# Introduction to Data Science - Topic 1


- Instructor: Professor Henry Horng-Shing Lu,
  Institute of Statistics, National Yang Ming Chiao Tung University, Taiwan
  Email: henryhslu@nycu.edu.tw

- WWW: http://misg.stat.nctu.edu.tw/hslu/course/DataScience.htm
- Classroom: ED B27 (新竹市大學路1001號工程四館B27教室)
- References:
  M. A. Pathak, Beginning Data Science with R, 2014, Springer-Verlag.
  K.-T. Tsai, Machine Learning for Knowledge Discovery with R: Methodologies for Modeling, Inference, and Prediction, 2021, Chapman and Hall/CRC.
- Evaluation: Homework: 70%, Term Project: 30%
- Office hours: By appointment

# Course Outline

**10 Topics and 10 Homeworks:**

- **Introduction of Data Science**
- **Introduction of R and Python**
- **Getting Data into R and Python**
- **Data Visualization**
- **Exploratory Data Analysis**
- **Regression (Supervised Learning)**
- **Classification (Supervised Learning)**
- **Text Mining**
- **Clustering (Unsupervised Learning)**
- **Neural Network and Deep Learning**

# Topic 1: Introduction to Data Science

**References:**

- References: Ch. 1, M. A. Pathak, Beginning Data Science with R, 2014, Springer-Verlag.

  https://en.wikipedia.org/wiki/Data_science

# What Is Data Science?

- Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining. [wiki]

- Video: The Human Face of Big Data

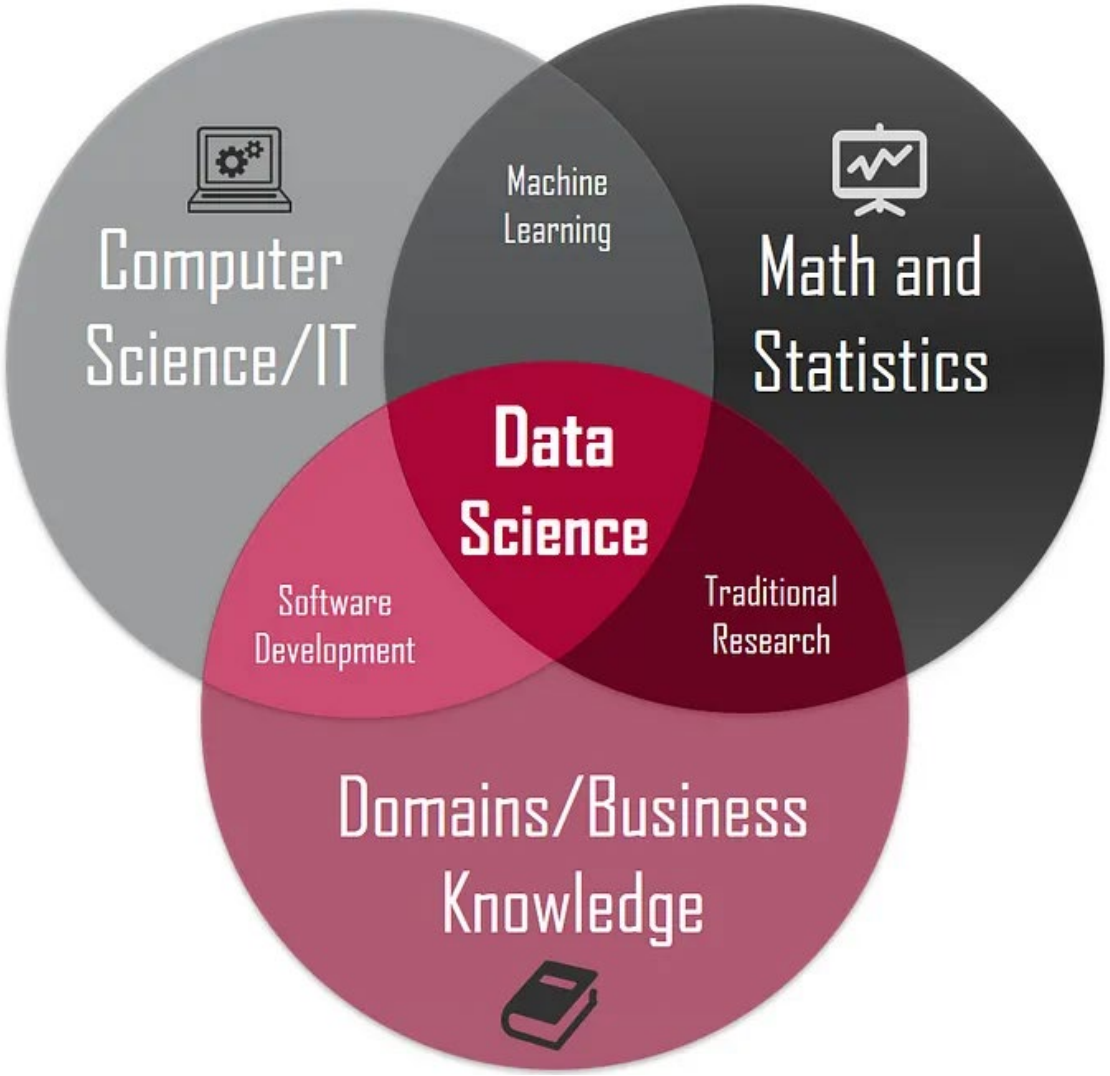
  https://www.youtube.com/watch?v=4VeITe6EJDU

