# METHODS OF DATA MINING 1

**Vo Luong Hong Phuoc**

Day 1

# INTRODUCE

**1. Number of**

**periods: 30 periods (1 TC LT, 1 TCTT)**

**Ms. Phuoc, Ms. Hoa Tien, Teacher Tien Thanh CBBM**

## 2. Study time:

- **Theory: 3 periods/session**
- **Practice: details will be announced later**

## 3. Location:

**Theory + exercises: B35**

**Practice: B35 (if students have a laptop)**

# Exercises in grade 1

1. What methods can be used in data mining?

2. In the subject "measuring... DLNN", what part do you find most difficult to understand and which part is the most interesting?

## 5. Knowledge needed:

- Probability theory and mathematical statistics;
- Numerical method
- Programming: Fortran, mathlab

## 6. Some requirements for students

- Must prepare assignments before practice -
Practice: 1 student/submitted assignment,

**Sinh viên phải tham gia lớp học như sau:**

- Dự lớp lý thuyết tối thiểu: 60%
- Dự lớp thực hành tối thiểu: 80%
- Dự lớp bài tập tối thiểu: 80%
- Dự lớp thảo luận tối thiểu: 80%
- Yêu cầu khác: đến lớp đúng giờ

| Ký hiệu mục tiêu của học phần (MH) | Mô tả/nội dung mục tiêu học phần | Mức độ năng lực đạt được (theo thang đánh giá Bloom) | Ghi chú |
|---|---|---|---|
| **KIẾN THỨC** | | | |
| MH1.1 | Sinh viên trang bị kiến thức sâu về các phương pháp phân tích dữ liệu | **3** | |
| MH1.2 | Sinh viên cũng sẽ được trang bị những kỹ năng lập trình (Matlab, Fortran), sử dụng phần mềm ứng dụng (SPSS, R) và kỹ năng phân tích số liệu | **2.5** | |
| **KỸ NĂNG** | | | |
| MH2.1 | Sinh viên sẽ có kỹ năng về lập trình, xử lý số liệu về các yếu tố khí tượng, thủy hải văn và môi trường | **2.5** | |
| MH2.2 | Sinh viên có khả năng tư duy nghiên cứu sáng tạo, độc lập, và khả năng làm việc nhóm. | **2.5** | |
| **THÁI ĐỘ** | | | |
| MH3.1 | Sinh viên được rèn luyện tính kỷ luật, chính xác, cẩn thận trong công việc, sự trung thực với số liệu | **3** | |
| MH3.2 | Nghiêm túc và trung thực trong học tập và thi cử | **3.5** | |

**Chuẩn đầu ra (CĐR) của học phần**

| Thứ tự các CĐR | Ký hiệu CĐR học phần (CHP) | Mô tả/nội dung CĐR học phần | Mức độ giảng dạy (I, T, U) * | Liên kết giữa CĐR học phần và mục tiêu học phần | Liên kết giữa CĐR học phần và CĐR chương trình đào tạo |
|---|---|---|---|---|---|
| | | **KIẾN THỨC** | | | |
| 1 | CHP1 | Hiểu được quy trình khai thác dữ liệu | I,T | MH1.1 | CCT1.2 |
| 2 | CHP2 | Hiểu được các phương pháp khai thác dữ liệu trong Khoa học Trái đất (phương pháp biến đổi Fourier, phương pháp phân tích phổ không tham số, phương pháp nội suy và các phép lọc) | I, T | MH1.2 | CCT1.2 |
| 3 | CHP3 | Ứng dụng matlab, Fortran, R, SPSS trong nghiên cứu thống kê, phân tích phổ, nội suy và phân tích số liệu | T, U | MH1.2 | CCT1.2 |
| | | **KỸ NĂNG** | | | |
| 4 | CHP4 | Ứng dụng các công cụ chuyên ngành (R, SPSS) trong phân tích và dự báo | T, U | MH2.1 | CCT3.1 |
| 5 | CHP5 | Kỹ năng về lập trình (Fortran, Matlab), xử lý số liệu về các yếu tố khí tượng, thủy hải văn và môi trường | T, U | MH2.1 | CCT2.1 |
| 6 | CHP6 | Khả năng tư duy nghiên cứu và khả năng làm việc nhóm. | U | MH2.2 | CCT2.3 |
| | | **THÁI ĐỘ** | | | |
| 7 | CHP7 | Sinh viên được rèn luyện tính kỷ luật, chính xác, cẩn thận trong công việc, sự trung thực với số liệu | I, U | MH3.1 | CCT4.1 |
| 8 | CHP8 | Nghiêm túc và trung thực trong học tập và thi cử | U | MH3.2 | CCT4.1 |

*(*) I (Introduce): giới thiệu; T (Teach): dạy; U (Utilize): sử dụng*

## 5. Hình thức, phương pháp và trọng số đánh giá kết quả học phần

| Hình thức đánh giá | Nội dung chi tiết | Phương pháp đánh giá (đánh dấu X) | | | | Ký hiệu bài đánh giá | Trọng số đánh giá | Ghi chú |
|---|---|---|---|---|---|---|---|---|
| | | Viết | Trắc nghiệm | Vấn đáp | Thực hành | | | |
| Đánh giá quá trình | Tổng điểm quá trình | | | | | ĐG1 (tổng điểm từ ĐG1.1 đến ĐG1.6) | **60** | Quy định từ 40% đến 60% |
| | Điểm kiểm tra giữa kỳ | | | | X | ĐG1.1 | 30 | |
| | Điểm kiểm tra thường xuyên | x | | | | ĐG1.2 | 10 | |
| | Điểm thảo luận | | | | | | | |
| | Điểm thực hành | | | | x | ĐG1.3 | 10 | |
| | Điểm báo cáo nhóm | | | x | | ĐG1.4 | 10 | |
| | Điểm chuyên cần | | | | | | | |
| Đánh giá tổng kết | Thi cuối học kỳ | x | | | | ĐG2 | 40 | Quy định tối thiểu 40% |

| The name of the lecture of the phn student | Second week |
|---|---|
| Chapter 1: Data mining Lesson | first |
| 1.1: Concepts Lesson 1.2 | first |
| Data mining process Lesson 1.3 Data | first |
| sampling and presentation process Chapter 2: | first |
| Statistical methods in mining data | 2 |
| Lesson 2.1: Statistical methods | 2 |
| Lesson 2.2: Central limit theorem Lesson | 2 |
| 2.3: Application of statistical methods in Earth science | 3 |
| Chapter 3: Spectroscopic Methods in Earth Science land | 4 |
| 3.1 Spectroscopic method | 4 |
| 3.2 Filtering and noise reduction method | 4 |
| 3.3 Application of spectral method in Science Earth | 5 |
| Chapter 4: Data mining applications in the Faculty study Earth | 6 |
| Lesson 4.1: Introduction to R and | 6 |
| applications Lesson 4.2: Introduction to SPSS | 7 |
| and applications Lesson 4.3: Application of interpolation method in using data management | 8 |
| Lesson 4.4: Some practice problems using programming | 9, 10 |
| Total number of periods | 45 |

**Tài liệu học tập**

| STT | Tên tác giả | Năm xuất bản | Tên giáo trình | Tên Nhà xuất bản | Giáo trình chính/Tài liệu tham khảo/Khác | Nơi có thể có tài liệu/trang web |
|---|---|---|---|---|---|---|
| 1 | Richard E. Thomson, William J Emery | 2014 | Data Analysis Methods in Physical Oceanography (3rd Edition) | Elsevier Science | Tài liệu giảng dạy | Thư viện |
| 2 | Phạm Văn Huấn | 2003 | Tính toán trong hải dương học | NXB Đại học Quốc gia Hà Nội | Tài liệu giảng dạy | Thư viện |
| 3 | Julius S. Bendat, Allan G. Piersol | 2010 | Random Data: Analysis and Measurement Procedures (4th Edition) | Wiley | Tài liệu tham khảo | Thư viện |
| 4 | Nguyễn Văn Tuấn | 2014 | Phân tích dữ liệu với R | NXB Tổng hợp TP. Hồ Chí Minh | Tài liệu tham khảo | Thư viện |
| 5 | Stanislaw R. Massel | 1999 | Fluid Mechanics for Marine Ecologists | Springer | Tài liệu tham khảo | Thư viện |
| 6 | UNESCO | 1983 | Algorithms for computation of fundamental properties of seawater. | Unesco technical papers in marine science. | Tài liệu tham khảo | researchgate.net/publication/335494 03_Algorithms_for_Computation_o f_Fundamental_Properties_of_Seaw ater |

