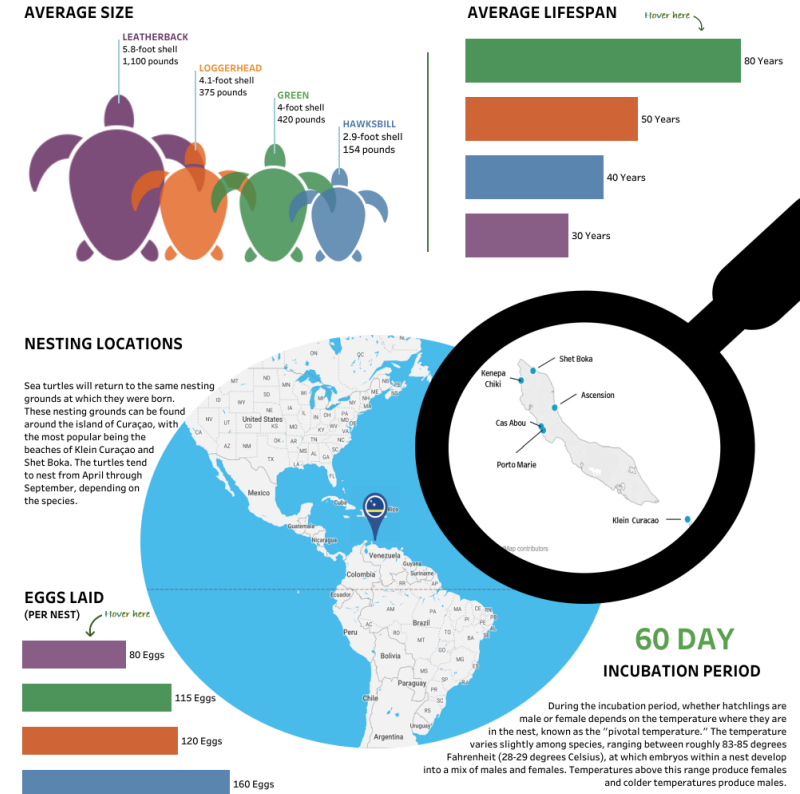
- In-depth Analytics create predictive models/algorithms
- Skills Required Machine Learning algorithms, Linear algebra etc.
- Predictive Power Example Walmart predicted that they would sell out all of their Strawberry Pop-tarts during the hurricane season in one of their store location.
  - Historical data showed that the most popular item sold before the event of a hurricane was Pop-tarts.