  (with Chinese caption)

  https://ihavenotv.com/the-human-face-of-big-data-nova-pbs

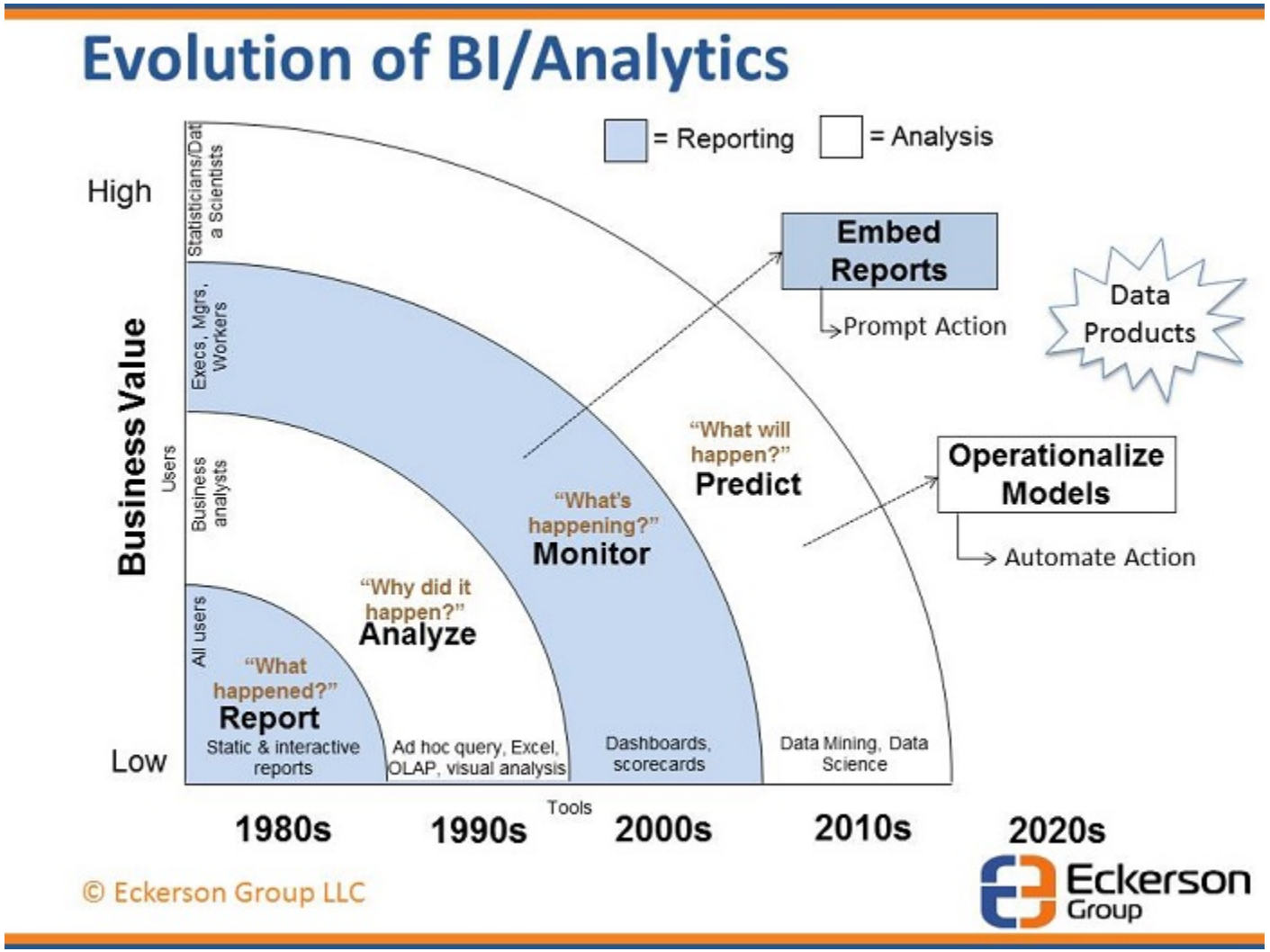
  (without caption)

