



INFH 5000 – TA Presentation

Data Science in Action: Insights from Industry and Academia

Present by Changlun LI, INFH-DSA-PhD Student

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Overview



Data Science Bootcamp

Definition
Evolution



Industry Viewpoint

Project lifecycle
Career outlook



Academy Viewpoint

Education
Trending Research



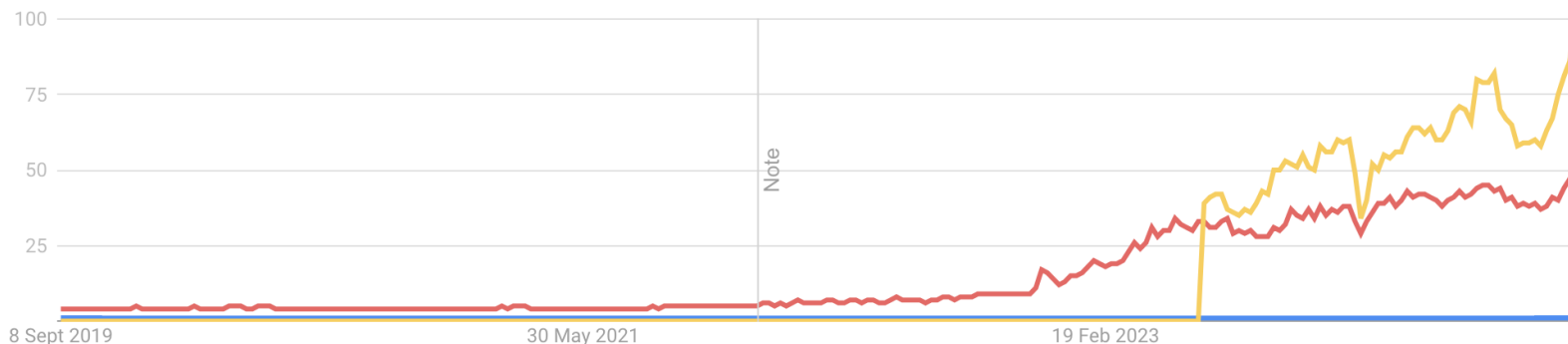
Motivation

Community
Tools

Remark: All images in the following slides credit to Google Images.

“Data Science” on Google Trends

● Data science ● Artificial intelligence ● ChatGPT



Baidu 指数 首页 最新动态 行业排行

关键词 **数据科学** 未被收录, 如要查看相关数据, 购买创建新词的权限后, 您可以添加自己关注的

Baidu 指数 首页 最新动态 行业排行

关键词 **data science** 未被收录, 如要查看, 购买创建新词的权限后, 您可以添加自己关注的

Any observation?

A vibrant blue background featuring the words "DATA SCIENCE" in large, bold, white letters. Surrounding the text are various icons and smaller text elements related to data science, including "ANALYSIS", "VISUALIZATION", "SYSTEM", "PROCESS", "STRUCTURE", "PROGRAMMING", "SOLVING", and "KNOWLEDGE". There are also illustrations of people working on laptops and a pie chart.



Data science combines math and statistics, specialized programming, advanced analytics, AI and ML with specific subject matter expertise to uncover actionable insights hidden in an organization's data.

The Evolution of Data Science



20th century: Statistics and early data analysis methods.



2000s: Rise of big data and machine learning.



Today: AI and deep learning advancements.



Major turning points: cloud computing, IoT, big data, LLM.

Distinct Objective and Focus



Industry

- **Outcome-Driven:** The primary goal is to drive business value, such as increasing profits, and improving customer experience.
- **Practical Application:** Tools and methods are chosen based on their ability to provide quick and effective solutions.
- **Performance Metrics:** Success is often measured in terms of business metrics like return on investment (ROI), customer penetration rates, or KPIs.



