Session 1: Mitr Phol

- Skills and tools
    - Fundamental and Advanced Skill
        - Statistical analysis
        - Data management
        - Machine learning
    - Soft Skill
        - Problem-solving
        - Communication
        - Business insight
    - Tools
        - Analysis and visualization
        - Big data technologies
- A day in life of a Data Scientist
    - Morning routine -> Check data pipelines -> Data review -> Model Monitoring
    - Collaborative Effort -> Stand-up meeting -> Cross Functional Collaboration
- Data Preparation
    - Data source -> Data Cleaning -> Data Transformation -> Data Warehouse -> Data Mart -> Data Visualization/Analytics
- Career Paths
    - Data Engineer: See data and data flow first.
    - ML Engineer: Use ML to do something with data.
    - Data Scientist: Engages in more complex problems and predictive analytics.
    - Data Analyst: Focus on interpreting data, generating report and dashboard.

Session 2: Data Engineer

- แนวคิด
    - Model-centric คือการพยายามปรับโครงสร้างให้ดีที่สุด
    - Data-centric คือพยายามทำ data ให้ดีที่สุด (ต่อให้นำไปใช้ใน model ห่วยๆ)
- บทบาท
    - Data Engineer เป็นคนหาข้อมูล (เหมือนคนเตรียมวัตถุดิบ เครื่องปรุง)
    - Data Analysts เป็นคนที่กระทำกับข้อมูล (เหมือนคนปรุงอาหาร)
    - Data Scientist เป็นคนคิด/สร้างโมเดล (คนคิดสูตรอาหาร)
- การพัฒนา AI ต้องแบ่ง work load ระหว่าง Application กับ Analytics ไม่งั้น Database อาจล่ม
    - Application workload
        - Application + Backend Service
        - Transactional data
        - Read/Write/Update (Databse ACID)
    - Analytics workload
        - Query
        - Dashboard / Historical Data
        - Write once, Read many (Data warehouse, Data Lake, Data Mart)
- AI system = Software + Model
    - Gathering Data -> Learning -> Applying the model
    - Data : Data engineering pipeline – Data quality is a key
    - Model: Machine learning pipeline – Train, Evaluate, Test <- Automatic
    - Software: Model serving and predictions (Usage)

- Rule of Machine Learning
    - To make great products: Do machine learning like the great engineer you are, **not like the great machine learning expert you aren't**

Session 3: Open Thai GPT

- ทำ OpenThaiGPT ทำไม
- ChatGPT ของต่างประเทศมีการเกิด **Lost in Translation** (การแปลผิด) เพราะขณะที่เรา input ภาษาไทยลงใน model ต่างประเทศจะทำให้ model ต้องแปลภาษาเป็นภาษาอังกฤษก่อนจะนำเข้า model และ output ของ model ก็ต้องถูกแปลกลับอีกที
    - ข้อมูลบางอย่างเป็นความลับ หรือ sensitive data
- หลักการของ **GPT (Generative Pretrained Transformer)** คือ AI ที่เรียนแบบการเติมคำของมนุษย์
- **LLM (Large Language Model)** คือ Transformer Model โดยหลักการคือนำ input text (prompt) ผ่าน transformer และสร้าง output text โดยใช้วิธีการเดาคำจากบริบท
    - Word Collocations, Phrase Structure, Translation Alignment (การ map ระหว่าง input และ output ซึ่งสิ่งนี้สำคัญที่สุด)
    - การเทรน model ต้องการอย่างน้อย 1 billion words
- ภายใน Transformer model คือ Scaled Dot-Product Attention ซึ่งมี concept คล้าย search engine โดยการใช้ query ซึ่งคือ vector ของแต่ละคำมา dot เพื่อหาความคล้ายกับ keys ที่เป็น vector ของแต่ละคำในประโยคจะได้ออกมาเป็น weight แล้วนำ weight นี้ไปคูณกับ value ที่โดน scale แล้ว จึงมานำมารวมกันเป็น combined results ที่สามารถนำมาใช้หา idioms ของ query ได้ด้วย self-attention
- Cross attention คือการทำ Scaled Dot-Product Attention ที่มี queries กับ keys เป็นคนละภาษากัน คล้ายกับการแปลภาษา
- Multihead Attention คือการทำ Attention หลายตัวซ้อนกันเพื่อหา idiom ใน idiom หรือ idiom ซ้อน idiom เมื่อเราซ้อนไปเรื่อย ๆ ก็จะทำให้เราสามารุเรียนรู้ Phrase Structure ได้
- Alignment of Phrase Structure คือการ multihead attention ของ 2 ภาษาแยกกัน และนำผลลัพธ์ของ 2 multihead attention นี้มาเป็น query ของ multihead attention อีกตัว

- Google Gemini คือ Retrieval-Augmented Generation (RAG) หรือ Search Engine + LLM โดยมี model ในการตัดสินว่าจะค้นหา หรือ ตอบด้วยความรู้ที่มี
- Multiview Knowledge Distillation คือการเทรนโมเดลเพื่อโดนความรู้โดยใช้ข้อมูลจากหลายมุมมองหรือหลายแหล่ง เพื่อช่วยลดการ overfitting
- Multimodal LLM คือ LLM ที่สามารถเรียนรู้อย่างอื่นนอกจากข้อความได้ เช่น ภาพ โดยการเปลี่ยนเป็น vector
- Challenging Problem ในตอนนี้
    - AI hallucination คือ AI หลอนตอบไม่ตรงคำถาม แรกพูกอีกอย่างหลังๆพูดอีกอย่าง พูดไปเรื่อย ตอบผิดอย่างชัดเจน