



Vietnam National University of HCMC
International University
School of Computer Science and Engineering



Data Analysis **(IT137)**

Nguyen Trung Ky, PhD

✉ ntky@hcmiu.edu.vn

Basic Information about course

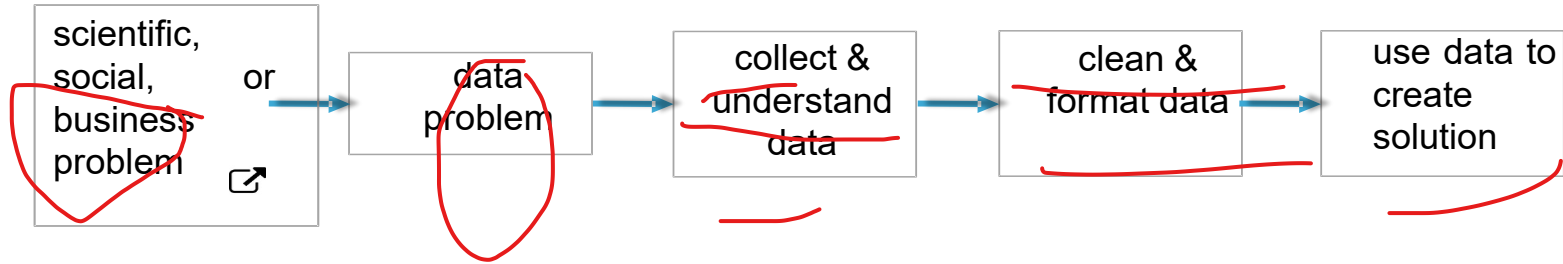


- Instructor: Dr. Nguyen Trung Ky.
 - Ph.D. Grenoble Alpes University 2019; second year at IU.
 - Research on Computational Linguistics (Natural Language Processing, Natural Language Generation) and Machine Learning.
 - Office: O1.610
 - Ask immediately after class or by appointment via email ntky@hcmiu.edu.vn
- Every Friday, 10:35 - 13:05 from 22/09/2023 - 14/01/2024.
- Previous course: Intro to data science
- Course credit: 4
 - Lecture: 3 (from 22/09/2023 - 12/01/2024)
 - Laboratory: 1
 - (Group 1 from 05/10/2023 - 21/12/2023)
 - (Group 2 from 07/10/2023 - 23/12/2023)

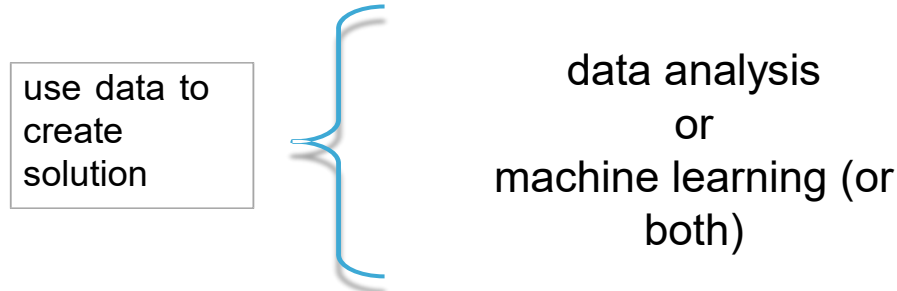
WHAT IS DATA SCIENCE?



...solving problems with data...



...which step is most challenging?



WHAT IS DATA ANALYSIS?

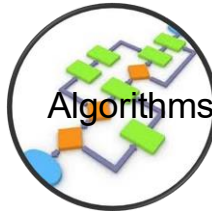
...using data to discover useful information...



- data: anything you can **measure** or **record**



- statistics: summarize (and visualize) **main characteristics** of the data



- algorithms: apply algorithms to find **patterns** in the data

What types of jobs related data in job market today?



Data Analyst

"The Astronaut"

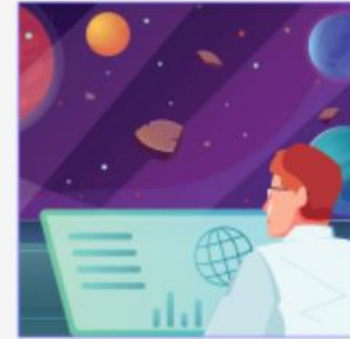
Better, faster
decision-making



Analytics Engineer

"Mission Control"

Better data for
decision-making



Data Scientist

"The Astrophysicist"

R&D on new
capabilities



Data Engineer

"The Aerospace Engineer"

Build data syncs
and data models



What skills and tools are needed?