# Exercises in class

1. Where can the data come from?

2. Let's give an example of a data object to exploit.

3. Please suggest steps for exploitation

# Chapter 1. Data mining

- **Definition:**
+ data (data, data bank) +
data mining: data mining: process

- **Origin of data:** + from
data bank + from direct
measurement data (web-site) +
from forecast model
+ from actual measured data

***Principle of measurement: depends***

+ purpose +

measuring object +

measuring tools and equipment

+ Measurement rules

# 2. Overview of data mining

- Origin of data
- Principles of measurement
- Measurement method to get data
- Correction and use of data
- Methods of calculation and data processing

# 1. Concepts

## DATA ANALYSIS

Data analysis is concerned with a variety of different tools and methods that have been developed to query existing data, discover exceptions, and verify hypotheses

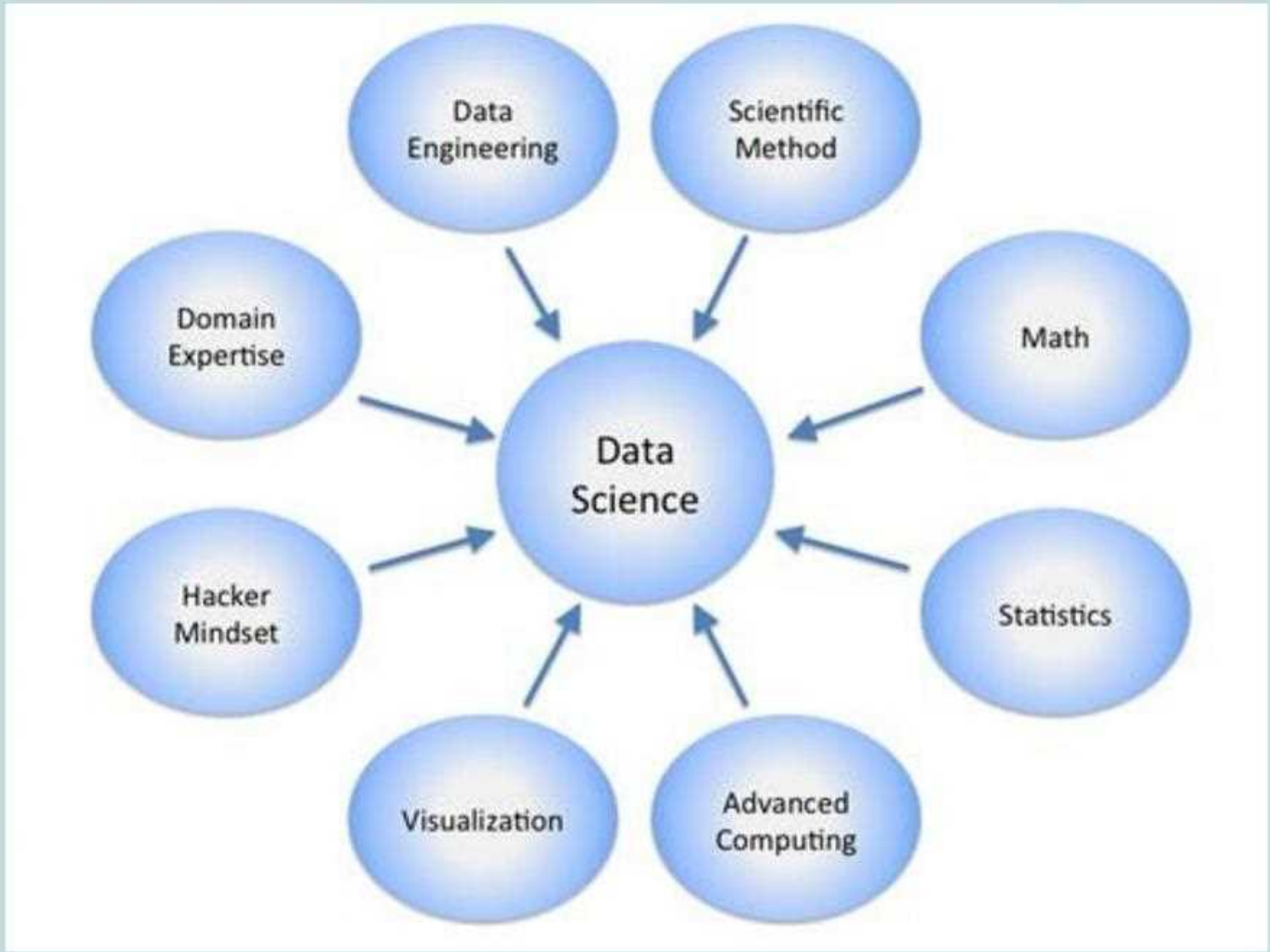
**DATA MINING** can be considered as a collection of methods for drawing inferences (infer / draw conclusions / implications) from data. The aims of data mining and some of its methods overlap with those of classical statistics.

ÿIt should be kept in mind that both data mining and statistics are not business solutions; they are ***just technologies.***

ÿAdditionally, there are still some philosophical and methodological ***differences*** between them.

# What is data mining?

- Extracting ("mining") knowledge from large amounts of data. (KDD: Knowledge Discovery from Data)




Data mining focuses on the discovery of (previously) *unknown* properties on the data.