# Data Science (1)



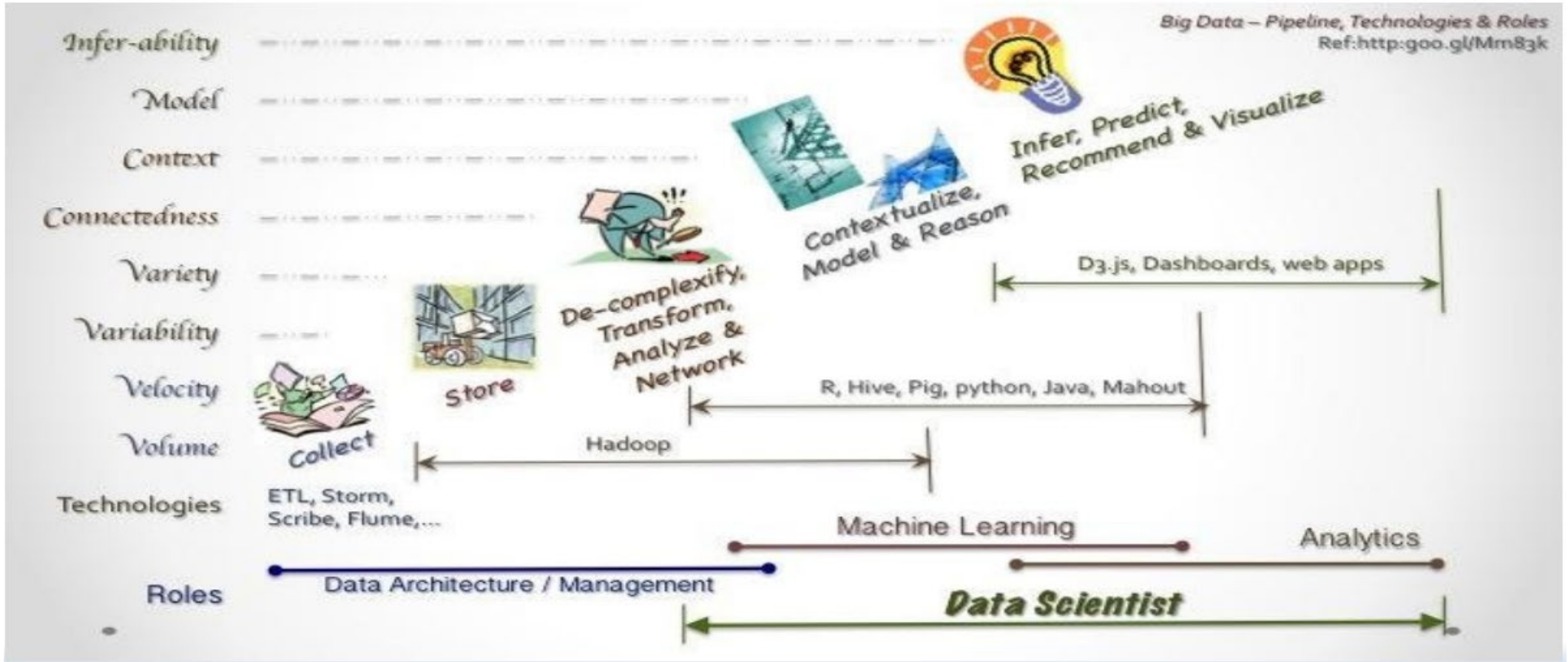

https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

# Data Science (2)



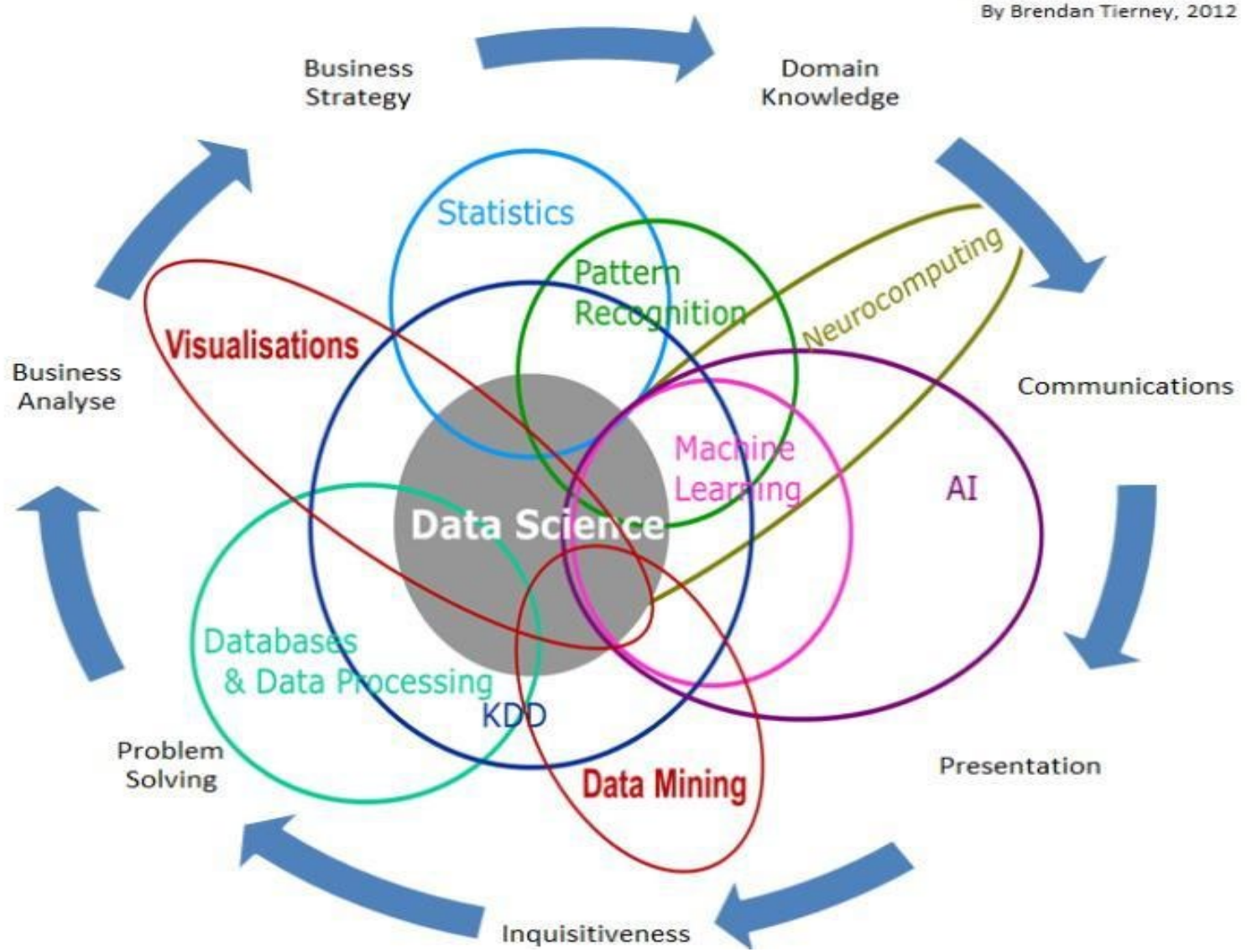

https://tdwi.org/articles/2016/09/08/embedded-analytics-future-of-bi.aspx

# Data Science (3)

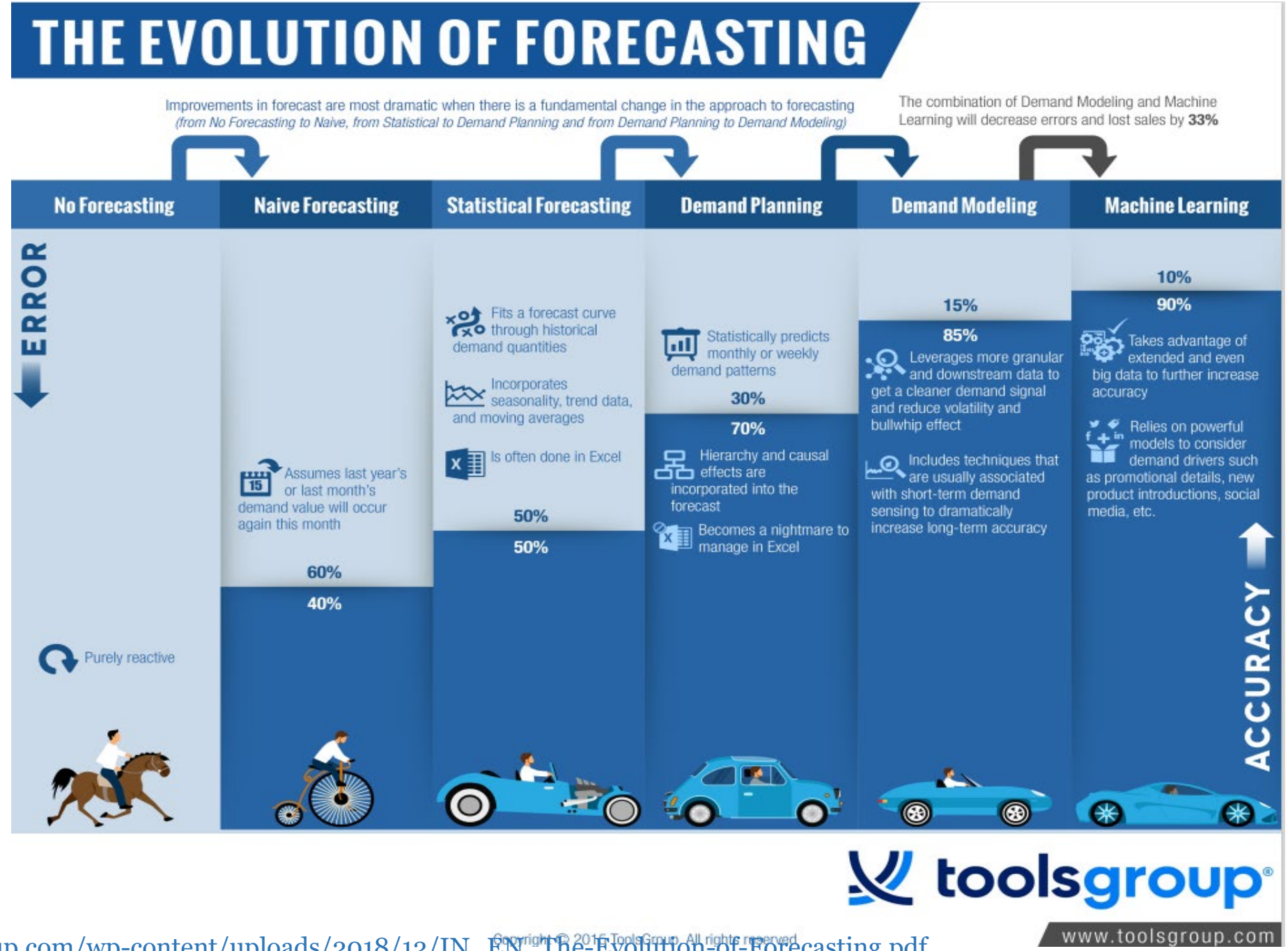



https://doubleclix.wordpress.com/category/data-science/page/3/

# Data Science Is Multidisciplinary




https://www.datasciencecentral.com/difference-of-data-science-machine-learning-and-data-mining/






https://www.toolsgroup.com/wp-content/uploads/2018/12/IN_EN_The-Evolution-of-Forecasting.pdf

# Why R ?

- Open Source
- R works on most platforms: GNU/Linux, OS X, Windows
- There are many high-level data types and functions available in R
- Complete documentation and discussion area
- The program is easy to modify according to the needs

# Introduction to ChatGPT

# ChatGPT

- ChatGPT is an advanced AI chatbot trained by OpenAI that interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

- **How to use ChatGPT**

  Step 1: **Create an OpenAI account**. Go to chat.OpenAi.com and register for an account with an email. You need to create an account on the OpenAI website to log in and access ChatGPT.