Academia

- **Knowledge-Driven:** The focus is often on advancing theoretical understanding and developing new methodologies.
- **Research and Innovation:** Strong emphasis on creating novel, generalizable approaches that can contribute to the broader scientific community.
- **Publication and Peer Review:** Success is typically measured by the quality and quantity of published research, citations, and contributions to academic knowledge.

Data Science in Industry



香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB

From Industry

The **accelerating volume** of data sources, and subsequently data, has made data science one of the fastest-growing fields across every industry.

Organizations are increasingly reliant on data experts to interpret data and provide actionable recommendations to **improve business outcomes**.



MICROSOFT Company size: large Focus areas: data, analytics, AI Unique benefits: variety of roles ecologically focused opportunities	
 Company size: large Focus areas: eCommerce, supply chain, forecasting Unique benefits: real-world problems to solve	AMAZON
AIRBNB Company size: large Focus areas: finance, trust, guest experiences Unique benefits: customer-centric, self-reflective culture	
 Company size: large Focus areas: optimization of chips and graphic cards Unique benefits: multiplicative effect of work, helping other data scientist	NVIDIA
ORACLE Company size: large Focus areas: advertising, designing data-science focused products Unique benefits: large number of open positions	
 Company size: large Focus areas: optimizing customer experience Unique benefits: customer-focused, first of its kind	DELTA
SPLUNK Company size: medium Focus areas: outlier detection, forecasting, event clustering, and predictive analytics Unique benefits: multiplicative effect of work, helping other data scientist	
CLOUDERA Company size: medium Focus areas: data management and their machine learning platform Unique benefits: most powerful and scalable cloud data platform	
NUMERATOR Company size: medium Focus areas: customer insights for ecommerce Unique benefits: client-focused, open-ended question to explore	
 Company size: medium Focus areas: algorithms and data optimization Unique benefits: academic approach to their work and validation	TERADATA
DATABRICKS Company size: medium Focus areas: forecasting, product analysis, churn prediction and insights, segmentations and recommendations Unique benefits: built on open source platforms	

How do Business Embrace Data Science?

Business operation emphasizes data-driven decision-making and investing in infrastructure.

It follows a typical project lifecycle in the industry, which involves moving from problem identification to solution deployment.

When assessing production, it's important to consider KPIs, cost savings, and revenue generation.

Challenges may arise from a need for business understanding, data quality, and privacy concerns.

Data Science Lifecycle

A data science lifecycle is defined as the **iterative set** of data science steps required to deliver a project or analysis.

There are **no one-size-fits** that define data science projects. Hence you need to determine the one that best fits your business requirements.

Each step in the lifecycle should be performed carefully. Any **improper execution** will affect the following step, ultimately impacting the entire process.



However, it might happen...



Data Science Job Market

Data Analyst

Data Engineer

Data Scientist

ML/DL Engineer

Data Product Manager

Chief Data Officer

More...



Hong Kong Exchanges and Clearing Limited (HKEX)



Data Analytics Project Manager - Vice President - Chief Data Office (6 Months Contract) ✓