# DATA MINING

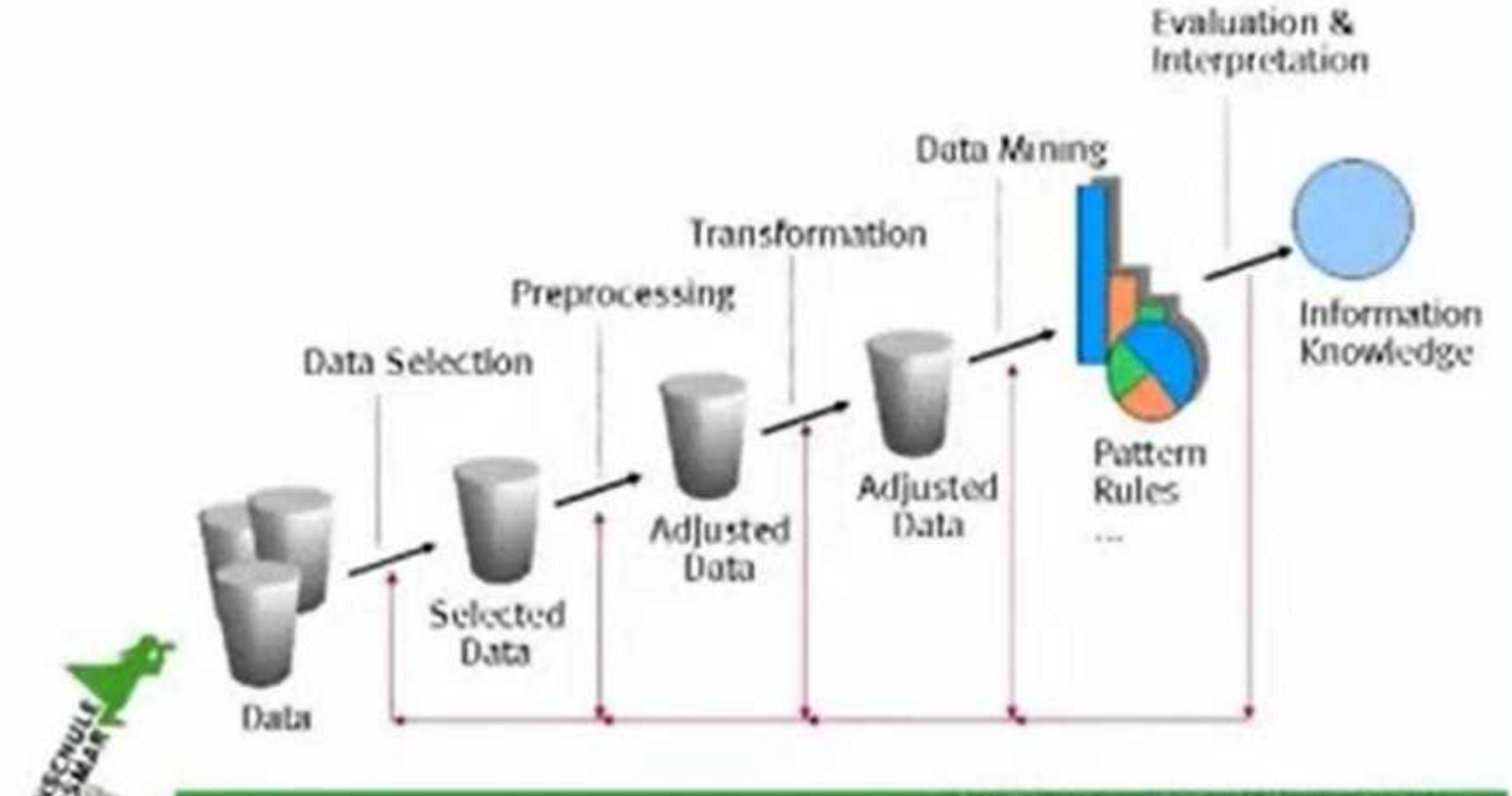
## CONCEPT MINING

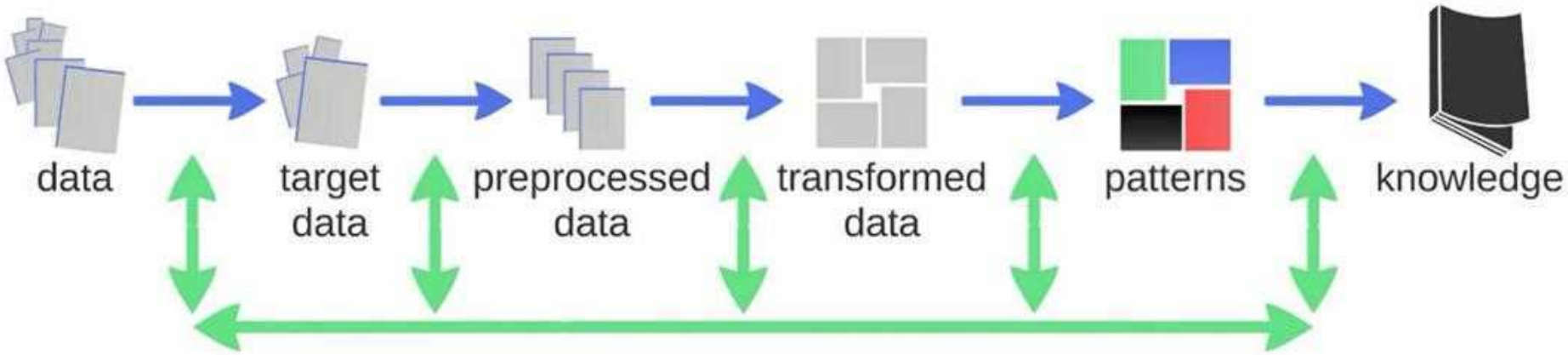
Concept mining is a process that focuses on extracting ideas and concepts found in documents.