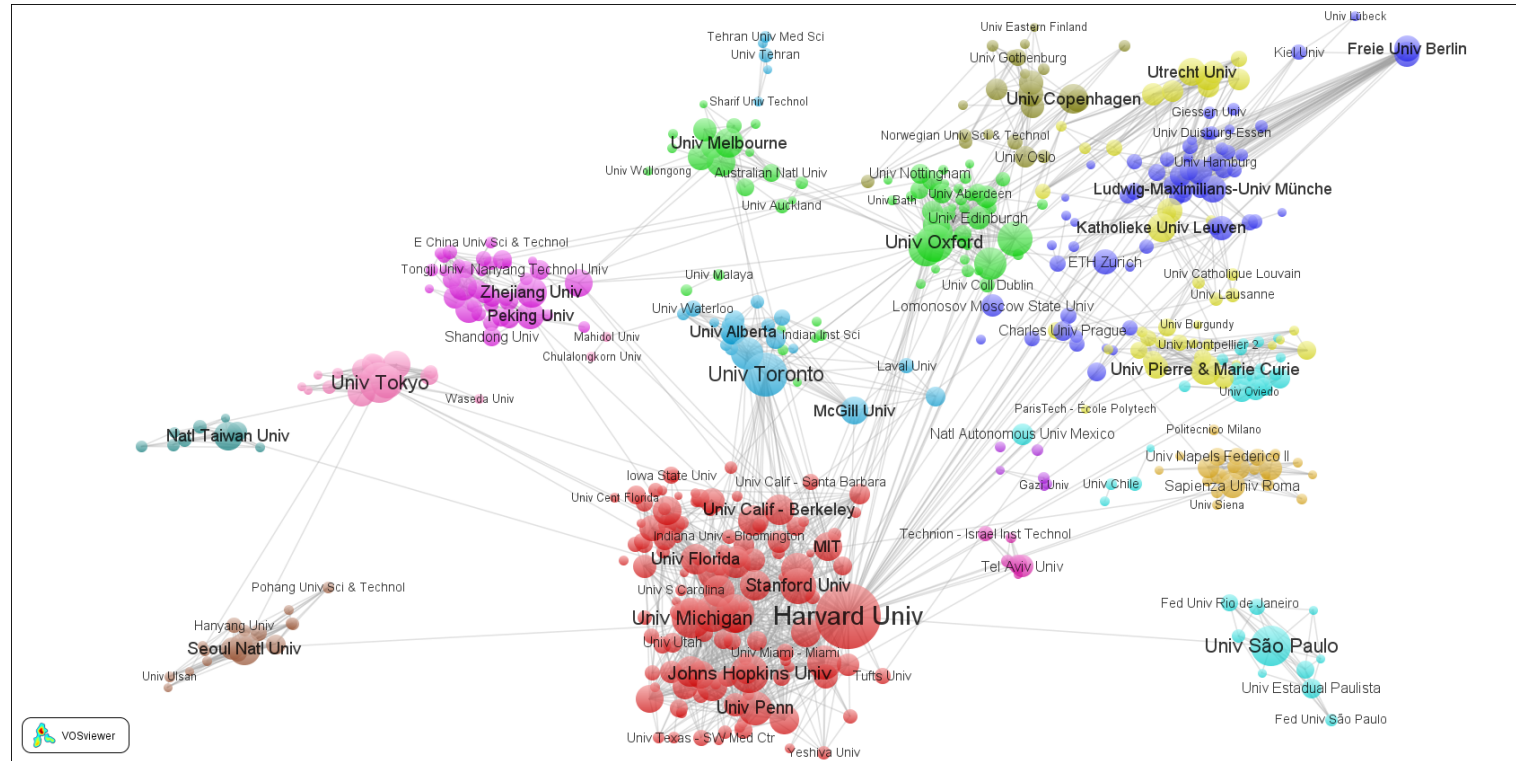


# Data science pipeline

## 5. Interpret Data (Data Storytelling)

- Visualize your findings accordingly: keep it simple and priority driven
- Tell a clear and actionable story: effectively communicate to non-technical audience





Co-author + university collaboration network

# Data acquisition and processing

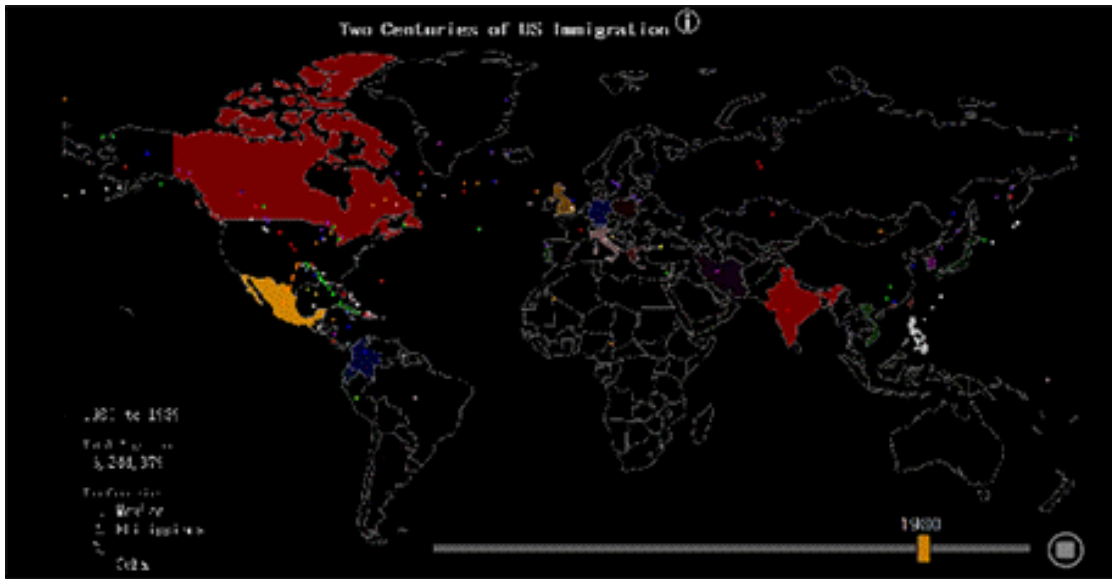
- What is social media (web) data acquisition?
  - Web data collection, data acquisition, screen scraping, data mining, web data harvesting, web crawlers...
  - “Social media” – web data, cultural and social artifacts
- Web scraping
  - It is accomplished by writing an automated program that queries a web server, requests data (usually in the form of HTML and other files that compose web pages), and then parses that data to extract needed information (Mitchell, 2018).
- Automated data collection via social media or institutions
  - The usage of APIs (Twitter, Fb, Weibo)
- It is the process of taking unstructured information from the web (webpages, sites, social media services, institutions) and turning it in to structured information that can be used in a subsequent stage of analysis.