  Step 2: **Accept ChatGPT terms.**

  Step 3**: Start writing.**

https://chat.openai.com/c/e7ad2919-0e3a-495a-ab1c-5e1f65589b19

# **ChatGPT** (Chat Generative Pre-trained Transformer)

Answer questions based on existing knowledge.

I am a highly intelligent question-answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery or has no clear answer, I will respond with "Unknown".

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was the president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?
A: He belonged to the Republican Party.

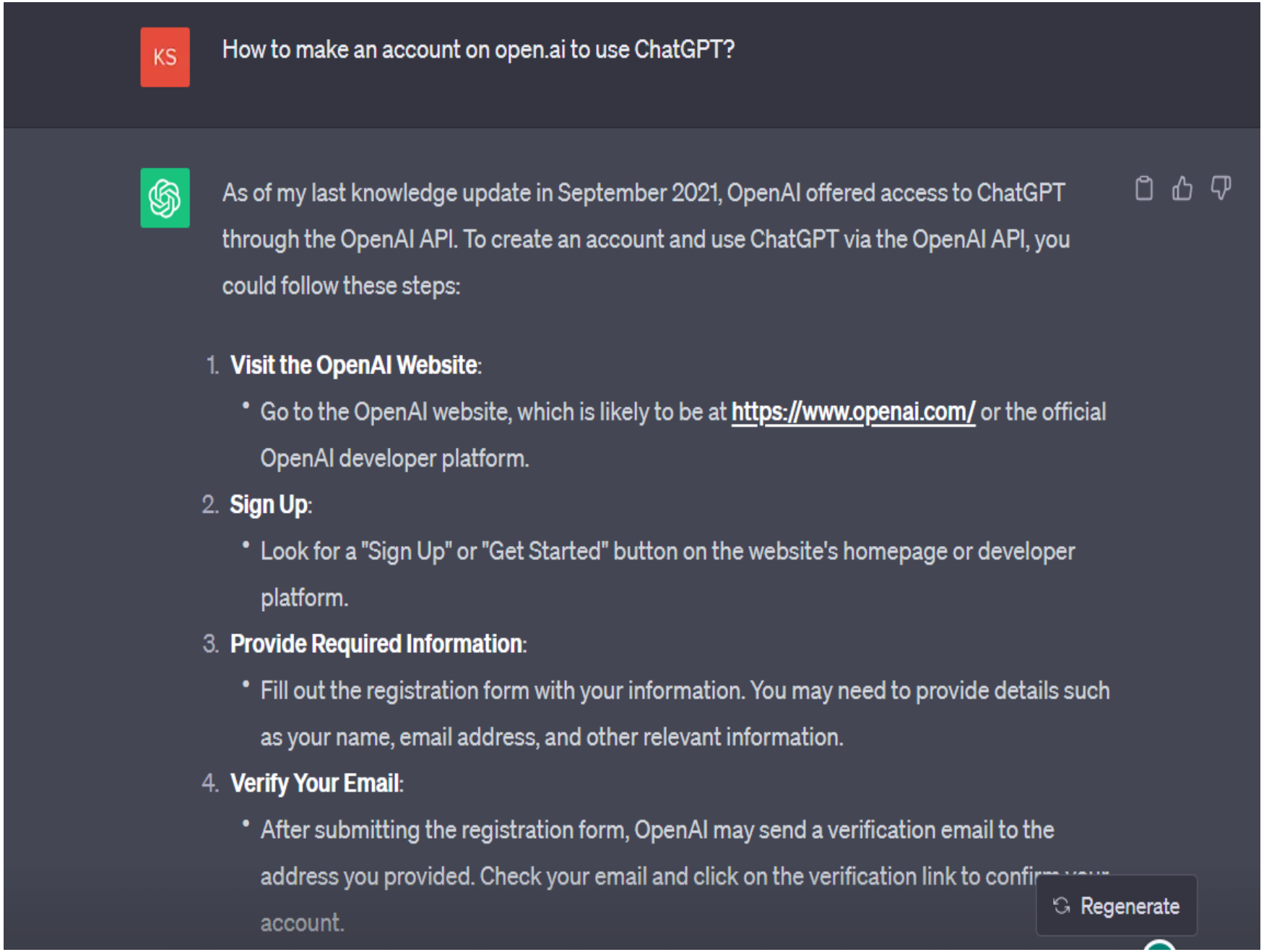
Q: What is the square root of banana?
A: Unknown