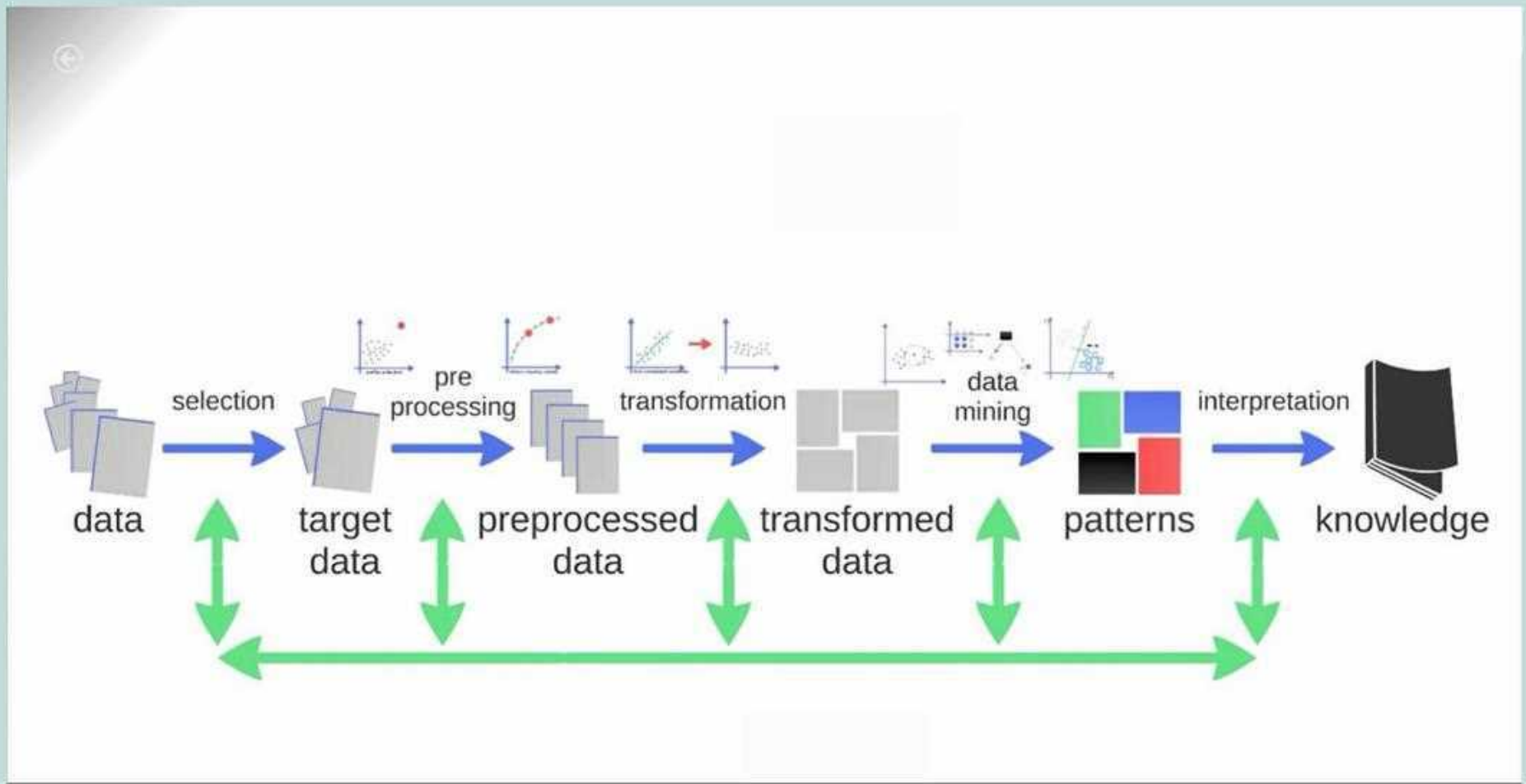
## DATA MINING MODELS

# Data Mining Process Model

### Data Mining Model

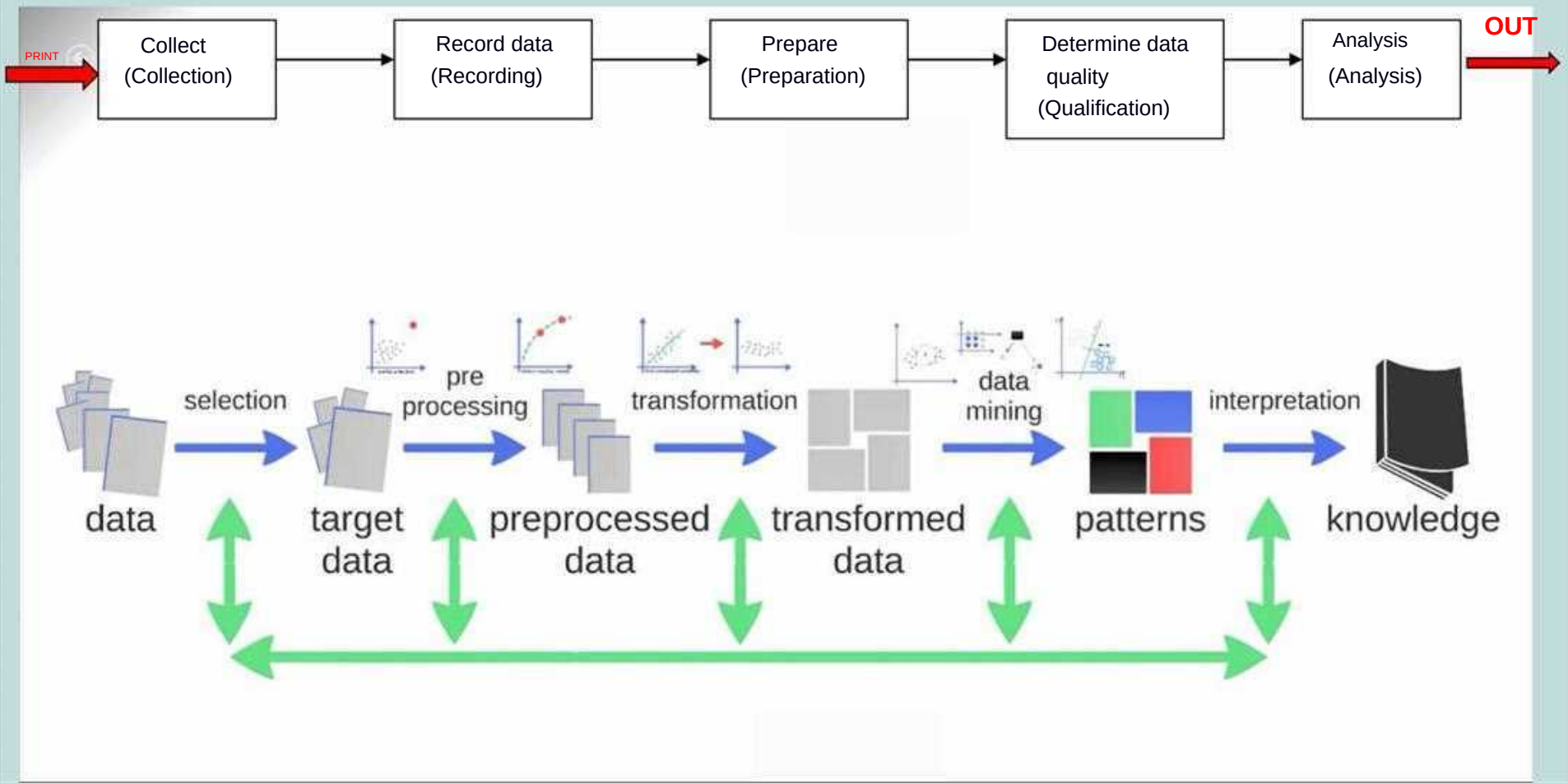




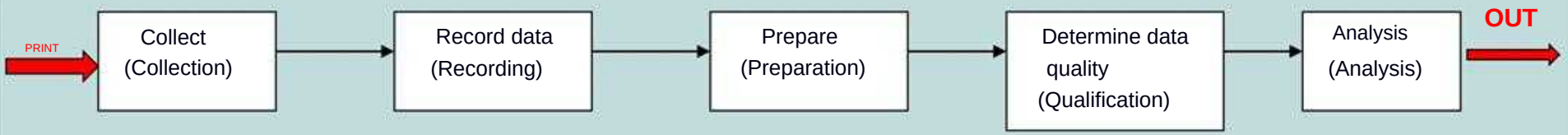


# 2. Sampling procedures

## Data mining steps



# 2. Sampling procedures



# 3. Sampling procedure and data presentation

**Error:** describes the accuracy of measurement, showing
the distance between the measured value and the actual value

-Classification errors

-+ due to: Raw error, systematic error, random error

-+ due to evaluation method: absolute ss, relative ss

-How to calculate SS

-+ direct measurements: random SS, instrumental SS, synthetic SS

-+ indirect SS:

-+ SS tools

-+ SS from mo^ picture

-- How to round SS

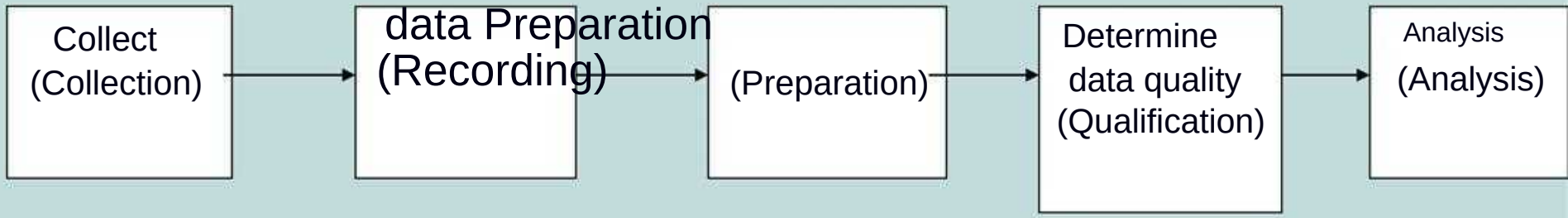
-- How to represent graphs

Software used:

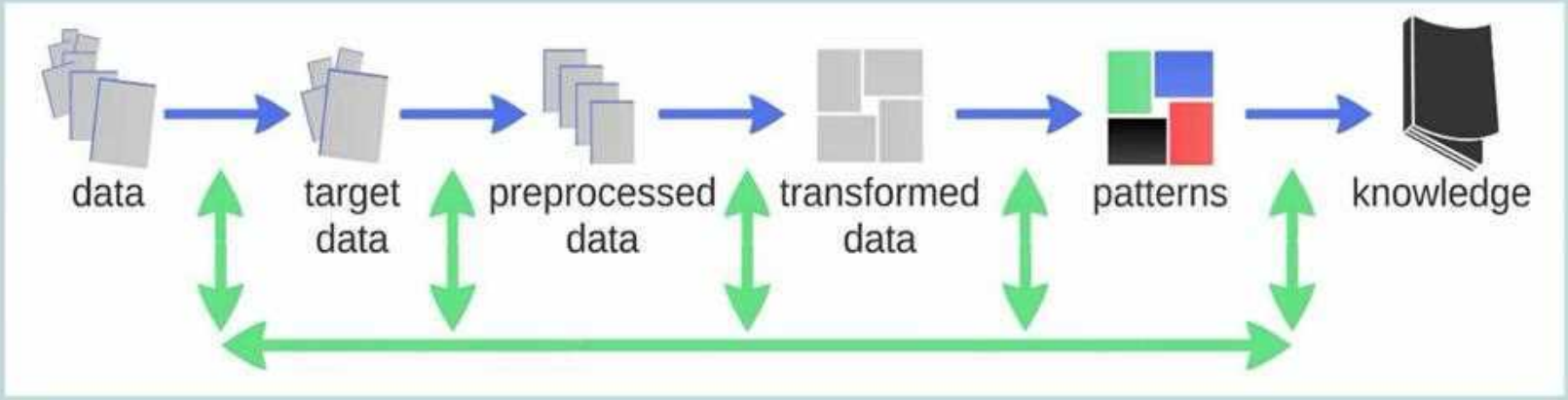
- Excel
-Grapher
-Sufer
-Mapper

# Exercise (15 minutes)

1. From the above example, let's analyze the data mining process according to Bendat Recording




2. Bendat's economic management with the following management:



# Preparation lesson

Review probability theory and statistics: +

Average value, variance, Max, Min...