Hong Kong, Hong Kong SAR · Reposted 2 weeks ago · 45 applicants



Microsoft

Senior Data Scientist ✓

Hong Kong, Hong Kong SAR · 3 days ago



Remote · Full-time



The Hong Kong Jockey Club

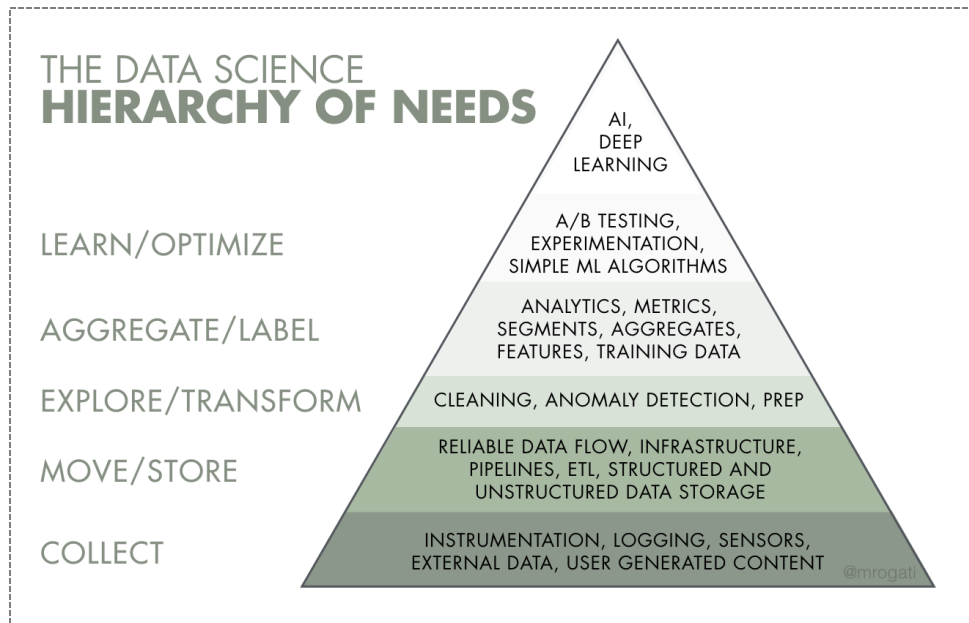
Data Scientist Lead ✓

Hong Kong, Hong Kong SAR · 6 days ago



Full-time · Mid-Senior level

Hierarchy of Needs



For medium/large company:

➔ Machine Learning Engineer

➔ Data Scientist
Data Analyst

➔ Data Engineer

➔ Data Infrastructure
Engineer

What about a small company or even a start-up? **Everything!**

Voice: The 3 biggest problems with Data Science

Cassie Kozyrkov, Google (former) Chief Decision Scientist



Link: <https://www.youtube.com/watch?v=kjzDo5f7Ecc>

Quick Recap: What are the problems?



MAKE: Who makes data useful? A well-organized company structure for data scientists and other roles.



DATA: Data should be usable then can be useful. The data engineering team should process large-scale of data at first, then the data science team can help with it.



USEFUL: Who makes the decision? Data science team needs skilled decision maker and business leader.

Tips when looking for a job:

- ✓ Carefully go through the job post and check if the job duties are clear.
- ✓ Think about the team's organization and leadership.
- ✓ Dare to try new stuff and learn by doing.