# Why taking this class?

- Increasing market demands
- The availability of digital footprints (digital traces)
- Internet as a rich database
- The field itself is becoming more interdisciplinary.
- The price and difficulty of collecting and storing massive online data has dramatically reduced.

# Antisyllabus

- Basic data visualization in Python, as opposed to professional data visualization tools

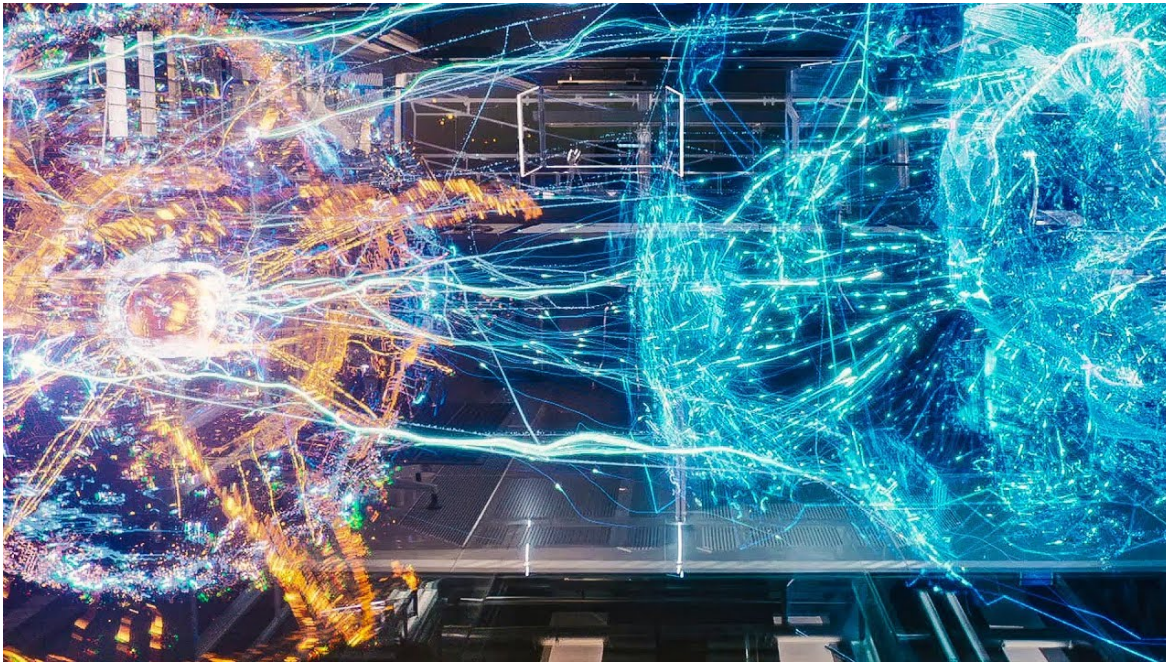


<http://metrocosm.com/us-immigration-history-map.html?ref=producthunt>[http://metrocosm.com/global-immigration-map/?utm\\_content=buffer113b2&utm\\_medium=social&utm\\_source=twitter.com&utm\\_campaign=buffer](http://metrocosm.com/global-immigration-map/?utm_content=buffer113b2&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer)



# Antisyllabus

- Data analytics, without AI/machine learning





# Getting started: Tools

- Talking to your computer: Command line interface (CLI)
- Text Editor: vscode, sublime, vim, emacs, or others
- Platform for publishing and socializing: Git and GitHub, and Markdown language
- The tool: Python 3.x (Anaconda 3) and Jupyter Notebook

# Thank You