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

https://platform.openai.com/examples/default-qa

# Some Questions Answered by ChatGPT


https://chat.openai.com/chat



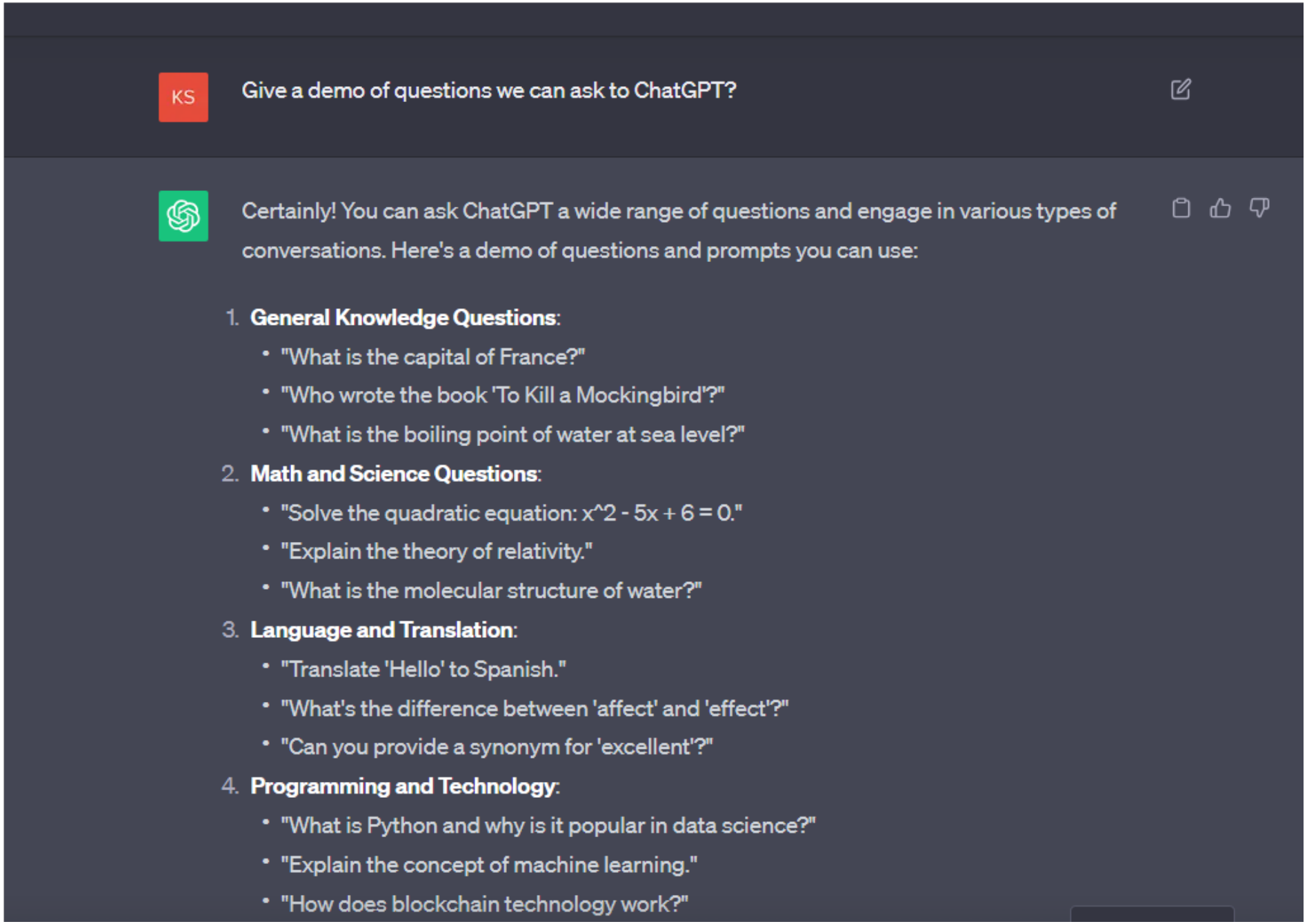

https://chat.openai.com/chat

5. **Log In**:
   - Once your account is verified, you can log in to your OpenAI account using the email address and password you used during registration.
6. **Navigate to the API Section**:
   - Explore the OpenAI website or developer platform to find information related to the API and ChatGPT access.
7. **Request API Access**:
   - Follow the instructions provided by OpenAI to request access to the ChatGPT API. This may involve agreeing to terms of use, privacy policies, and potentially payment or subscription details, depending on OpenAI's policies at the time.
8. **Get API Key**:
   - If your API access request is approved, OpenAI will typically provide you with an API key or authentication token. This key will be used to authenticate your requests to the ChatGPT API.
9. **Read Documentation**:
   - Familiarize yourself with the API documentation, which will provide details on how to make API requests, what parameters to use, and how to handle responses.
10. **Start Using ChatGPT**:
    - Use the API key and the documentation to integrate ChatGPT into your applications or use it for various natural language processing tasks.

Please note that OpenAI's policies, pricing, and procedures may change over time. Therefore, it's essential to visit the OpenAI website or developer platform and refer to their most current documentation and terms of use for the most up-to-date information on creating an account and accessing ChatGPT through the OpenAI API.

Regenerate

https://chat.openai.com/chat




https://chat.openai.com/chat

5. **Health and Medical Questions:**
   - "What are the symptoms of COVID-19?"
   - "How can I lower my cholesterol naturally?"
   - "What is the recommended daily intake of vitamins?"
6. **Travel and Geography:**
   - "Tell me about popular tourist attractions in Tokyo."
   - "What are the top-rated restaurants in Paris?"
   - "How can I apply for a tourist visa to the United States?"
7. **Entertainment and Pop Culture:**
   - "Who won the Academy Award for Best Actor in 2022?"
   - "What are some must-watch TV shows on Netflix?"
   - "Tell me a joke."
8. **History and Historical Figures:**
   - "Who was Leonardo da Vinci?"
   - "What were the main causes of World War II?"
   - "When was the Declaration of Independence signed?"
9. **Opinion-Based Questions:**
   - "What is your opinion on climate change?"
   - "Do you think artificial intelligence will replace human jobs?"
   - "What's the best way to motivate oneself to exercise?"
10. **Creative Writing Prompts:**
    - "Write a short poem about nature."
    - "Create a fictional story about a detective solving a mystery."
    - "Compose a haiku about the changing seasons."