Data Science in Academia

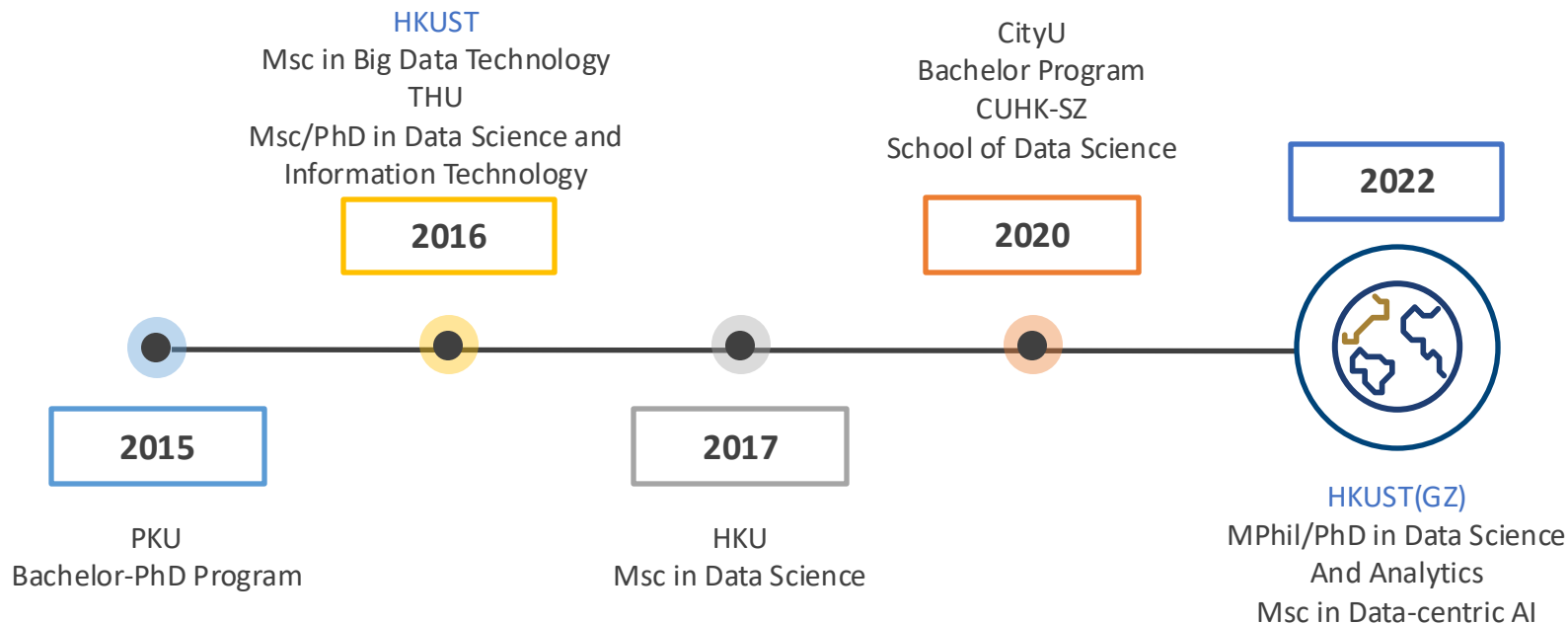


香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB

Growing demand for Data Science Education

Mainland China & Hong Kong



Data Science Research

Top conferences:

KDD, SIGIR, ICDE, SIGMOD,
WSDM, VLDB, ACL, ICML ...

Trending topics:

Good data for AI (a.k.a. data-centric AI)

AI for Good data (preparation & cleaning)

Data visualization

Harness LLM

More ...

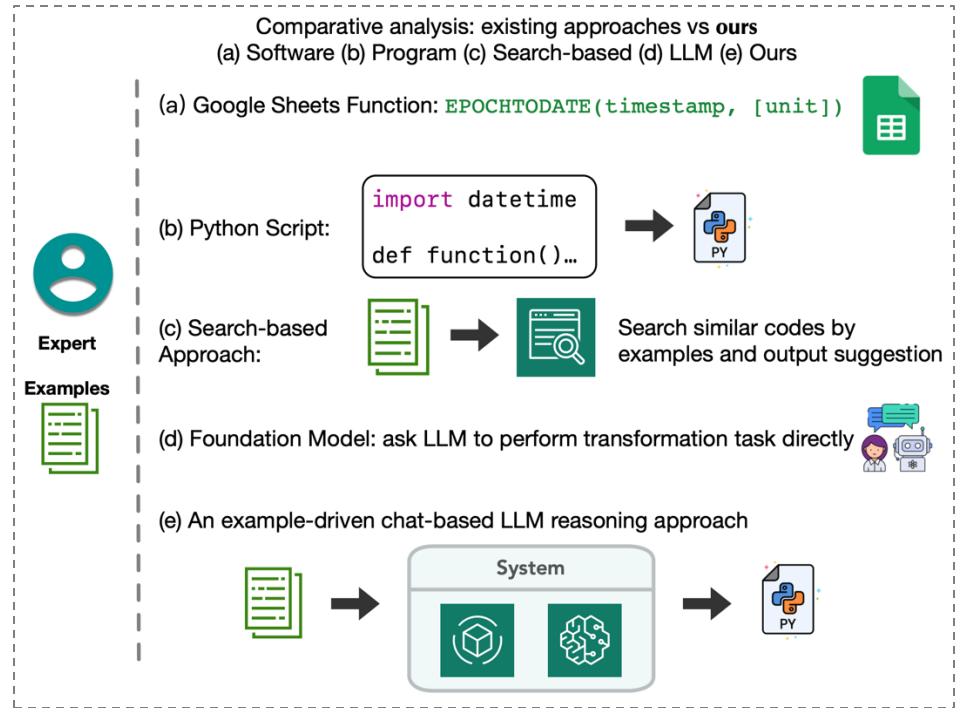


My Research

Focus on chat-based data preparation solutions using the Large Language Models.