Data Scientist

"The Astrophysicist"

MISSION: R&D on new analytics capabilities

SKILLS

- AI/ML
- Statistical analysis
- Research

TOOLS: Python, R

RESPONSIBILITIES

- Build forecasting and other predictive models
- Detect anomalies and outliers
- Cluster look-alikes
- Research and test new AI/ML techniques



SAVANT

Data Analyst

"The Astronaut"

MISSION: Better, faster decision-making

SKILLS

- Creative problem solving
- Domain knowledge (ex. marketing)
- Quantitative analysis
- Compelling communication

TOOLS: Spreadsheet, SQL, BI

RESPONSIBILITIES

- Inform and guide decision-makers
- Build dashboards and analyses
- Translate business needs into data
- Focus on datasets < 10M rows



SAVANT

More on job of data scientist



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

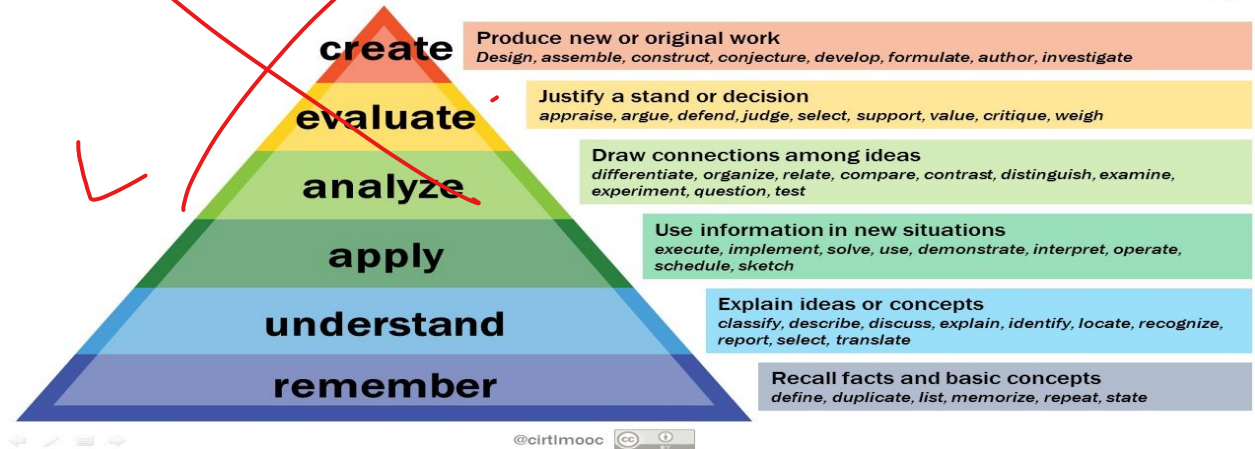
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



Learning outcomes

1. Understand fundamental concepts of data analysis.
2. Explain how to perform data analysis with **descriptive statistics** and **inferential statistics**.
3. Apply data analysis techniques and tools to some practical cases in business/engineering.

Bloom's Taxonomy



Topics to be covered in this course



Week 1	Course Overview
Week 2	Basic of R
Week 3	Data types & wrangling
Week 4	Data types & wrangling (continue)
Week 5	Summary statistics
Week 6	Summary statistics (continue)
Week 7	Data Plotting
Week 8	Data Plotting (continue)

Topics to be covered in this course



Week 9	Probability Basics
Week 10	Models & parameter inference
Week 11	Hypothesis testing
Week 12	Hypothesis testing (continue)
Week 13	Model comparison
Week 14	Linear regression
Week 15	Linear regression (continue)

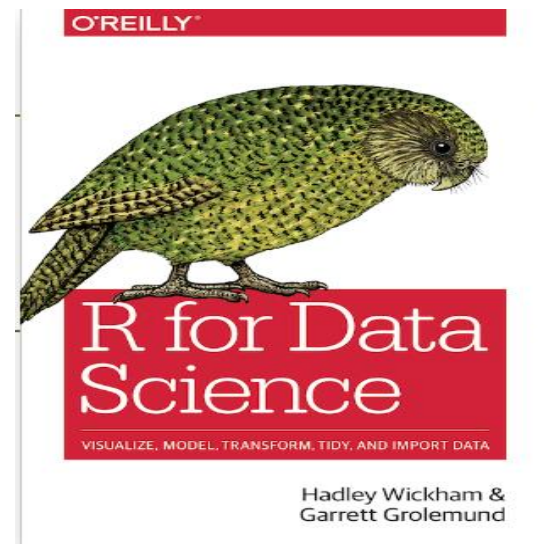
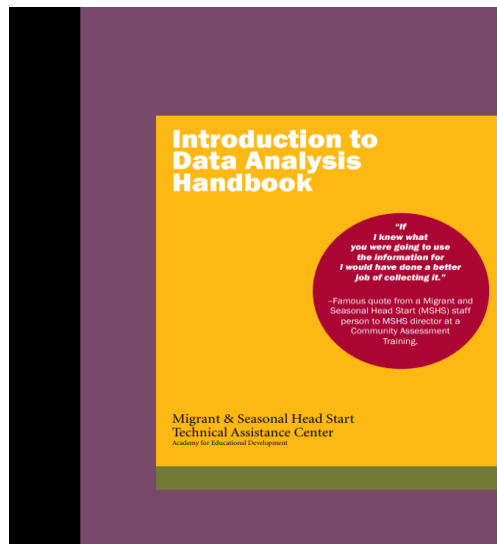
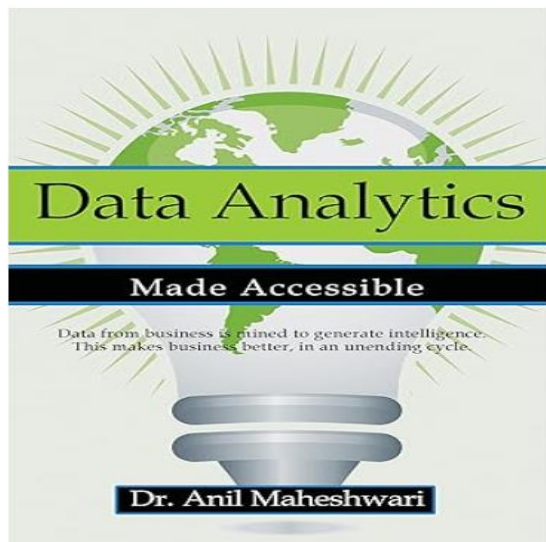
Materials/Books



[1] . Anil Maheshwari, **Data Analytics**, 2022

[2]. Migrant & Seasonal Head Start Technical Assistance Center. **Introduction to Data Analysis Handbook**, non-commercial uses only.

[3]. Hadley Wickham & Garret Grolemund, **R for Data Science**. O'reilly 2023.



Some useful websites for R

- [1] . <https://www.w3schools.com/r/>
- [2]. <https://www.tutorialspoint.com/r/index.htm>
- [3]. <https://www.r-bloggers.com/2021/04/tidyverse-in-r-complete-tutorial/>
- [4]. <https://www.datacamp.com/tutorial/tidyverse-tutorial-r>

Why we choose to learn R for Data Analysis?



Applications of R



Why we choose to learn R for Data Analysis?



Companies that use R for Analytics



Why we choose to learn R for Data Analysis?

1. Who Uses R? Companies That Use R and What R Is Used For

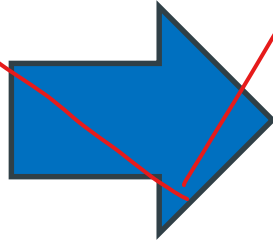
<https://careerkarma.com/blog/who-uses-r/>

2. Why Top Companies are using R Programming

<https://data-flair.training/blogs/r-applications/>

Why we choose to learn R for Data Analysis?

Please share your experiences on R Language using code **6118 5657** on **menti.com** or the following QR code



- Course information, announcements
 - [IT137IU_1_2023-2401: Data Analysis S1 2023-24 G01](#)(KyNguyen)
- Upload lectures, quizzes or homework

Grading policies

1. Quizzes + Lab Assignments or Project : 30%
2. Midterm: 30%
3. Final: 40%



Thank you for your listening!