Remember that while ChatGPT can provide information and generate text, it's important to critically evaluate its responses, especially for factual accuracy and ethical considerations. ChatGPT generates text based on patterns in the data it was trained on and may not always provide completely accurate or up-to-date information. Additionally, it's crucial to use AI responsibly and be mindful of the content and context of your questions and prompt

Regenerate

https://chat.openai.com/chat

KS

example of conversion on ChatGPT?

Certainly! Here's an example of a conversation with ChatGPT, where a user interacts with the model by asking questions and having a dialogue:

User: Hello! What's your name?

ChatGPT: Hi there! I'm ChatGPT, a language model created by OpenAI. How can I assist you today?

User: Nice to meet you, ChatGPT. Can you tell me a joke?

ChatGPT: Of course! Here's one: Why don't scientists trust atoms? Because they make up everything!

User: Haha, that's a good one! How about a fun fact?

ChatGPT: Sure thing! Did you know that honey never spoils? Archaeologists have even found pots of honey in ancient Egyptian tombs that are over 3,000 years old and still perfectly edible.

User: That's fascinating! Thanks for sharing. What's the weather forecast for New York City tomorrow?

ChatGPT: I'm sorry, but I don't have access to real-time data like weather forecasts. I recommend checking a weather website or using a weather app for the most up-to-date information.

https://chat.openai.com/chat

User: No problem. How does photosynthesis work?

ChatGPT: Photosynthesis is the process by which plants, algae, and some bacteria convert sunlight into energy. It involves capturing light energy using chlorophyll and using it to convert carbon dioxide and water into glucose and oxygen. It's a crucial process for sustaining life on Earth.

User: Thanks for the explanation! What are some good books to read for someone interested in science fiction?

ChatGPT: If you're into science fiction, you might enjoy classics like "Dune" by Frank Herbert, "Neuromancer" by William Gibson, and "Foundation" by Isaac Asimov. These are just a few recommendations, and there are many more great sci-fi books out there to explore.

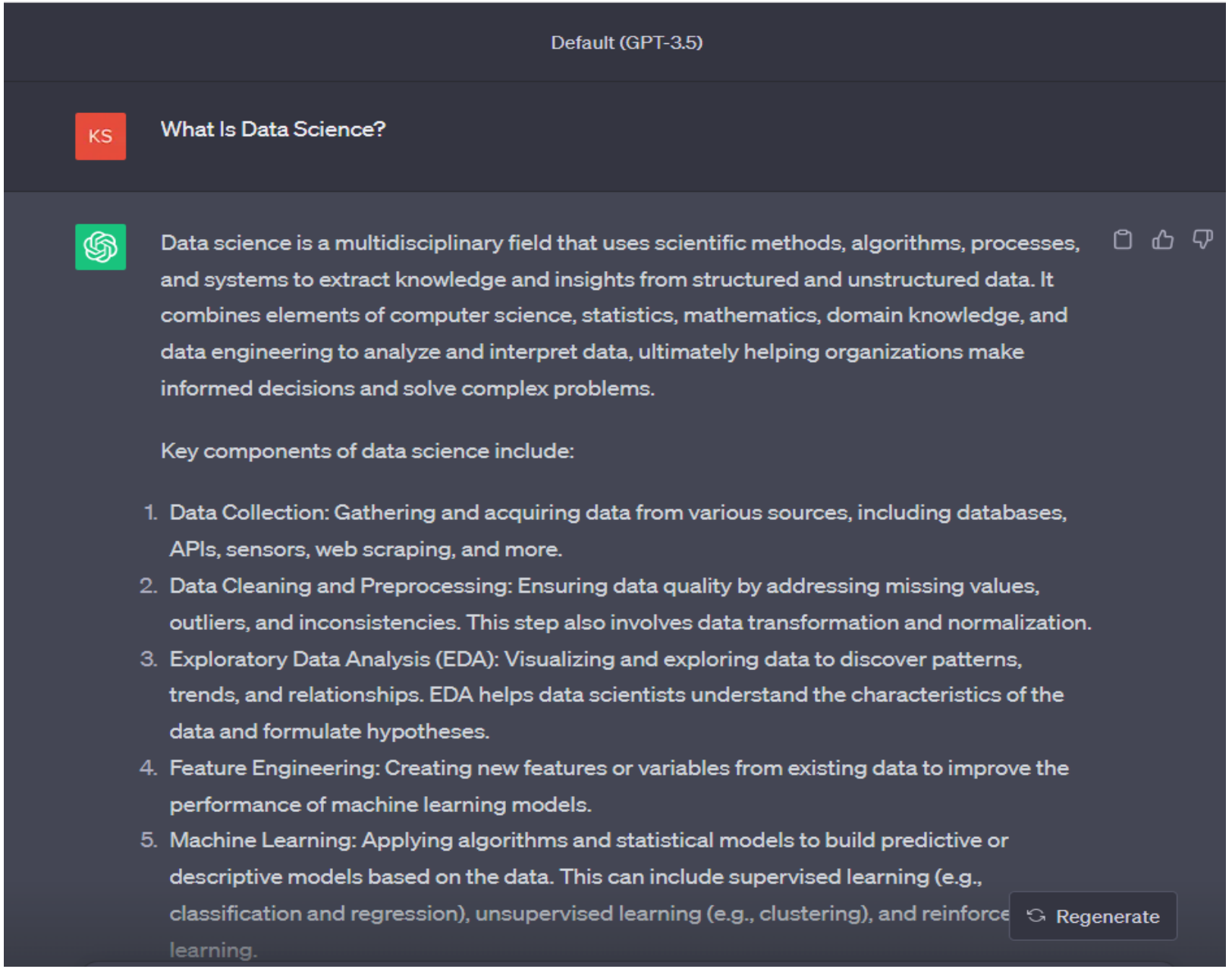
User: Great suggestions! I'll check those out. One last question: Can you summarize the plot of "The Great Gatsby" by F. Scott Fitzgerald?