Below are some examples:

- LLM-based Missing value imputation
- Chat-based EDA
- Data Transformation 



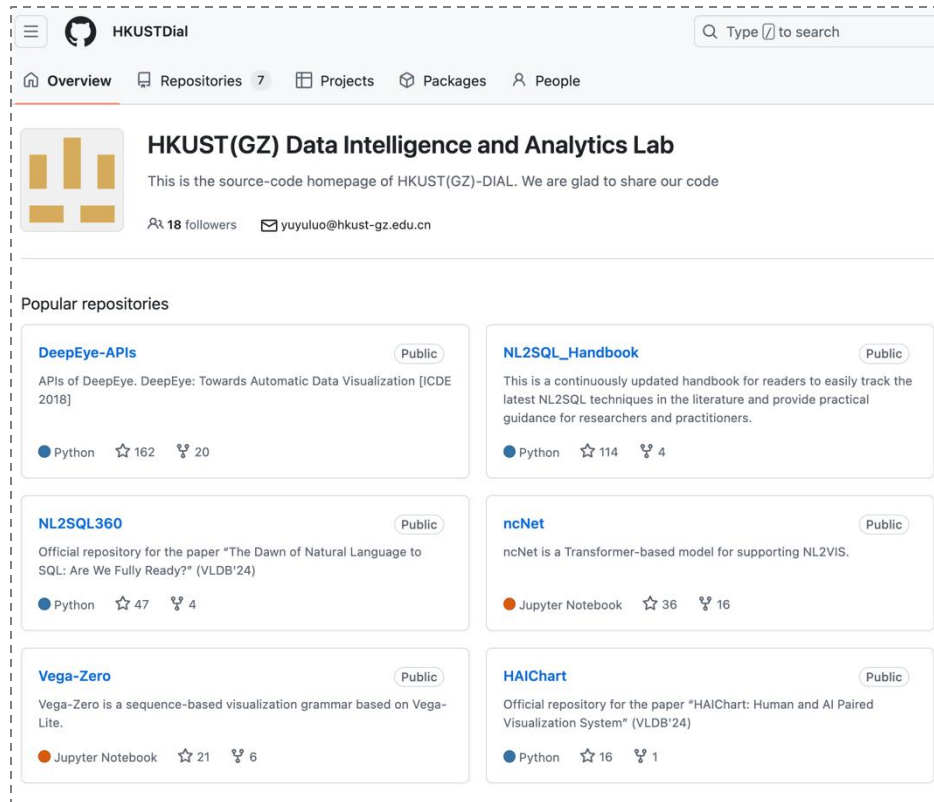
Our Lab Research

Data Intelligence and Analytics Lab, DIAL@HKUST(GZ)

We aim to reduce overhead in the world of data science and lower the human cost, such as:

- NL2SQL
- Table retrieval and analysis
- Data Preparation
- RAG meets LLM
- Visualization

<https://github.com/HKUSTDial>



Motivation: Why I Choose DSA?

Bottlenecks occurred at work

Accumulated long TODO list
but seldom handled

Dealing with clients
occupied the major
schedule

Human resource shortage,
especially developer



Transition to academia

Professional supervision
and self-paced learning

Collect problems from the
industry

Explore solutions by doing
my research

GET HANDS DIRTY

Community

kaggle

stackoverflow



Online Courses

TOP 10 FREE ONLINE COURSES WEBSITES



<https://opportunitiespedia.com>

Tool & Library





Thank You!



香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB