

<div style="display: flex; flex-direction: column; align-items: center;"> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>D</span></div> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>I</span></div> </div>	<b>PHÒNG THÍ NGHIỆM THỰC HÀNH ỨNG DỤNG</b> <b>KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO</b> Applied Data Science and Artificial Intelligence Lab
--	---



## LẬP TRÌNH PYTHON CHO KHOA HỌC DỮ LIỆU

### PYTHON FOR DATA SCIENCE

Hà Nội ★ 11.2020

Nothing is faster than Innovation together

1

<div style="display: flex; flex-direction: column; align-items: center;"> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>D</span></div> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>I</span></div> </div>	<h2>Thông tin giảng viên</h2>	
--	-------------------------------	--



**TS. TẠ QUANG CHIẾU**

**Vị trí & Đào tạo**

- Trưởng bộ môn Hệ thống thông tin và Tri thức, Khoa Công nghệ Thông tin
- PhD & Post-doc in Polytech Tours, France

**Lĩnh vực giảng dạy, nghiên cứu**

- Big data, Data science, machine learning, AI
- Các phương pháp cho bài toán lập kế hoạch

**Contact:**

(E): [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com)

(M): 0913 522 275

Tạ Quang Chiếu – (E): [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – (M): 0913 522 275

2/78

2

<div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>D</span> </div> <div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>I</span> </div>	<h2 style="margin: 0;">Objectives</h2>
<ul style="list-style-type: none"> <li>• Hiểu được tầm quan trọng của Khoa học dữ liệu</li> <li>• Vận dụng được các bước trong quy trình thực hiện một dự án về khoa học dữ liệu.</li> <li>• Kiến thức, kỹ năng cần thiết để trở thành một nhà khoa học dữ liệu</li> <li>• Áp dụng được Python và các thư viện phổ biến trong giải quyết một số bài toán cơ bản của Khoa học dữ liệu.</li> </ul>	
<div style="display: flex; justify-content: space-between;"> <span>Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</span> <span style="background-color: #000080; color: white; padding: 2px 5px;">3/78</span> </div>	

3

<div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>D</span> </div> <div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>I</span> </div>	<h2 style="margin: 0;">References</h2>
<ul style="list-style-type: none"> <li>[1]. Introducing Data Science</li> <li>[2]. Python Data Science Handbook</li> <li>[3]. Data Science from Scratch</li> <li>[4]. Python crash course</li> <li>[5]. Bài giảng Khoa học dữ liệu</li> </ul>	
<div style="display: flex; justify-content: space-between;"> <span>Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</span> <span style="background-color: #000080; color: white; padding: 2px 5px;">4/78</span> </div>	

4

AD

AI

## PHẦN 2: NHẬP MÔN KHOA HỌC DỮ LIỆU

### FUNDAMENTALS OF DATA SCIENCE

### CONTENT

1

CHAPTER 1: INTRODUCTION

2

CHAPTER 2: PYTHON LIBRARIES FOR DATA SCIENCE

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

5/112

5

AD

AI

## Chapter 1: Introduction

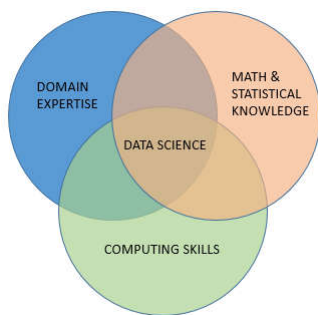
### 1.1 Data Science in Practice

### 1.2 What is Data Science?

### 1.3 The big data and data science

### 1.4 Process of data science project

### 1.5 Data Scientist Knowledge and Skills



The diagram consists of three overlapping circles. The top-left circle is blue and labeled 'DOMAIN EXPERTISE'. The top-right circle is orange and labeled 'MATH & STATISTICAL KNOWLEDGE'. The bottom circle is green and labeled 'COMPUTING SKILLS'. The central area where all three circles overlap is labeled 'DATA SCIENCE'.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

6/78

6

A

D

Content of chapter 1



## “DATA IS THE NEW GOLD”

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

7/78

7

A

D

1.1 Data Science in Practice

### WHO USES DATA SCIENCE?



Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

8/78

8

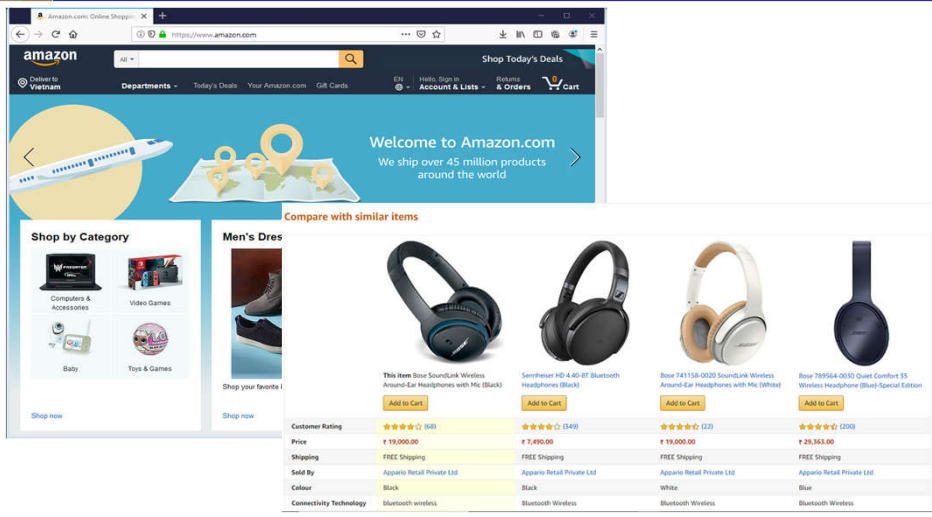


A

D

I

## E-commerce: Amazon



**Compare with similar items**

Product	Customer Rating	Price	Shipping	Sold By	Colour	Connectivity Technology
This item Bose SoundLink Wireless Around-Ear Headphones with Mic (Black)	★★★★☆ (164)	₹ 19,000.00	FREE Shipping	Appario Retail Private Ltd	Black	Bluetooth wireless
Sennheiser HD 440-BT Bluetooth Headphones (Black)	★★★★☆ (149)	₹ 7,490.00	FREE Shipping	Appario Retail Private Ltd	Black	Bluetooth Wireless
Bose 741158-0020 SoundLink Wireless Around-Ear Headphones with Mic (Silver)	★★★★☆ (123)	₹ 19,000.00	FREE Shipping	Appario Retail Private Ltd	White	Bluetooth Wireless
Bose 789564-0030 Quiet Comfort 35 Wireless Headphone (Black-Special Edition)	★★★★☆ (108)	₹ 29,363.00	FREE Shipping	Appario Retail Private Ltd	Blue	Bluetooth Wireless

**Khi tìm một sản phẩm, Amazon gợi ý cho ta một loạt sản phẩm liên quan như thế nào? So sánh giá? Nhận diện màu sắc....?**

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

11/78

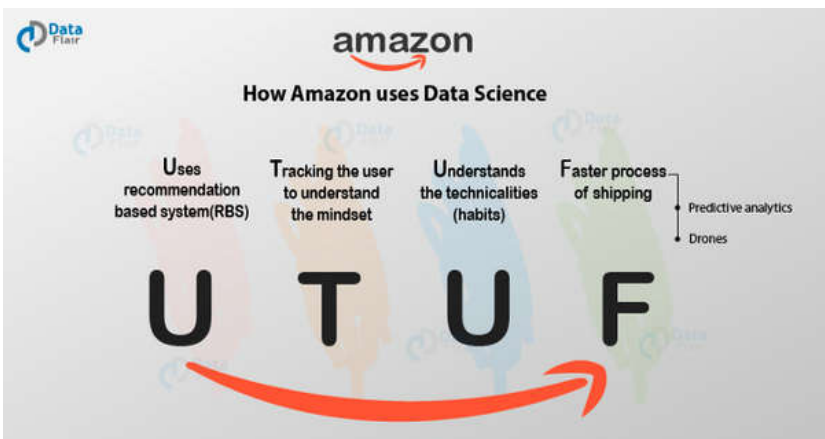
11

A

D

I

## Amazon



**How Amazon uses Data Science**

Uses recommendation based system(RBS)    Tracking the user to understand the mindset    Understands the technicalities (habits)    Faster process of shipping

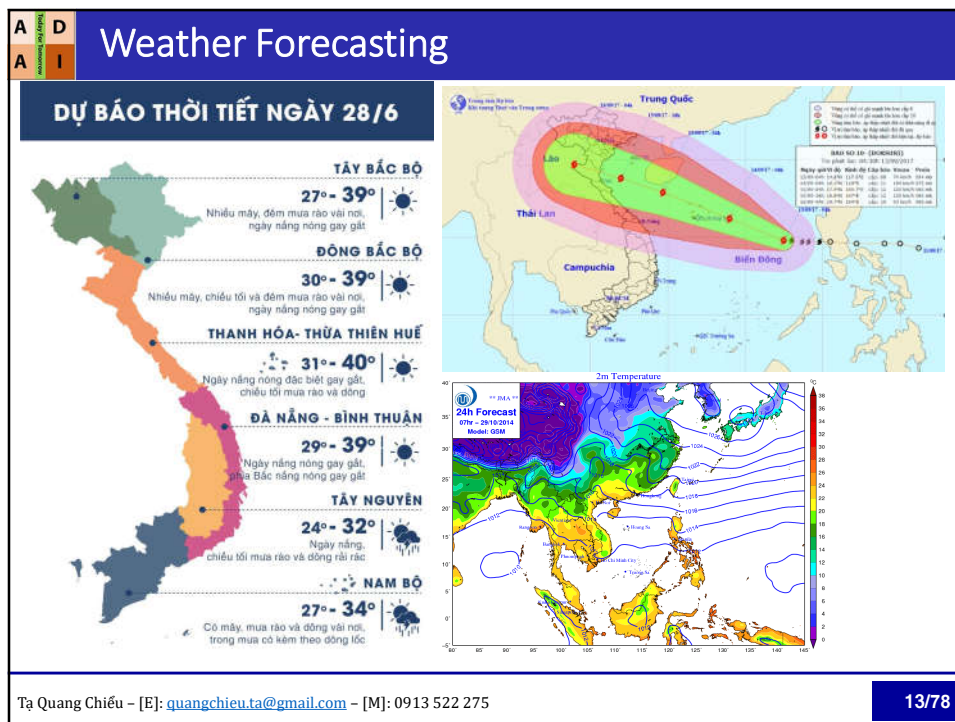
UTUF

Predictive analytics  
Drones

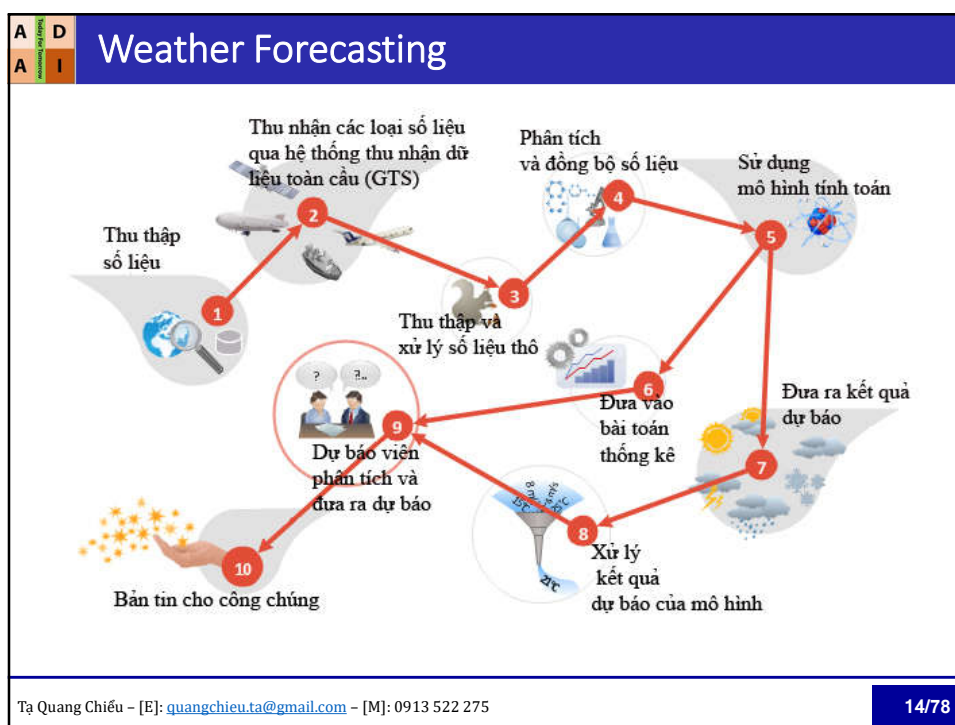
Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

12/78

12

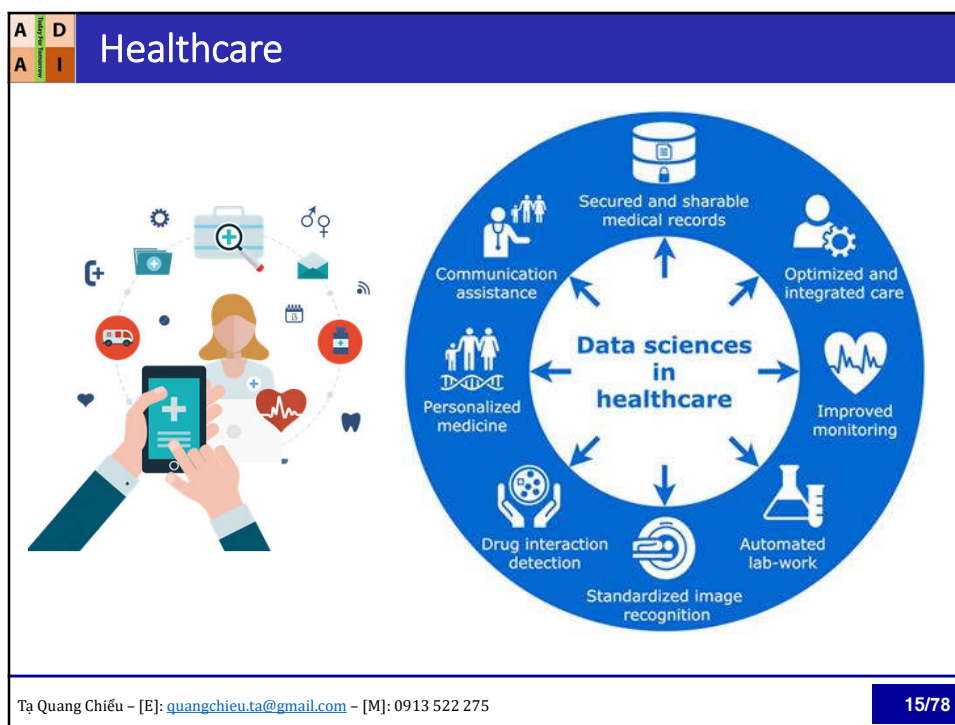


13

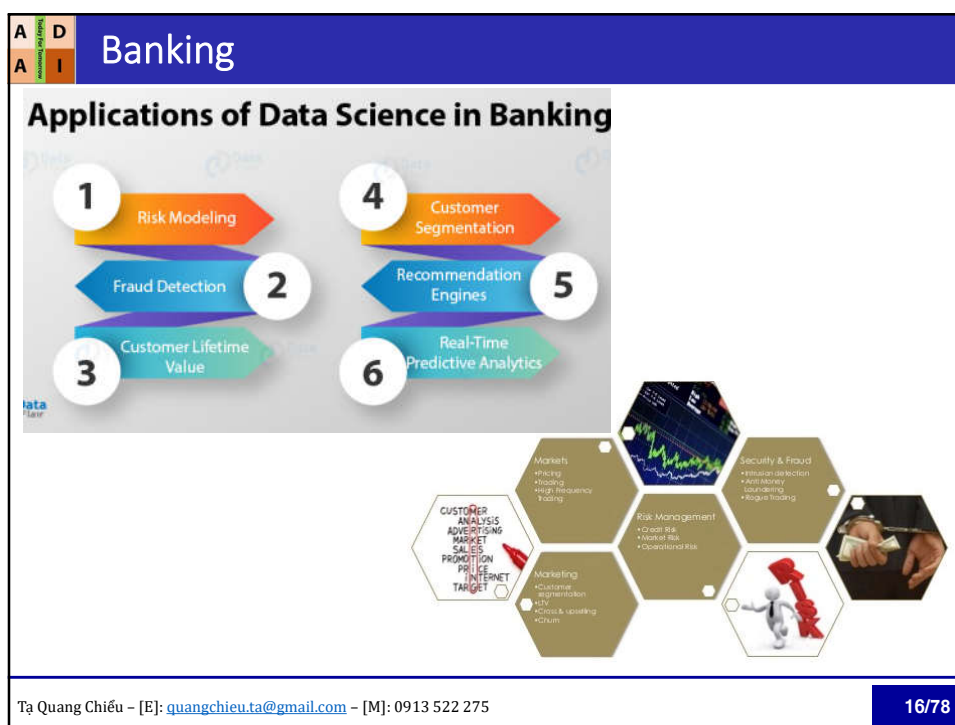


14





15



16



A
D

A
I

Airlines

Using Data Science, we can achieve the following:

- Route Planning: Whether to schedule direct or connecting flights
- Predictive analytics model can be built to foresee flight delays
- Promotional offers depending on customer booking patterns
- Deciding which class of planes to purchase for better performance

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

17/78

17

A
D

A
I

Transport

Logistics companies like FedEx are using Data Science models for operational efficiency

- Discover the best routes to ship
- The best suited time to deliver
- The best mode of transport

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

18/78


18


A
D

A
I

## 1.1 Data Science in Practice

- Các công ty thương mại sử dụng khoa học dữ liệu để hiểu rõ hơn về khách hàng, quy trình, nhân viên, hoàn thiện và sản phẩm của họ.
- Các tổ chức tài chính sử dụng khoa học dữ liệu để dự đoán thị trường chứng khoán, xác định rủi ro cho vay tiền và tìm hiểu cách thu hút các nhân viên mới cho dịch vụ của họ.
- Các tổ chức chính phủ sử dụng khoa học dữ liệu để khai phá thông tin, phát hiện các hoạt động tội phạm, tối ưu hóa các hoạt động.
- Các tổ chức sử dụng Khoa học dữ liệu để tăng hiệu quả các dự án phát triển gây quỹ, bảo vệ tài nguyên.
- Các trường đại học sử dụng khoa học dữ liệu trong nghiên cứu, phát triển và đánh giá các lớp học trực tuyến.






Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

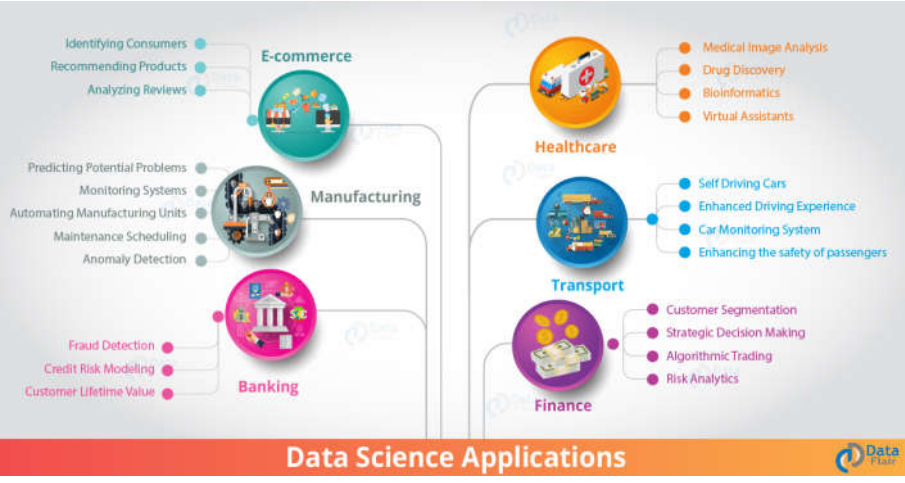
19/78

19

A
D

A
I

## 1.1 Data Science in Practice



**Data Science Applications**

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275


20/78

20

AD

AI

## 1.2. What is Data Science



# Data Science

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

21/78

21


AD

AI


## Data – Information - Knowledge

- Dữ liệu **Data**: là các yếu tố thô, chưa được xử lý, bao gồm: văn bản, số liệu, ký hiệu, hình ảnh, âm thanh,...
- Thông tin **Information** là dữ liệu đã được xử lý để đáp ứng yêu cầu của người dùng
- Tri thức/kiến thức **Knowledge**: bao gồm những dữ kiện, thông tin, sự mô tả hay kỹ năng có được nhờ trải nghiệm hay thông qua giáo dục.

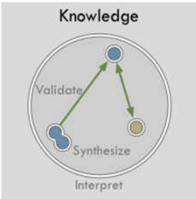
**Data**



**Information**



**Knowledge**



Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

22/78

22

A D  
A I

## 1.2. What is Data Science?

Chưa có một định nghĩa về Khoa học dữ liệu (Data Science) được tất cả mọi người chấp nhận.

**Một vài định nghĩa:**

**Viện Tiêu chuẩn và Công nghệ Quốc gia Mỹ (National Institute of Standards and Technology - NIST):**

- "Data science is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing"
- Khoa học dữ liệu là trực tiếp trích rút tri thức hành động từ dữ liệu qua quá trình phát hiện, thiết lập và kiểm nghiệm các giả thiết.

**Microsoft:**

- "Data science is about using data to make decisions that drive actions"
- Khoa học dữ liệu là sử dụng dữ liệu tạo ra các quyết định dẫn dắt hành động.

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

23/78

23

A D  
A I

## 1.2. What is Data Science?

- Khoa học dữ liệu là khoa học về quản trị, phân tích dữ liệu để tìm ra các hiểu biết, các tri thức, hành động, các quyết định để dẫn đến hành động

**Nhiệm vụ của DS:**

- Thu thập để xử lý dữ liệu để tìm ra những insight giá trị.
- Giải thích, trình bày những insight cho các bên liên quan để chuyển hóa các insight thành hành động

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

24/78


24

A
D


A
I


## 1.2. What is Data Science?


Công cụ mới và hiệu quả trong khám phá **thông tin ẩn chứa** từ dữ liệu



Phương pháp tự **động phân tích** và **trích xuất** thông tin từ **khối lượng dữ liệu lớn**







Một lĩnh vực mới kết hợp giữa **thống kê, toán học, kỹ thuật lập trình, mô phỏng** và **biểu diễn** nhằm chuyển đổi **dữ liệu thành thông tin**

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

25/78


25

A
D

A
I

## 1.2. What is Data Science?

- **Khoa học dữ liệu** ≠ **Khoa học thông thường** ở quan điểm: Tìm tri thức (**insights**) từ dữ liệu (dẫn dắt bởi dữ liệu - “data-driven”)
  - Rút ra tri thức bằng việc tìm tòi, khám phá từ dữ liệu chứ không nhất thiết phải chứng minh nó.
  - Tri thức tìm ra phải có tính ổn định (luôn có cùng kết quả nếu sử dụng cùng một phương pháp).



Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

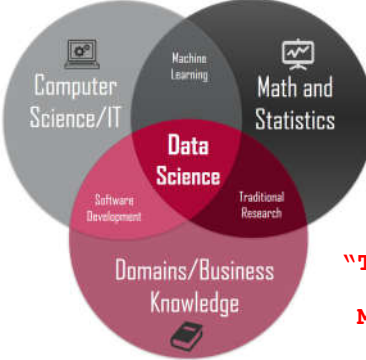
26/78

26

A
D


A
I

## 1.2. What is Data Science?



**In God we trust, all others bring data.**

-William E. Deming



**"Ta chỉ tin vào thượng đế,  
Mọi thứ khác phải dựa vào dữ liệu"**

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

27/78

27

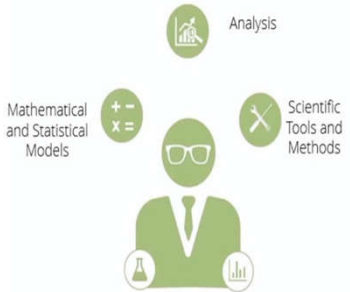
A
D

A
I


## Components of Data Science

- Khoa học dữ liệu là sự kết hợp của **Miền chuyên môn** (Domain Expertise) và **Các phương pháp khoa học** (Scientific Methods) với **Công nghệ** (Technology)

Domain Expertise and Scientific Methods



Technology



**Data Science**

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

28/78

28

A
D

A
I

## Khoa học dữ liệu cần cho?

Đưa ra các quyết định cái nào tốt hơn:  
Giữa A và B?

Phân tích đưa ra các dự đoán:  
Chuyện gì sẽ xảy ra tiếp theo?

Nhận dạng mẫu: Có bất kỳ thông tin ẩn  
quan trọng nào trong mẫu không?

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

29/78

29

A
D

A
I

## Ví dụ

Want to buy online furniture?

```

graph TD
    Q1{Does website sell furniture?} -- YES --> Q2{Rating > 4 out of 5}
    Q1 -- NO --> A1[Close website]
    Q2 -- NO --> A2[Close website]
    Q2 -- YES --> Q3{Discount > 20%}
    Q3 -- NO --> A3[Close website]
    Q3 -- YES --> A4[Purchase Product]
    
```

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

30/78

30



**1.3. The big data and data science**

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

31/78

31

**What is Big Data?**

The "three V's", i.e. the Volume, Variety and Velocity of the data coming in is what creates the challenge.

**VOLUME**

> 3,500 NORTH AMERICA  
> 2,000 EUROPE  
> 250 CHINA  
> 200 MIDDLE EAST  
> 50 LATIN AMERICA  
> 50 INDIA  
> 400 JAPAN

**VARIETY**

PEOPLE TO PEOPLE: ARCHIVES, MEDICAL RECORDS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...

PEOPLE TO MACHINE: ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...

MACHINE TO MACHINE: SENSORS, GPS DEVICES, SIM CARD SCANNERS, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...

**VELOCITY**

2.9 MILLION PHOTOS SENT EVERY SECOND  
20 HOURS OF VIDEO UPLOADED EVERY MIN  
50 MILLION TEXTS PER DAY

**VALUE**

Industry	Productivity Increase	Sales Increase
RETAIL	49%	\$9.6B
CONSULTING	30%	\$5.0B
AIR TRANSPORTATION	21%	\$4.2B
CONSTRUCTION	20%	\$4.2B
FOOD PRODUCTS	20%	\$7
STEEL	20%	\$7
AUTOMOBILE	19%	\$2B
INDUSTRIAL INSTRUMENTS	18%	\$1.2B
PUBLISHING	18%	\$0.8B
TELECOMMUNICATIONS	17%	\$0.4B

**CASE STUDY - Healthcare**

165B CLINICAL  
\$9B PUBLIC HEALTH  
\$108B R&D  
\$47B ACCOUNTS  
\$5B BUSINESS MODEL

40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR

5% PROJECTED GROWTH IN GLOBAL IT SPENDING PER YEAR

**Dữ liệu lớn (Big Data) là thuật ngữ sử dụng cho tập hợp các dữ liệu quá lớn và phức tạp khiến cho việc xử lý các dữ liệu này trở nên khó khăn khi sử dụng các kỹ thuật quản lý dữ liệu truyền thống.**

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

32/78

32

A
D

A
I

## Characteristics Of Big Data:

- Khối lượng (**V**olume): lượng dữ liệu được tạo ra
- Tốc độ (**V**elocity): Tốc độ dữ liệu được tạo ra và tốc độ chuyển đổi dữ liệu.
- Đa dạng (**V**ariety): Các kiểu dữ liệu được sử dụng
- Độ chính xác (**V**eracity): Độ tin cậy của dữ liệu
- Giá trị (**V**alue): Giá trị của dữ liệu

(Gartner.2014)

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

33/78

33

A
D

A
I

## 1.3. The big data and data science

**trên Internet diễn ra những gì?**

**60 giây**

- 1 triệu lượt đăng nhập facebook
- 87.500 người viết tweet
- 342.000 lượt tải ứng dụng Google Play
- 347.222 người dùng lượt Instagram
- 694.444 giờ xem NETFLIX
- 996.956 USD Tiêu dùng trực tuyến
- 41,6 triệu tin nhắn được gửi đi từ Facebook Messenger và Whatsapp
- 18,1 triệu tin nhắn được gửi đi
- 4,5 triệu lượt xem YouTube
- 3,8 triệu lượt tìm kiếm Google
- 2,1 triệu snap được tạo ra

Tweet: Mẩu tin đăng ở trang cá nhân trên Twitter  
Snap: Hình ảnh hoặc video được ghi lại bằng Snapchat  
Nguồn: Visual Capitalist  
<https://infographics.vn>

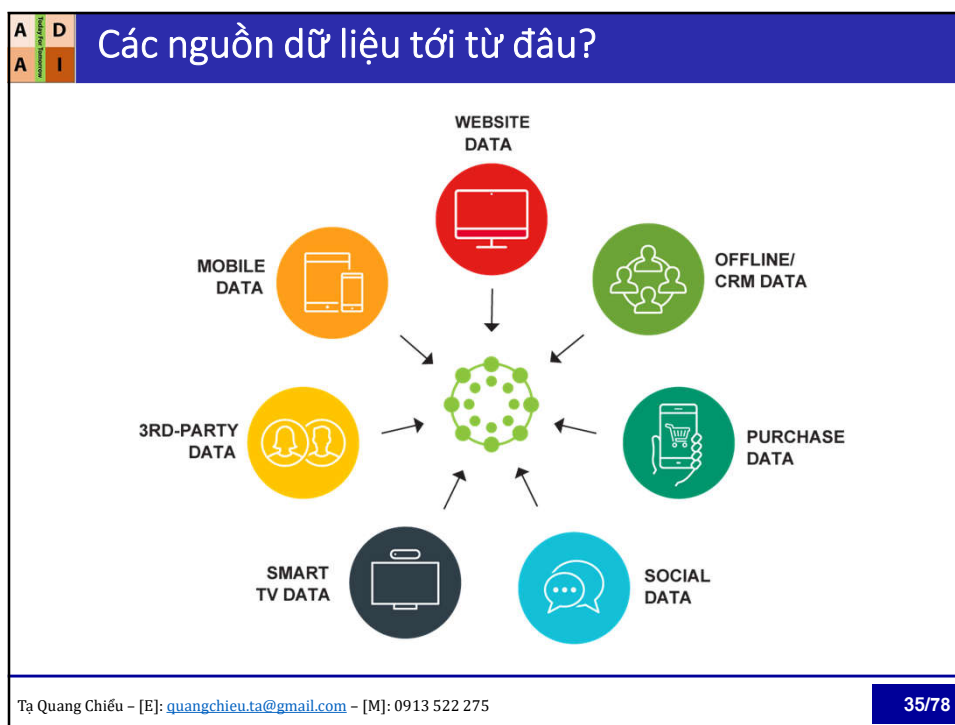
**60 SEC**

© TTXVN

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

34/78

34



35

**Data type**

- Các kiểu dữ liệu trong khoa học dữ liệu**
  - Dữ liệu có cấu trúc (Structured)**
    - Dữ liệu phụ thuộc vào mô hình dữ liệu và nằm trong một trường cố định trong một bản ghi.
    - Lưu trữ trong cơ sở dữ liệu.

Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Int
1	214390830	Total (Age-adjusted)	2008	74.6%	73.8%
2	214390833	Aged 18-44 years	2008	59.4%	58.0%
3	214390831	Aged 18-24 years	2008	37.4%	34.6%
4	214390832	Aged 25-44 years	2008	66.9%	65.5%
5	214390836	Aged 45-64 years	2008	88.6%	87.7%
6	214390834	Aged 45-54 years	2008	86.3%	85.1%
7	214390835	Aged 55-64 years	2008	91.5%	90.4%
8	214390840	Aged 65 years and over	2008	94.6%	93.8%
9	214390837	Aged 65-74 years	2008	93.6%	92.4%
10	214390838	Aged 75-84 years	2008	95.6%	94.4%
11	214390839	Aged 85 years and over	2008	96.0%	94.0%
12	214390841	Male (Age-adjusted)	2008	72.2%	71.1%
13	214390842	Female (Age-adjusted)	2008	76.8%	75.9%
14	214390843	White only (Age-adjusted)	2008	73.8%	72.9%
15	214390844	Black or African American only (Age-adjusted)	2008	77.0%	75.0%
16	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%	57.1%
17	214390846	Asian only (Age-adjusted)	2008	80.5%	77.7%
18	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU	
19	214390848	2 or more races (Age-adjusted)	2008	75.6%	69.6%

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

36/78

36

A
D
A
I
Data type

- Các kiểu dữ liệu trong khoa học dữ liệu
  - Dữ liệu phi cấu trúc (Unstructured)
    - Dữ liệu không phụ thuộc vào mô hình dữ liệu vì nội dung theo các ngữ cảnh, cách thức và ngôn ngữ khác nhau.
    - Email.

← → Delete Move Spam ↑ ↓ ×

• New team of UI engineers

• CDA@engineer.com  
To xyz@program.com Today 10:21 ★

An investment banking client of mine has had the go ahead to build a new team of UI engineers to work on various areas of a cutting-edge single-dealer trading platform.

They will be recruiting at all levels and paying between 40k & 85k (+ all the usual benefits of the banking world). I understand you may not be looking. I also understand you may be a contractor. Of the last 3 hires they brought into the team, two were contractors of 10 years who I honestly thought would never turn to what they considered "the dark side."

This is a genuine opportunity to work in an environment that's built up for best in industry and allows you to gain commercial experience with all the latest tools, tech, and processes.

There is more information below. I appreciate the spec is rather loose – They are not looking for specialists in Angular / Node / Backbone or any of the other buzz words in particular, rather an "engineer" who can wear many hats and is in touch with current tech & tinkers in their own time.

For more information and a confidential chat, please drop me a reply email. Appreciate you may not have an updated CV, but if you do that would be handy to have a look through if you don't mind sending.

← Reply → Reply to All → Forward

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
37/78

37

A
D
A
I
Data type

- Các kiểu dữ liệu trong khoa học dữ liệu
  - Dữ liệu thiết lập từ máy tính (Machine-generated)
    - Thông tin được thiết lập tự động bởi máy tính, các quá trình, ứng dụng hoặc các loại máy móc khác mà không cần sự can thiệp của con người.
    - Nguồn thiết lập dữ liệu chính (The Internet of Things).
    - Yêu cầu các công cụ và kỹ thuật xử lý mạnh mẽ.
    - Nhật ký máy chủ web, bản ghi chi tiết cuộc gọi, nhật ký sự kiện mạng.

CSIPERF:TXCOMMIT:313236 2014-11-28 11:36:13, Info 69), objectname [6]"(null)" 2014-11-28 11:36:13, Info result 0x00000000, handle @0x4e54 2014-11-28 11:36:13, Info Beginning NT transaction commit... 2014-11-28 11:36:13, Info trace: CSIPERF:TXCOMMIT:273983 2014-11-28 11:36:13, Info 70), objectname [6]"(null)" 2014-11-28 11:36:13, Info result 0x00000000, handle @0x4e5c 2014-11-28 11:36:13, Info Beginning NT transaction commit... 2014-11-28 11:36:14, Info trace: CSIPERF:TXCOMMIT:386259 2014-11-28 11:36:14, Info 71), objectname [6]"(null)" 2014-11-28 11:36:14, Info result 0x00000000, handle @0x4e5c 2014-11-28 11:36:14, Info Beginning NT transaction commit... 2014-11-28 11:36:14, Info trace: CSIPERF:TXCOMMIT:375581	CSI 00000153 Creating NT transaction (seq CSI 00000154 Created NT transaction (seq 69) CSI 00000155@2014/11/28:10:36:13.471 CSI 00000156@2014/11/28:10:36:13.705 CSI perf CSI 00000157 Creating NT transaction (seq CSI 00000158 Created NT transaction (seq 70) CSI 00000159@2014/11/28:10:36:13.764 CSI 0000015a@2014/11/28:10:36:14.094 CSI perf CSI 0000015b Creating NT transaction (seq CSI 0000015c Created NT transaction (seq 71) CSI 0000015d@2014/11/28:10:36:14.106 CSI 0000015e@2014/11/28:10:36:14.428 CSI perf
--	--

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
38/78

38

A

A

D

I

## Data type

- **Các kiểu dữ liệu trong khoa học dữ liệu**
  - **Dữ liệu dạng đồ thị hoặc mạng lưới (Graph-based or Network data)**
    - Dữ liệu phụ tập trung vào mối quan hệ hoặc tính phụ thuộc của các đối tượng.
    - Sử dụng các nút, cạnh và thuộc tính để biểu diễn và lưu trữ dữ liệu đồ thị.
    - Dữ liệu dựa trên đồ thị là một cách tự nhiên để thể hiện các mạng xã hội và cấu trúc của nó. **LinkedIn Twitter Facebook**

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

39/78

39

A


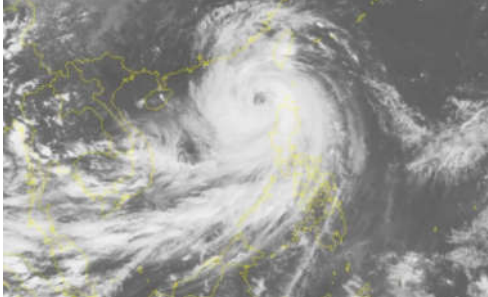
A

D

I

## Data type

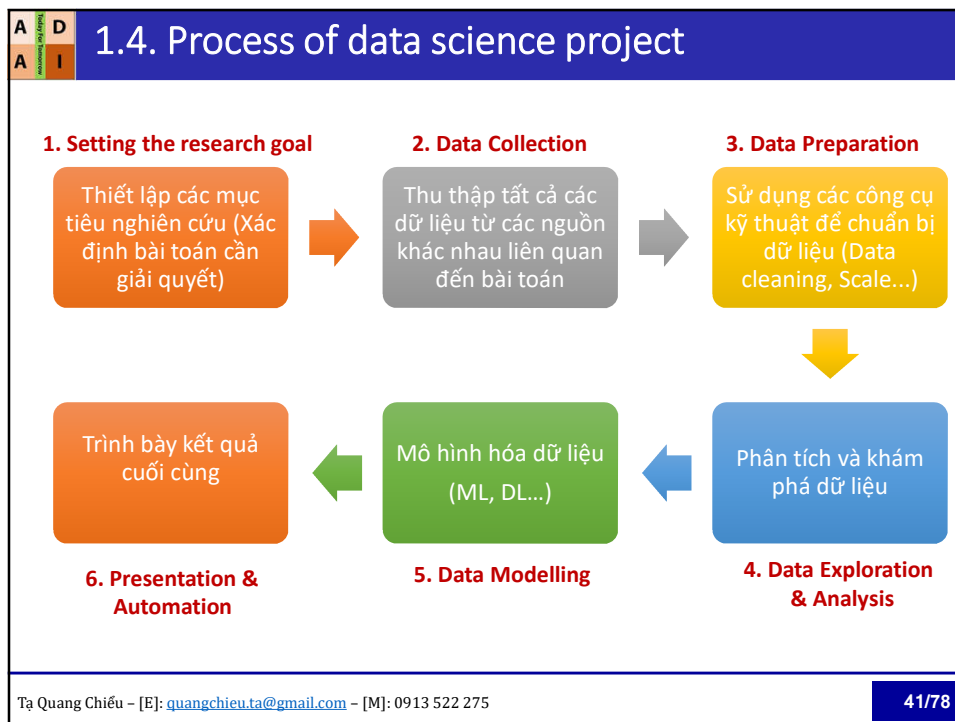
- **Các kiểu dữ liệu trong khoa học dữ liệu**
  - **Dữ liệu âm thanh, hình ảnh và video (Audio, Image, Video)**
    - Thách thức đối với các nhà khoa học dữ liệu.
    - Tự động nhận dạng đối tượng cụ thể qua âm thanh, hình ảnh.
    - Tính toán chuyển động của đối tượng theo thời gian thực.
    - Thu nhận thông tin thông qua các video sử dụng thuật toán Deep Learning.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

40/78

40



41

**B1: Thiết lập mục tiêu nghiên cứu**

- **Setting the research goal**
  - Một dự án data science bắt đầu bằng việc tìm hiểu xác định mục tiêu (what), lý do cần thực hiện (why) và thực hiện nó như thế nào (how)

- Trả lời ba câu hỏi này là mục tiêu của giai đoạn đầu tiên, để mọi người biết phải làm gì và có thể đồng ý về hướng hành động tốt nhất.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

42/78

42

A	D	B1: Thiết lập mục tiêu nghiên cứu
A	I	
<ul style="list-style-type: none"> <li>• <b>Kết quả của quá trình này bao gồm:</b> <ul style="list-style-type: none"> <li>• Mục tiêu nghiên cứu</li> <li>• Hiểu nhiệm vụ và bối cảnh dự án</li> <li>• Dự án cần dữ liệu gì? lấy như thế nào? Ở đâu? Số lượng như thế nào?</li> <li>• Tài nguyên nào bạn muốn sử dụng (nhân lực, thời gian)</li> <li>• Sản phẩm của dự án là gì, sẽ được sử dụng ở đâu?</li> <li>• Đây là thước đo dự án có thành công hay không?</li> <li>• Một kế hoạch hành động với thời gian biểu (timetable)</li> </ul> </li> </ul>		
Tạ Quang Chiêu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		43/78

43

A	D	B1: Thiết lập mục tiêu nghiên cứu
A	I	
<ul style="list-style-type: none"> <li>• <b>Mục tiêu nghiên cứu</b> <ul style="list-style-type: none"> <li>• Mục tiêu nghiên cứu có thể được bắt nguồn từ nhu cầu hoặc nhiệm vụ từ công ty, xã hội hoặc phát sinh trong quá trình thực hiện một dự án <ul style="list-style-type: none"> <li>• Tìm danh sách khách hàng tiềm năng đang cần vay vốn</li> <li>• Từ tag người vào ảnh có trước (facebook)</li> <li>• Dự đoán khả năng bị bệnh của một người</li> <li>• Hệ thống recommend mua hàng cho khách hàng (Amazon)</li> </ul> </li> </ul> </li> </ul>		
Tạ Quang Chiêu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		44/78

44



A

D

A

I

## B1: Thiết lập mục tiêu nghiên cứu

- **Ví dụ**
  - Mục tiêu: dự đoán giá viên kim cương 1.35 carats
  - Tìm hiểu về ngành công nghiệp kim cương



Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

45/78

45

A

D

A

I

## B2: Thu thập dữ liệu

- **Thu thập dữ liệu (Data collection)**
  - Mục tiêu: thu thập tất cả các dữ liệu cần cho dự án
  - Dữ liệu có thể được lưu trữ ở nhiều dạng, từ các tệp văn bản đơn giản đến các bảng trong cơ sở dữ liệu
  - Ví dụ: thu thập giá kim cương tại các cửa hàng bán lẻ




Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

46/78

46

A

D

A

I

## B3. Chuẩn bị dữ liệu (Data Preparation)

**Là bước quan trọng, chiếm nhiều thời gian và nguồn lực nhất trong bất kỳ một dự án khoa học dữ liệu nào (80%)**

**Làm sạch dữ liệu**  
Chỉnh sửa dữ liệu bằng cách bổ sung các dữ liệu còn thiếu, thay thế và hiệu chỉnh các dữ liệu nhiễu

**Giảm kích thước dữ liệu**  
Đảm bảo chất lượng ban đầu của dữ liệu

**Chuyển đổi dữ liệu**  
Bao gồm chuẩn hóa, chuyển đổi và tổng hợp dữ liệu sử dụng các phương pháp ETL.

**Tích hợp dữ liệu**  
Xử lý sự không tương thích giữa các dữ liệu

**80%**  
of time & resources spent on any data project is data preparation\*

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

47/78

47

A

D

A

I

## B3: Chuẩn bị dữ liệu (Data Preparation) (2)

• **Chuẩn bị dữ liệu**

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	
0.6	4172
Two	21764
1.1	4682
1.31	6171

Dữ liệu không hợp lệ (Improper)

Dữ liệu bị thiếu (missing)

Dữ liệu bị thiếu (missing)

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

48/78

48

**B3: Chuẩn bị dữ liệu (Data Preparation) (3)**

- Chuẩn bị dữ liệu**

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	
0.6	4172
Two	21764
1.1	4682
1.31	6171

Dữ liệu không hợp lệ (Improper)

Dữ liệu bị thiếu (missing)

Dữ liệu bị thiếu (missing)

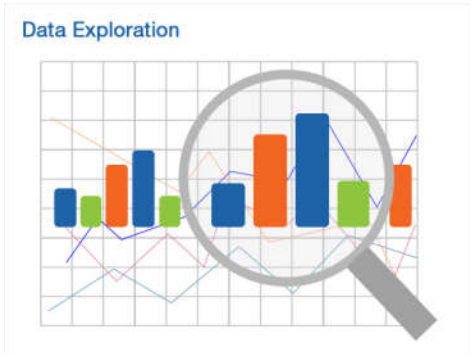
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	693
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	4325
0.6	4172
Two	21764
1.1	4682
1.31	6171

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

49/78

49

**B4: Khám phá dữ liệu (Data Exploration)**



Data Exploration

- Khám phá dữ liệu để **hiểu rõ** hơn về **mối qua hệ giữa các biến** và nhận biết được các **thông tin** được truyền tải **từ dữ liệu**.
- Lựa chọn các **mô hình phù hợp** và các **biến quan trọng** để đưa vào mô hình.

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

50/78

50


A D  
A I

B4: Khám phá dữ liệu (Data Exploration)(2)

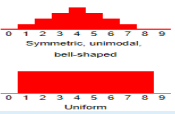

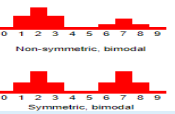
• **Thiết kế mô hình**

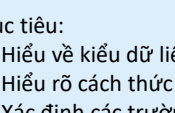
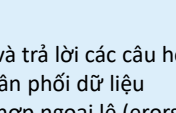
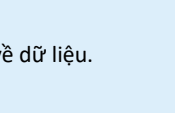
Phân tích dữ liệu khai phá

But what is Exploratory Data Analysis?



Exploratory Data Analysis (EDA) là phương pháp phân tích dữ liệu chủ yếu sử dụng kỹ thuật về biểu đồ, đồ thị.

Mục tiêu:

- Hiểu về kiểu dữ liệu và trả lời các câu hỏi về dữ liệu.
- Hiểu rõ cách thức phân phối dữ liệu
- Xác định các trường hợp ngoại lệ (errors)
- Xác định các quy luật có trong dữ liệu (pattern)

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

51/78

51

A D  
A I

B4: Khám phá dữ liệu (Data Exploration)(3)

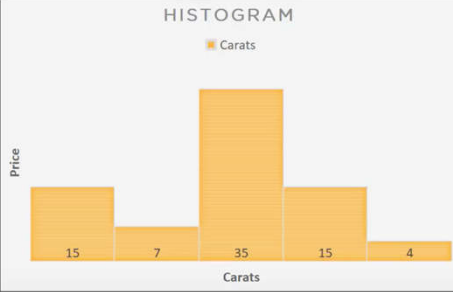
• **Phân tích khám phá dữ liệu EDA**

Techniques:

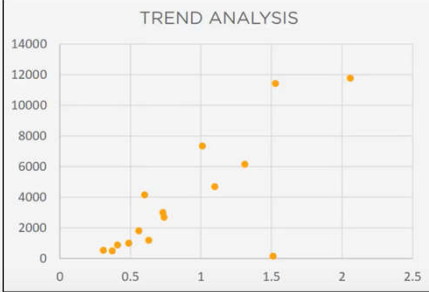
• Histogram

• Trend Analysis

HISTOGRAM



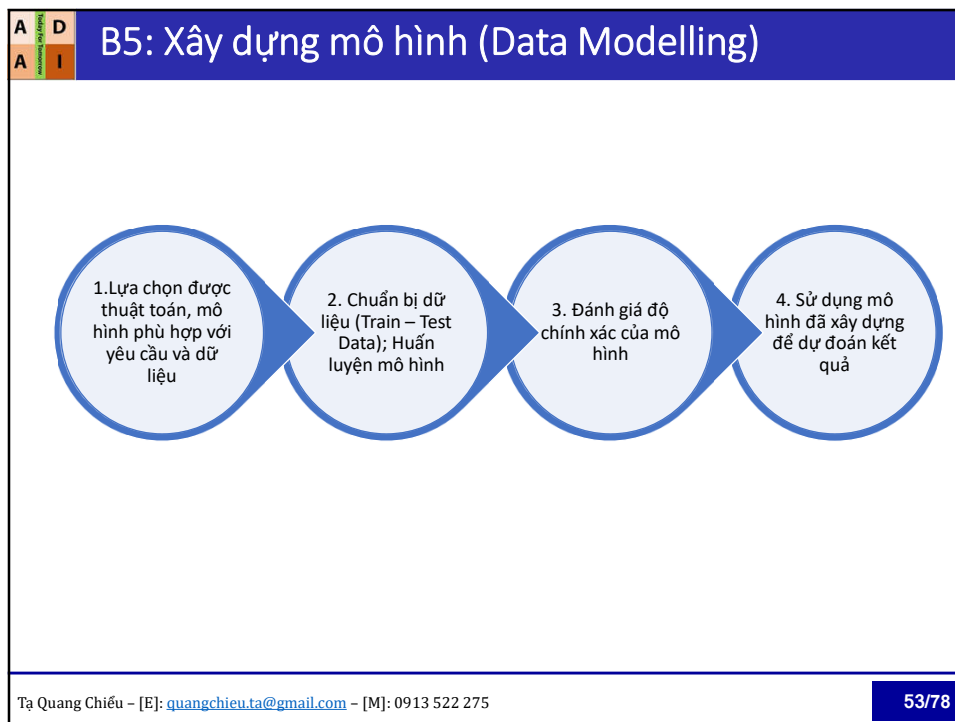
TREND ANALYSIS



Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

52/78

52



53

**B5: Xây dựng mô hình (Data Modelling) (2)**

**Lựa chọn thuật toán, mô hình:**

- Dựa vào dữ liệu của bài toán cụ thể, để lựa chọn được thuật toán phù hợp.
  - Với bài toán dự đoán giá Kim cương: kết quả phân tích tiến triển tuyến tính. Do đó thuật toán **Hồi quy tuyến tính** được lựa chọn để xây dựng mô hình trong trường hợp này.
  - Dữ liệu mẫu (train data) được sử dụng để chạy mô hình.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

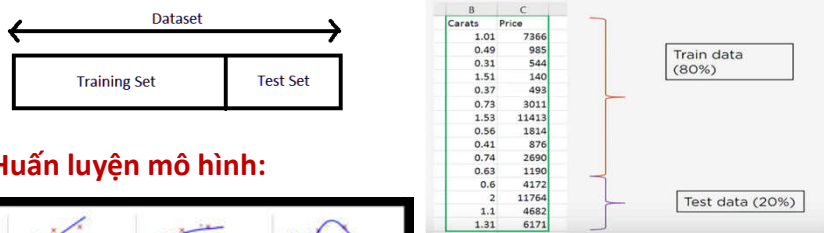
54/78

54

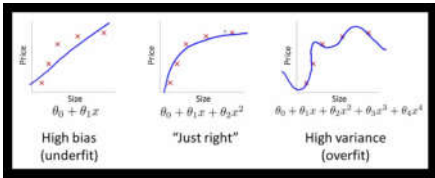
**B5: Xây dựng mô hình (Data Modelling) (3)**

**Chuẩn bị dữ liệu cho huấn luyện và kiểm thử mô hình:**

- Phân tách tập dữ liệu cho việc huấn luyện và kiểm thử mô hình. Thông thường tập dữ liệu sẽ được chia theo tỷ lệ **80:20** (80% dữ liệu được sử dụng để huấn luyện mô hình, 20% được sử dụng để kiểm thử mô hình)



**Huấn luyện mô hình:**



Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

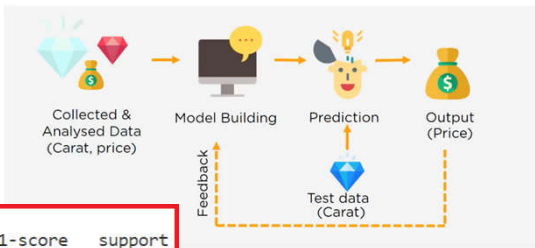
55/78

55

**B5: Xây dựng mô hình (Data Modelling) (4)**

**Đánh giá mô hình:**

- Sử dụng tập dữ liệu kiểm tra (Test Data), đánh giá độ chính xác của mô hình đã xây dựng



Classification report is as:

	precision	recall	f1-score	support
0	0.95	0.91	0.93	64
1	0.95	0.97	0.96	107
avg / total	0.95	0.95	0.95	171

Accuracy of model is 0.9473684210526315

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

56/78

56

A D  
A I

## B5: Xây dựng mô hình (Data Modelling) (5)

Chạy mô hình với dữ liệu mới và đưa ra các dự đoán:

- Kết quả dự đoán viên kim cương 1.35 carat có giá 10.000 với mô hình được xây dựng dựa vào thuật toán hồi quy tuyến tính.**

The figure is a scatter plot titled 'Price of diamond' on the y-axis and 'Carat' on the x-axis. The y-axis has labels at Rs. 5,000, Rs. 10,000, and Rs. 15,000. The x-axis has labels at 0.5, 1.0, and 1.5. Several data points are plotted as black dots. A solid orange line, labeled 'Regression line', shows a positive linear trend. A red dashed line extends from the regression line down to the x-axis at 1.35 and across to the y-axis at Rs. 10,000. Below the x-axis, there are three diamond icons: a red one at 0.5 carats, a blue one at 1.35 carats, and a light blue one at 1.5 carats.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

57/78

57

A D  
A I

## B6: Trình diễn kết quả (Presentation & Automation)

Trình bày kết quả của dự án với khách hàng, công ty

Tích hợp với các công cụ (ứng dụng) khác

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

58/78

58



A

D

A

I

## Một số lưu ý:

Không phải mọi dự án khoa học dữ liệu (Data science) sẽ tuân theo một quy trình giống nhau.

Các bước cụ thể trong các dự án khác nhau có thể khác nhau đôi chút

Các bước này phụ thuộc vào nhà khoa học dữ liệu, công ty và các yếu tố khác của dự án.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

59/78

59

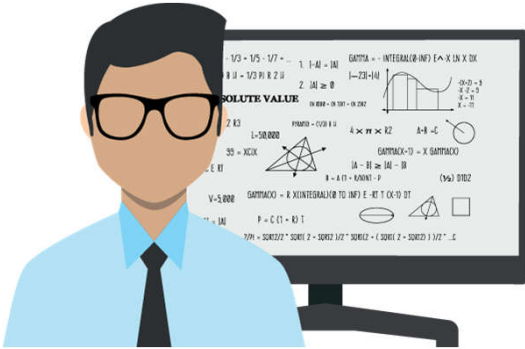
A

D

A

I

## 1.5. Data Scientist



Kiến thức, kỹ năng và cơ hội việc làm  
của một nhà Khoa học dữ liệu.

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

60/78


60

**A D**  
**A I**

**Nhà khoa học dữ liệu?**

### WHO AM I?

I am a part analyst & part artist. I use my analytical and technical abilities to extract meaning/insights from massive data sets.



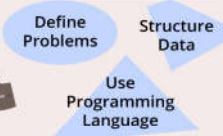
### WHAT DO I DO?

1. I cleaning existing raw data & build models to predict future data.
2. I go beyond merely collecting and reporting data, to look at data from multiple angles & give meaning to it.
3. I identify the correct business problem(s) & offer solutions (via visualizations, reports or blogs) by best applying the data.

### WHAT DO I RELY ON?

1. Analytics
2. Predictive Models
3. Statistical Analysis & Modeling
4. Data Mining
5. Sentiment Analysis
6. What-if-Analysis

### THE PROCESS I FOLLOW



### WHAT DO I EARN?

After oil & gas geologists, mine is the 2nd highest paid job in the world!

### HOW DO I HELP ORGANISATIONS TODAY?

- Increase data accuracy
- Develop strategies
- Improve operational efficiency
- Reduce costs
- Mitigate risks
- Offer personalized products/services

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

**61/78**

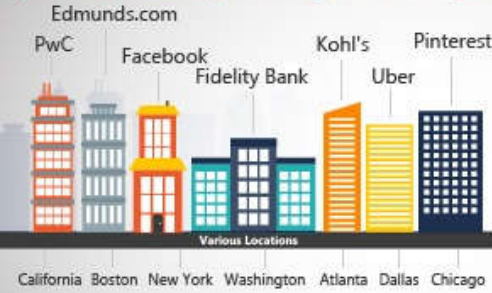
61

**A D**  
**A I**

**Tại sao bạn nên trở thành một nhà khoa học dữ liệu?**

## Why you should become a Data Scientist ?

### 2,339 data science job listings from companies




Average data science entry-level salary in US

\$91,000 (US)      \$110,000 (Silicon Valley)



### Average data science managerial level salary ranges from



\$140,000      \$240,000

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

**62/78**

62

**Tại sao bạn nên trở thành một nhà khoa học dữ liệu?**

### Why you should become a Data Scientist ?

**Demand**  
By 2026, the world will need 11.5M data scientists. This is a perfect time to get on board.

**Influence**  
In 2018, AI deployments will bring 58% revenue growth worldwide. Help create value!

**Salary**  
As a data scientist you could earn 20% more than a software engineer.

**Fun**  
What if your job was to teach AI to do cool stuff?

**Things that matter**  
Data scientists keep us healthy and safe - all from the comfort of their desks. Modern heroes!

**Make the world easier**  
AI guys build intelligent systems to help non-AI guys live better.

Do you know that AI plays Atari games better than any human?

Deep learning can make you an artist too. Just try style transfer!

Deep learning helps save endangered North Atlantic whales...

...diagnose diabetic retinopathy

and even predict dangerous seismic events!

They create smart bots

and fraud detection solutions.

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

63/78

63

**Phân biệt:**

- Phân biệt nhà Khoa học dữ liệu/ Phân tích dữ liệu/ Kỹ sư dữ liệu

Data Scientist	Data Analyst	Data Engineer
<ul style="list-style-type: none"> <li>Nhà khoa học dữ liệu là một người sử dụng trình độ kỹ thuật dữ liệu tiên tiến để đưa ra các quyết định chiến lược.</li> <li>Họ là những người có vị trí cao nhất trong nhóm có kiến thức chuyên sâu về thống kê, thao tác dữ liệu và học máy.</li> <li>Dựa trên các sản phẩm của các Kỹ sư dữ liệu và các nhà phân tích dữ liệu để đưa những giá trị và phương hướng hành động cụ thể cho doanh nghiệp.</li> </ul>	<ul style="list-style-type: none"> <li>Nhà phân tích dữ liệu giữ vị trí thấp nhất trong nhóm phân tích dữ liệu.</li> <li>Họ cần nắm vững các kỹ thuật về xử lý dữ liệu, mô hình hóa, xây dựng báo cáo.</li> <li>Các nhà phân tích dữ liệu có thể trở thành các nhà Khoa học dữ liệu hoặc Kỹ sư dữ liệu khi có nhiều kinh nghiệm trong lĩnh vực này.</li> </ul>	<ul style="list-style-type: none"> <li>Một Kỹ sư dữ liệu có vị trí trung gian giữa nhà phân tích dữ liệu và nhà khoa học dữ liệu.</li> <li>Họ cần có các kiến thức chuyên môn trong phát triển, xây dựng và bảo trì kiến trúc của hệ thống.</li> <li>Họ làm việc với các dữ liệu lớn và gửi báo cáo kết quả cho các nhà khoa học dữ liệu để thực hiện phân tích đánh giá.</li> </ul>

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

64/78

64

A D  
A I

Phân biệt:

- Phân biệt nhà Khoa học dữ liệu/ Phân tích dữ liệu/ Kỹ sư dữ liệu

### Skillset

Data Scientist

Programming Languages

Python, R, SQL, SAS, Java

Frameworks

Pig, Spark, Hadoop

Technologies

Machine Learning, Deep Learning

Data Analyst

Programming Languages

Python, R, SQL, SAS, JavaScript

Tools

SAS Miner, Microsoft Excel, SSAS, SPSS

Data Engineer

Programming Languages

Python, R, SQL, SAS, Java

Frameworks

Hadoop, MapReduce, Hive, Pig, Apache Spark, Data Streaming, NoSQL

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

65/78

65

A D  
A I

Phân biệt:

- Phân biệt nhà Khoa học dữ liệu/ Phân tích dữ liệu/ Kỹ sư dữ liệu

### Roles and Responsibilities

Data Scientist

- Làm sạch và khai phá dữ liệu, xử lý các dữ liệu phi cấu trúc.
- Thiết kế mô hình để làm việc với dữ liệu lớn.
- Suy luận và giải thích các phân tích dựa trên dữ liệu lớn.
- Lãnh đạo nhóm phân tích dữ liệu để đạt được các mục tiêu đặt ra.
- Đưa ra các kết luận có ảnh hưởng trực tiếp tới các hoạt động kinh doanh.

Data Analyst

- Thu thập thông tin từ cơ sở dữ liệu thông qua truy vấn
- Xử lý dữ liệu và cung cấp các báo cáo tóm tắt sử dụng các thuật toán cơ bản trong công việc.
- Có kỹ năng cốt lõi về thống kê, khai phá dữ liệu, trộn dữ liệu, trực quan hóa dữ liệu và phân tích dữ liệu khai phá

Data Engineer

- Khai phá dữ liệu phục vụ trích xuất thông tin.
- Chuyển đổi các dữ liệu lỗi sang các dạng phù hợp cho các phân tích tiếp theo.
- Xây dựng truy vấn dữ liệu.
- Bảo trì thiết kế và kiến trúc dữ liệu.
- Phát triển các kho dữ liệu lớn

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

66/78

66

A
D

A
I

## Nhà khoa học dữ liệu

• **Đặc trưng của một nhà Khoa học dữ liệu**



**Sự ham hiểu biết**

Đặt câu hỏi để hiểu rõ vấn đề. Tò mò muốn khám phá những gì ẩn giấu bên trong.



**Khả năng phán đoán**

Xác định cách thức mới để giải quyết vấn đề và chỉ rõ các tiêu chí quan trọng.



**Kỹ năng giao tiếp**

Trao đổi để truyền tải giá trị thu nhận được với người khác

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

67/78

67

A
D


A
I

## Một nhà khoa học dữ liệu cần gì?


Mô hình toán học giúp tăng hiệu suất tính toán và dự đoán

2

MATHEMATICAL MODELLING



Thống kê là nền tảng của KHD, trích xuất kiến thức và thu nhận kết quả phân tích từ dữ liệu




STATISTICS

3

Kiến thức, kỹ năng về lập trình để xây dựng các chương trình trích xuất, mô hình hóa dữ liệu

COMPUTER PROGRAMMING





4

ML là công cụ quan trọng trong KHD, cung cấp cách thức giải quyết các vấn đề

1


MACHINE LEARNING





Truy vấn trong CSDL giúp đặt ra các câu hỏi chính xác để giải quyết các vấn đề đặt ra trong KHD

DATABASES



5

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

68/78

68

**Những kiến thức cần trang bị?**

The infographic displays seven essential skills for data science, each with an icon and a label in a blue box:

- Kiến thức CSDL**: Represented by a database icon.
- Thống kê**: Represented by a bar chart icon.
- Công cụ lập trình**: Represented by a code editor icon with various programming languages listed.
- Xử lý dữ liệu**: Represented by a circular flow diagram icon.
- Học máy**: Represented by a brain icon with a gear inside.
- Trực quan hóa dữ liệu**: Represented by an eye icon looking at data blocks.
- Dữ liệu lớn**: Represented by the Hadoop and Spark logos.

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275 **69/78**

69

**Những kiến thức cần trang bị?**

**1 – Kiến thức về cơ sở dữ liệu**

The infographic focuses on database knowledge. It includes a definition of SQL and logos for major database systems:

**Kiến thức CSDL**: Represented by a database icon.

**SQL (Structured Query Language - Ngôn ngữ truy vấn cấu trúc)** là một ngôn ngữ cần thiết để trích xuất một lượng lớn thông tin từ tập dữ liệu. Kiến thức về SQL là bắt buộc cho một nhà khoa học dữ liệu

**Công cụ cần thiết**


Logos for **ORACLE DATABASE**, **Microsoft SQL Server**, **MySQL**, and **TERADATA**.

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275 **70/78**

70

A D  
A I

Những kiến thức cần trang bị?




**2 – Kiến thức về thống kê**


Thống kê

Thống kê là một tập con của toán học liên quan đến việc thu thập, phân tích và thể hiện dữ liệu; Nhà khoa học dữ liệu cần có hiểu biết về thống kê

Thống Kê



Xác Suất




Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
71/78

71

A D  
A I


Những kiến thức cần trang bị?




**3 – Kiến thức về lập trình**

Công cụ lập trình


Thành thạo ít nhất một trong số những ngôn ngữ lập trình dưới đây là cần thiết cho việc phân tích dữ liệu của bất kỳ nhà khoa học dữ liệu nào



- ❑ R is a free software environment for statistical computing and graphics
- ❑ Supports most Machine Learning algorithms for Data Analytics like regression, association, clustering, etc.



- ❑ Python is an open-source general purpose programming language
- ❑ Python libraries like NumPy and SciPy are used in Data Science



- ❑ SAS can mine, alter, manage, and retrieve data from a variety of sources
- ❑ Can perform statistical analysis on the data

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
72/78


72



A D  
A I

Những kiến thức cần trang bị?

### 4 – Kiến thức về thu thập và xử lý dữ liệu



Xử lý dữ liệu

Xử lý dữ liệu là quá trình chuyển đổi từ dữ liệu thô (raw data) thành định dạng phù hợp để làm cho nó hữu ích cho phân tích.

Bao gồm:

Làm sạch dữ liệu thô

➔

Phân tích cấu trúc dữ liệu thô

➔

Làm giàu dữ liệu thô

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275


73/78

73

A D  
A I

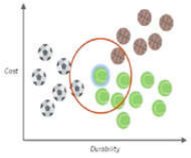
Những kiến thức cần trang bị?

### 5 – Kiến thức về Học máy

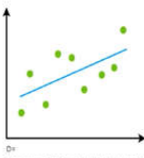


Học máy


Kiến thức về các kỹ thuật học máy như học có giám sát, cây quyết định, hồi quy tuyến tính, KNN...là hữu ích cho công việc của một nhà khoa học dữ liệu



KNN



Linear Regression



Decision Tree

Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

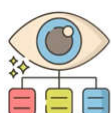
74/78

74

A D  
A I


Những kiến thức cần trang bị?

### 6 – Kiến thức về trực quan hóa dữ liệu



Trực quan hóa dữ liệu

Trực quan hóa dữ liệu là việc nghiên cứu và trình bày dữ liệu trực quan, thông qua các biểu đồ, hình vẽ... Đây là cách truyền tải thông tin một cách rõ ràng và hiệu quả nhất



Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275


75/78

75

A D  
A I


Những kiến thức cần trang bị?

### 7 – Kiến thức về Dữ liệu lớn



Dữ liệu lớn

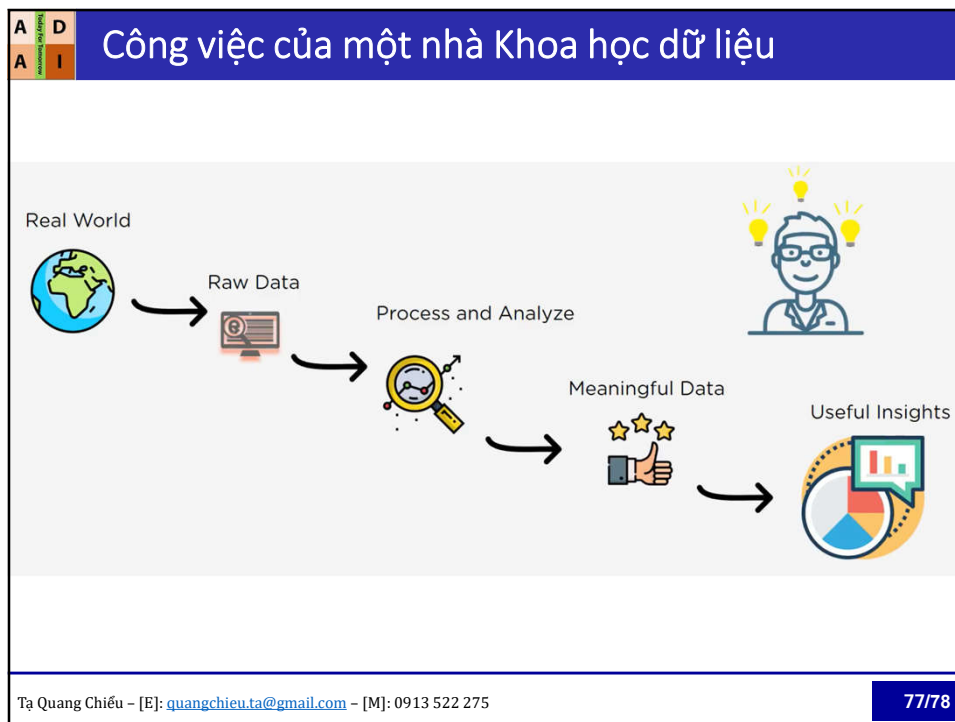
Dữ liệu lớn có nhiều lợi ích khác nhau như: Truy cập vào dữ liệu mạng xã hội có thể cho phép điều chỉnh chiến lược kinh doanh, cải thiện trải nghiệm của khách hàng...



Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

76/78

76



77

**Thảo luận**

- Khoa học dữ liệu trong hoạt động kinh doanh sách trên Amazon.com
- Khoa học dữ liệu trong hoạt động kinh doanh của GRAB

**Hãy nêu một vài bài toán xung quanh bạn có thể áp dụng khoa học dữ liệu để giải quyết nó?**

Tạ Quang Chiêu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

78/78

78