ChatGPT: Certainly! "The Great Gatsby" is a novel set in the 1920s and tells the story of Jay Gatsby, a mysterious millionaire known for his extravagant parties. The plot revolves around Gatsby's obsession with his lost love, Daisy Buchanan, and explores themes of wealth, decadence, and the American Dream.

User: Thanks for the summary, ChatGPT. It was a pleasure talking to you!

ChatGPT: You're welcome! I enjoyed our conversation as well. If you ever have more questions or just want to chat, feel free to reach out. Have a great day!

https://chat.openai.com/chat




https://chat.openai.com/chat

5. Machine Learning: Applying algorithms and statistical models to build predictive or descriptive models based on the data. This can include supervised learning (e.g., classification and regression), unsupervised learning (e.g., clustering), and reinforcement learning.
6. Data Visualization: Communicating insights and findings through visual representations such as charts, graphs, and dashboards to make data-driven decisions accessible to non-technical stakeholders.
7. Model Evaluation and Validation: Assessing the performance of machine learning models using metrics like accuracy, precision, recall, F1-score, and cross-validation to ensure their reliability.
8. Deployment and Productionization: Implementing data-driven solutions in real-world applications and maintaining them for ongoing use. This may involve integrating models into software systems or deploying them on cloud platforms.
9. Data Ethics and Privacy: Addressing ethical considerations and ensuring compliance with data privacy regulations to protect individuals' sensitive information.

Data scientists use a variety of programming languages and tools, such as Python, R, SQL, and specialized libraries and frameworks like TensorFlow and scikit-learn. They also work closely with domain experts to understand the context of the data and frame relevant questions for analysis.

The applications of data science are vast and include areas like finance, healthcare, marketing, e-commerce, social sciences, and many others, where data-driven insights can drive better decision-making, optimize processes, and uncover hidden opportunities.

https://chat.openai.com/chat

KS

why to use R?

R is a programming language and environment specifically designed for statistical analysis, data visualization, and data manipulation. It offers several advantages that make it a preferred choice for certain tasks and in specific contexts:

1. **Statistical Analysis**: R was developed by statisticians, and it excels in statistical modeling and analysis. It provides a wide range of statistical functions and packages for regression analysis, hypothesis testing, time series analysis, and more. Researchers, statisticians, and data analysts often find R to be a powerful tool for in-depth statistical work.
2. **Data Visualization**: R has a robust ecosystem for data visualization, including packages like ggplot2, which allows users to create highly customizable and publication-quality plots and charts. Data visualization is crucial for exploring data and conveying insights to stakeholders.
3. **Data Manipulation**: R offers powerful data manipulation capabilities through libraries like dplyr and tidyr. These libraries make it easy to filter, transform, and reshape data, facilitating the data preprocessing stage common in data analysis.
4. **Community and Packages**: R has a vibrant and active community of users and developers. The Comprehensive R Archive Network (CRAN) hosts thousands of packages for various data analysis and statistical tasks. This extensive package ecosystem makes it easy to find solutions for specific data science needs.

https://chat.openai.com/chat

solutions for specific data science needs.

5. **Reproducibility**: R scripts and analyses can be easily documented and shared, promoting transparency and reproducibility in research and data analysis. This is essential in scientific and research settings.
6. **Integration**: R can be integrated with other programming languages like C, C++, and Python, which can be useful when certain tasks or algorithms are better suited for those languages.
7. **Open Source**: R is open-source software, which means it is freely available to anyone. This accessibility has contributed to its widespread adoption in academia and industry.
8. **Cross-Platform Compatibility**: R is available for multiple platforms, including Windows, macOS, and various Linux distributions, ensuring broad compatibility.
9. **Community Support**: R has a large and active online community, which means that users can find help, tutorials, and resources easily through forums, mailing lists, and online documentation.

However, it's important to note that R may not be the best choice for all data science tasks. For example, for tasks involving deep learning, large-scale data processing, or web development, other languages like Python might be more suitable. The choice of programming language depends on the specific requirements of the project, the existing tools and skills of the team, and the nature of the data analysis tasks at hand. Many data scientists choose to use both R and Python in their work, taking advantage of the strengths of each language.

https://chat.openai.com/chat

# Homework 1 (submitted to e3.nycu.edu.tw before Sept 27, 2023)

- Find at least one data set that you plan to study for your future homeworks and final project.
- Explain the features in your data set.
- Discuss possible problems you plan to investigate based on the data sets you select.

**Possible sources of open datasets:**

- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php)
- Kaggle Datasets (https://www.kaggle.com/datasets)