+ XS distribution function

+…

=> Write programming

# KHAI THÁC DỮ LIỆU I

**Võ Lương Hồng Phước**

Day 1

# GIỚI THIỆU

**1. Số tiết:**

**30 tiết (1TC LT, 1TCTT)**

**Cô Phước, cô Hoa Tiên, Thầy Tiến Thành**

**CBBM**

**2. Thời gian học:**

- **Lý thuyết: 3 tiết/buổi**
- **Thực hành: sẽ thông báo cụ thể sau**

**3. Địa điểm:**

**Lý thuyết + bài tập: B35**

**Thực Hành: B35 (nếu SV có laptop)**

# Bài tập tại lớp 1

1. Phương pháp nào có thể dùng để sử dụng trong khai thác dữ liệu?

2. Trong môn học “đo đạc.. Các DLNN”, các bạn cảm thấy khó hiểu nhất va phần nào lý thú nhất?

5. Các kiến thức cần trang bị:

- Lý thuyết xác suất và thống kê toán học;
- Phương pháp số trị
- Lập trình: Fortran, mathlab

6. Một số yêu cầu cho sinh viên

- Phải chuẩn bị bài tập trước khi thực hành
- Thực tập: 1 sinh viên/bài làm nộp,

**Sinh viên phải tham gia lớp học như sau:**

- Dự lớp lý thuyết tối thiểu: 60%
- Dự lớp thực hành tối thiểu: 80%
- Dự lớp bài tập tối thiểu: 80%
- Dự lớp thảo luận tối thiểu: 80%
- Yêu cầu khác: đến lớp đúng giờ

| Ký hiệu mục tiêu của học phần (MH) | Mô tả/nội dung mục tiêu học phần | Mức độ năng lực đạt được (theo thang đánh giá Bloom) | Ghi chú |
|---|---|---|---|
| **KIẾN THỨC** | | | |
| MH1.1 | Sinh viên trang bị kiến thức sâu về các phương pháp phân tích dữ liệu | **3** | |
| MH1.2 | Sinh viên cũng sẽ được trang bị những kỹ năng lập trình (Matlab, Fortran), sử dụng phần mềm ứng dụng (SPSS, R) và kỹ năng phân tích số liệu | **2.5** | |
| **KỸ NĂNG** | | | |
| MH2.1 | Sinh viên sẽ có kỹ năng về lập trình, xử lý số liệu về các yếu tố khí tượng, thủy hải văn và môi trường | **2.5** | |
| MH2.2 | Sinh viên có khả năng tư duy nghiên cứu sáng tạo, độc lập, và khả năng làm việc nhóm. | **2.5** | |
| **THÁI ĐỘ** | | | |
| MH3.1 | Sinh viên được rèn luyện tính kỷ luật, chính xác, cẩn thận trong công việc, sự trung thực với số liệu | **3** | |
| MH3.2 | Nghiêm túc và trung thực trong học tập và thi cử | **3.5** | |

**Chuẩn đầu ra (CĐR) của học phần**

| Thứ tự các CĐR | Ký hiệu CĐR học phần (CHP) | Mô tả/nội dung CĐR học phần | Mức độ giảng dạy (I, T, U) * | Liên kết giữa CĐR học phần và mục tiêu học phần | Liên kết giữa CĐR học phần và CĐR chương trình đào tạo |
|---|---|---|---|---|---|
| | | **KIẾN THỨC** | | | |
| 1 | CHP1 | Hiểu được quy trình khai thác dữ liệu | I,T | MH1.1 | CCT1.2 |
| 2 | CHP2 | Hiểu được các phương pháp khai thác dữ liệu trong Khoa học Trái đất (phương pháp biến đổi Fourier, phương pháp phân tích phổ không tham số, phương pháp nội suy và các phép lọc) | I, T | MH1.2 | CCT1.2 |
| 3 | CHP3 | Ứng dụng matlab, Fortran, R, SPSS trong nghiên cứu thống kê, phân tích phổ, nội suy và phân tích số liệu | T, U | MH1.2 | CCT1.2 |
| | | **KỸ NĂNG** | | | |
| 4 | CHP4 | Ứng dụng các công cụ chuyên ngành (R, SPSS) trong phân tích và dự báo | T, U | MH2.1 | CCT3.1 |
| 5 | CHP5 | Kỹ năng về lập trình (Fortran, Matlab), xử lý số liệu về các yếu tố khí tượng, thủy hải văn và môi trường | T, U | MH2.1 | CCT2.1 |
| 6 | CHP6 | Khả năng tư duy nghiên cứu và khả năng làm việc nhóm. | U | MH2.2 | CCT2.3 |
| | | **THÁI ĐỘ** | | | |
| 7 | CHP7 | Sinh viên được rèn luyện tính kỷ luật, chính xác, cẩn thận trong công việc, sự trung thực với số liệu | I, U | MH3.1 | CCT4.1 |
| 8 | CHP8 | Nghiêm túc và trung thực trong học tập và thi cử | U | MH3.2 | CCT4.1 |

*(*) I (Introduce): giới thiệu; T (Teach): dạy; U (Utilize): sử dụng*

## 5. Hình thức, phương pháp và trọng số đánh giá kết quả học phần

| Hình thức đánh giá | Nội dung chi tiết | Phương pháp đánh giá (đánh dấu X) | | | | Ký hiệu bài đánh giá | Trọng số đánh giá | Ghi chú |
|---|---|---|---|---|---|---|---|---|
| | | Viết | Trắc nghiệm | Vấn đáp | Thực hành | | | |
| Đánh giá quá trình | Tổng điểm quá trình | | | | | ĐG1 (tổng điểm từ ĐG1.1 đến ĐG1.6) | **60** | Quy định từ 40% đến 60% |
| | Điểm kiểm tra giữa kỳ | | | | X | ĐG1.1 | 30 | |
| | Điểm kiểm tra thường xuyên | x | | | | ĐG1.2 | 10 | |
| | Điểm thảo luận | | | | | | | |
| | Điểm thực hành | | | | x | ĐG1.3 | 10 | |
| | Điểm báo cáo nhóm | | | x | | ĐG1.4 | 10 | |
| | Điểm chuyên cần | | | | | | | |
| Đánh giá tổng kết | Thi cuối học kỳ | x | | | | ĐG2 | 40 | Quy định tối thiểu 40% |

| Tên bài giảng của học phần | Tuần thứ |
|---|---|
| Chương 1: Khai thác dữ liệu | 1 |
| Bài 1.1:Các khái niệm | 1 |
| Bài 1.2 Quy trình khai thác dữ liệu | 1 |
| Bài 1.3 Quy trình lấy mẫu và trình bày dữ liệu | 1 |
| Chương 2: Phương pháp thống kê trong khai thác dữ liệu | 2 |
| Bài 2.1: Phương pháp thống kê | 2 |
| Bài 2.2: Định lý giới hạn trung tâm | 2 |
| Bài 2.3: Ứng dụng phương pháp thống kê trong Khoa học Trái đất | 3 |
| Chương 3: Phương pháp phổ trong Khoa học Trái đất | 4 |
| 3.1 Phương pháp phổ | 4 |
| 3.2 Phương pháp lọc và khử nhiễu | 4 |
| 3.3 Ứng dụng phương pháp phổ trong Khoa học Trái đất | 5 |
| Chương 4: Ứng dụng khai thác dữ liệu trong Khoa học Trái đất | 6 |
| Bài 4.1: Giới thiệu R và ứng dụng | 6 |
| Bài 4.2: Giới thiệu SPSS và ứng dụng | 7 |
| Bài 4.3: Ứng dụng phương pháp nội suy trong sử lý số liệu | 8 |
| Bài 4.4: Một số bài toán thực hành bằng lập trình | 9, 10 |
| Tổng cộng số tiết | 45 |

**Tài liệu học tập**

| STT | Tên tác giả | Năm xuất bản | Tên giáo trình | Tên Nhà xuất bản | Giáo trình chính/Tài liệu tham khảo/Khác | Nơi có thể có tài liệu/trang web |
|---|---|---|---|---|---|---|
| 1 | Richard E. Thomson, William J Emery | 2014 | Data Analysis Methods in Physical Oceanography (3rd Edition) | Elsevier Science | Tài liệu giảng dạy | Thư viện |
| 2 | Phạm Văn Huấn | 2003 | Tính toán trong hải dương học | NXB Đại học Quốc gia Hà Nội | Tài liệu giảng dạy | Thư viện |
| 3 | Julius S. Bendat, Allan G. Piersol | 2010 | Random Data: Analysis and Measurement Procedures (4th Edition) | Wiley | Tài liệu tham khảo | Thư viện |
| 4 | Nguyễn Văn Tuấn | 2014 | Phân tích dữ liệu với R | NXB Tổng hợp TP. Hồ Chí Minh | Tài liệu tham khảo | Thư viện |
| 5 | Stanislaw R. Massel | 1999 | Fluid Mechanics for Marine Ecologists | Springer | Tài liệu tham khảo | Thư viện |
| 6 | UNESCO | 1983 | Algorithms for computation of fundamental properties of seawater. | Unesco technical papers in marine science. | Tài liệu tham khảo | researchgate.net/publication/33549403_Algorithms_for_Computation_of_Fundamental_Properties_of_Seawater |

# Bài tập tại lớp

1. Dữ liệu có thể lấy từ đâu?
2. Hãy ví dụ một đối tượng dữ liệu để khai thác.
3. Hãy đề nghị các bước để khai thác

# Chương 1. Khai thác dữ liệu

- **Định nghĩa**:

+ dữ liệu (data, data bank)
+ khai thác dữ liệu: data mining: quy trình

- **Nguồn gốc của số liệu:**

+từ ngân hàng dữ liệu
+ từ số liệu đo trực tiếp (web-site)
+ từ mô hình dự báo
+ từ số liệu thực đo

***Nguyên tắc đo đạc: phụ thuộc***

+ mục đích
+ đối tượng đo đạc
+ dụng cụ, thiết bị đo đạc
+ Quy phạm đo đạc

# 2.Tổng quan về khai thác dữ liệu

- Nguồn gốc của số liệu
- Nguyên tắc đo đạc
- Phương pháp đo đạc lấy số liệu
- Chỉnh lý và sử dụng số liệu
- Phương pháp tính toán và xử lý số liệu


DATA MINER
DATA
Knowledge

# 1.Các khái niệm

**DATA ANALYSIS**

Data analysis is concerned with a variety of different tools and methods that have been developed to query existing data, discover exceptions, and verify hypotheses

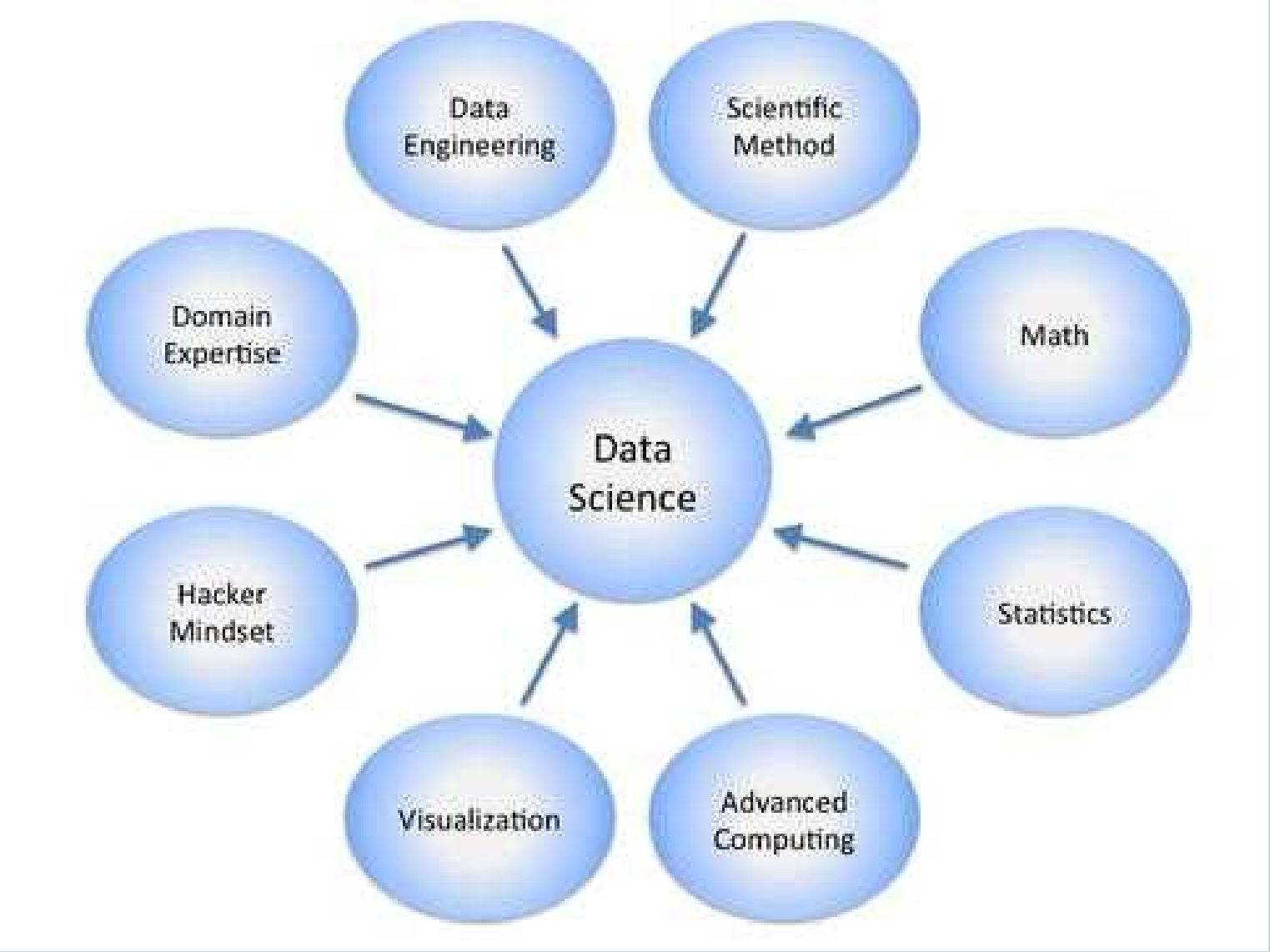
**DATA MINING** can be regarded as a collection of methods for drawing inferences (suy ra / rút ra kết luận/hàm ý) from data. The aims of data mining and some of its methods overlap with those of classical statistics.

⇒ It should be kept in mind that both data mining and statistics are not business solutions; they are ***just technologies***.

⇒ Additionally, there are still some philosophical and methodological ***differences*** between them.

# What is data mining?

- Extracting ("mining") knowledge from large amounts of data. (KDD: Knowledge Discovery from Data)




Data mining focuses on the discovery of (previously) *unknown* properties on the data.

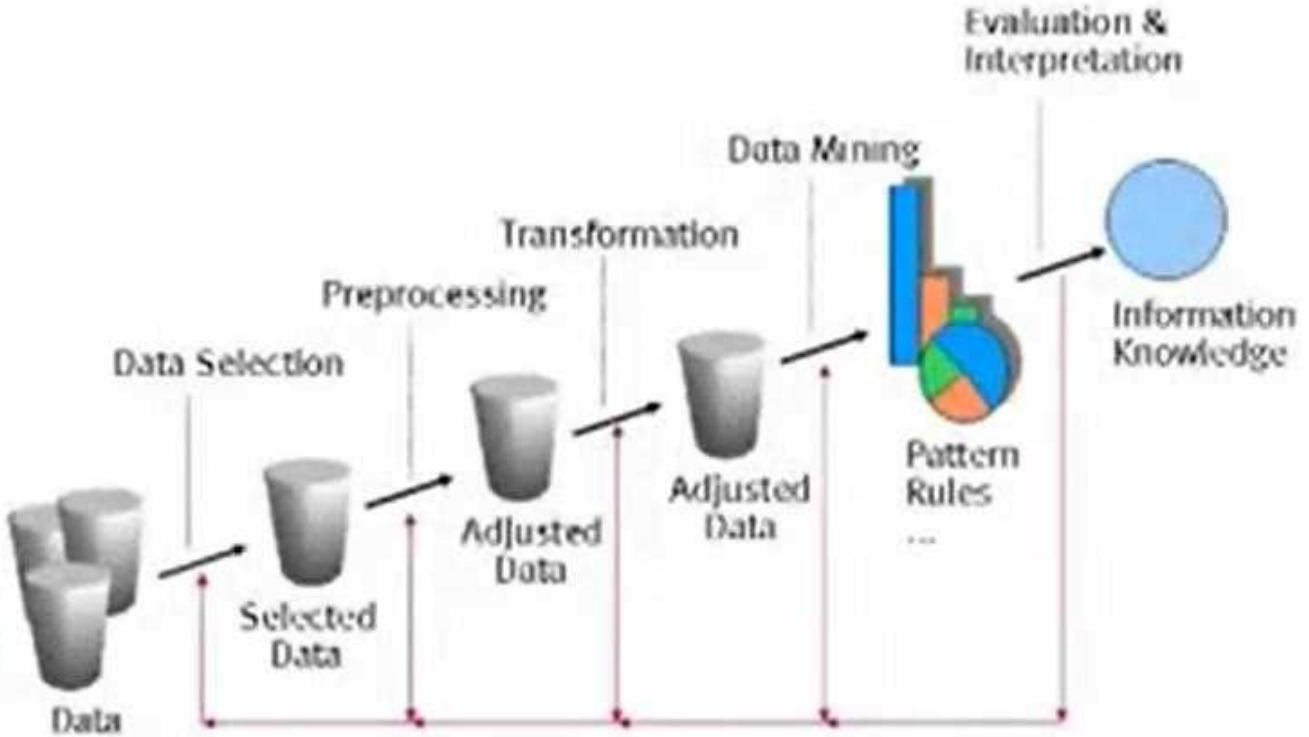
## CONCEPT MINING

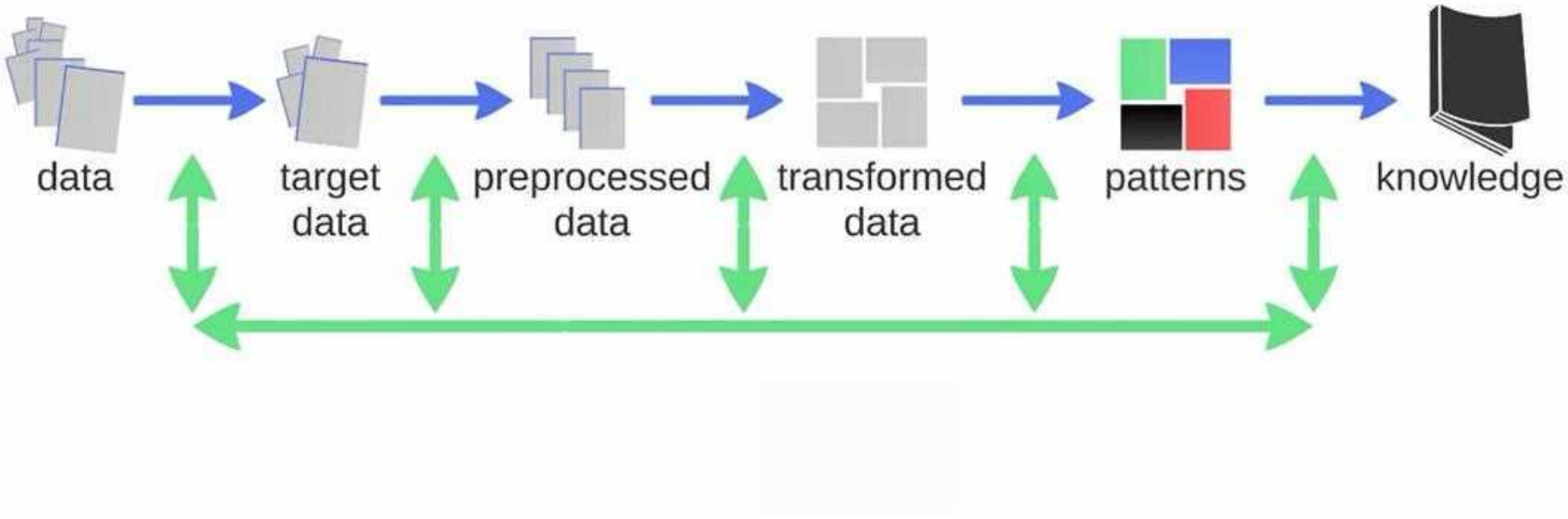
Concept mining is a process that focuses on extracting ideas and concepts found in documents.

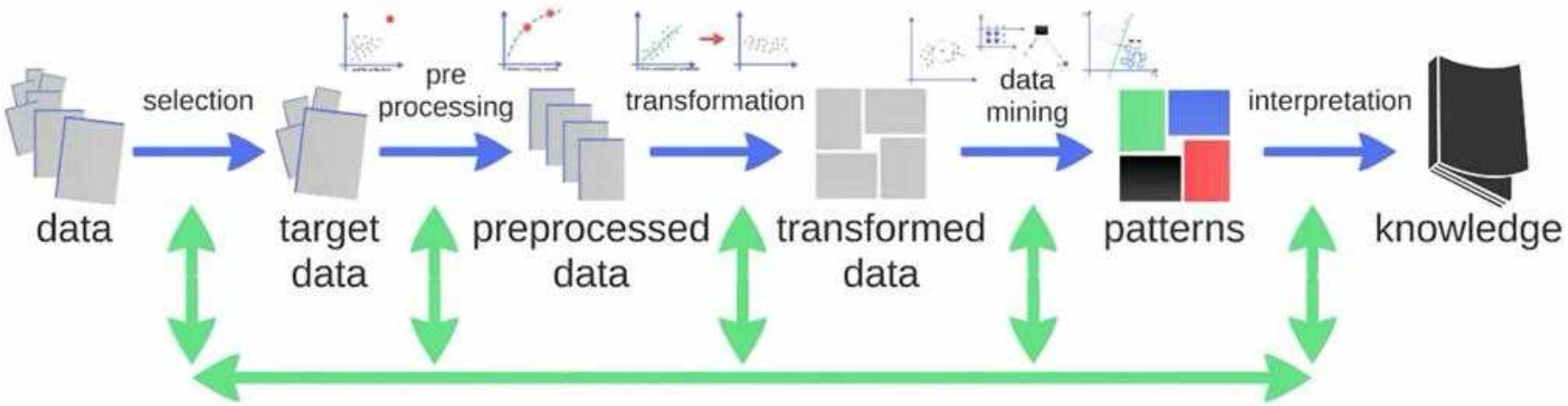
## DATA MINING MODELS

# Data Mining Process Model

### Data Mining Model

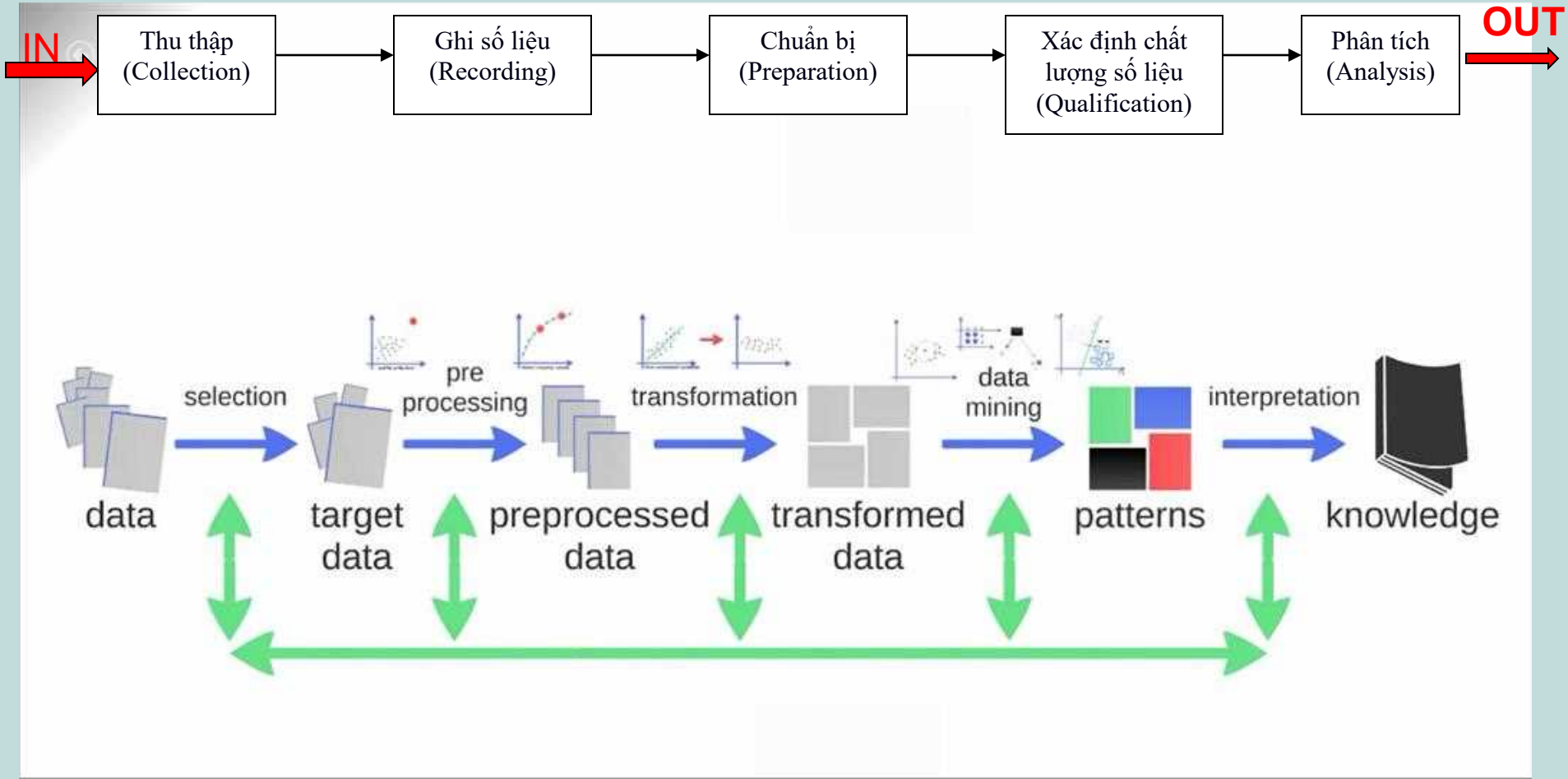



data
target data
preprocessed data
transformed data
patterns
knowledge

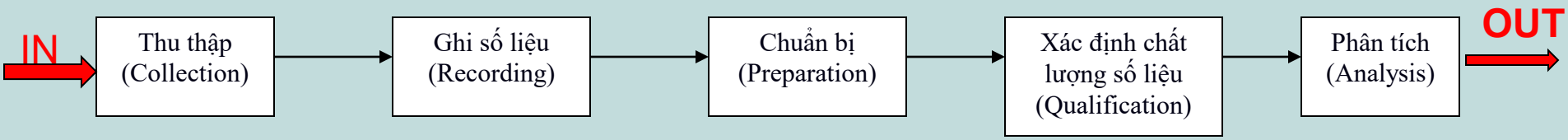

selection
pre
processing
transformation
data
mining
interpretation
data
target
data
preprocessed
data
transformed
data
patterns
knowledge

# 2.Các quy trình lấy mẫu

## Các bước khai thác dữ liệu



# 2.Các quy trình lấy mẫu



# 3.Quy trình lấy mẫu và trình bày dữ liệu

**Sai số:** diễn tả sự chính xác cuả phép đo,
biển thị khoảng cách giữa giá trị đo và giá trị thực

-Phân loại sai số

-+ do nguyên nhân: Sai số thô, sai số hệ thống, sai số ngẫu nhiên

-+ do cách đánh giá: ss tuyệt đối, ss tương đối

-Cách tính SS

-+ phép đo trực tiếp: SS ngẫu nhiên, SS dụng cụ, SS tổng hợp
-+ SS gián tiếp:
-+ SS dụng cụ
-+ SS từ mo^ hình

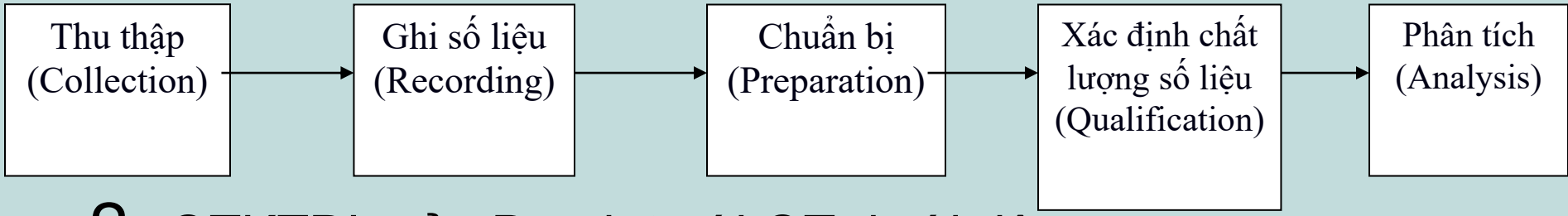
-- Cách làm tròn SS
-- Cách biểu diễn đồ thị

Phần mềm sử dụng:

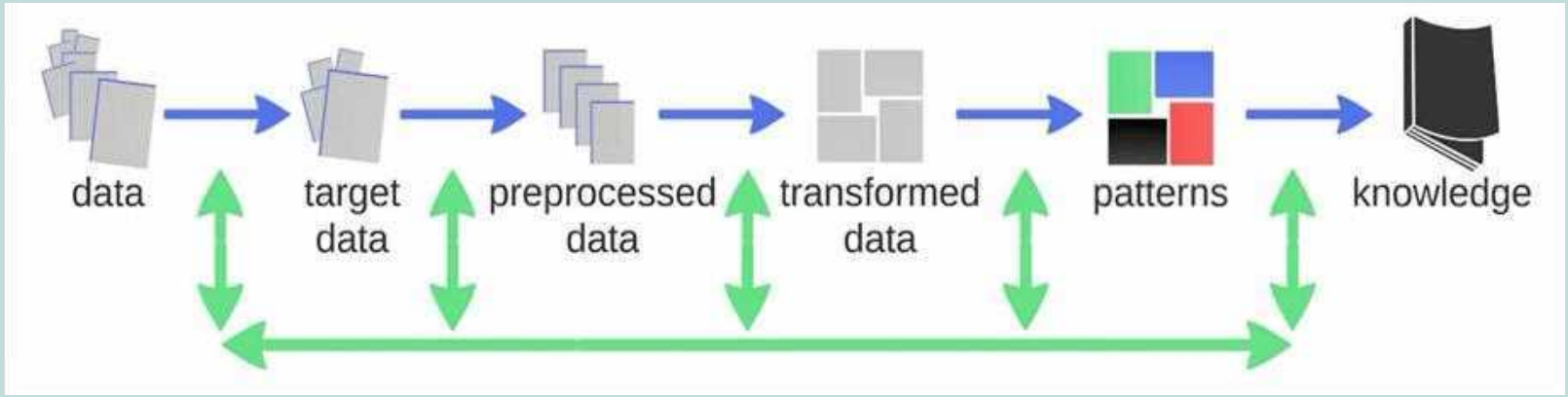
- Excel
-Grapher
-Sufer
-Mapper

# Bài tập (15 phút)

1. Từ ví dụ trên, hãy phân tích theo quy trình khai thác dữ liệu theo Bendat




2. QTKTDL của Bendat với QT dưới đây:



# Bài chuẩn bị

Ôn tập về lý thuyết xác suất, thống kê:

+ GTTB, phương sai, Max, Min…

+ Hàm phân bố XS

+…

=> Viết lập trình

# Methods of Data Mining 1

| No | Student ID | Student' Name | Theory | | | Practice | | | Final Grade |
|---|---|---|---|---|---|---|---|---|---|
| | | | Progress | Final Exam | Total | Progress | Final Exam | Total | |
| 1 | 20210004 | Nguyễn Thị Kim Huệ | 8.7 | 6.5 | 7.2 | 6.4 | 7.5 | 6.8 | **7.0** |
| 2 | 20210014 | Nguyễn Lâm Nhật Quang | 8.6 | 6.5 | 7.1 | 8.6 | 7.5 | 8.2 | **7.6** |
| 4 | 20210026 | Trần Kiên Nhẫn | 6.8 | 7 | 7 | 7 | 7 | 7 | **7.0** |
| 3 | 20210034 | Bùi Minh Thiện | 9 | 8.75 | 8.8 | 9.5 | 8.5 | 9.1 | **9.0** |

20210034 - BUI MINH THIEN

CLASS EXERCISE 5

```
program btBlackTurkey
parameter (n=1500, m=100)
real a(0:n), tong, k(0:m) , uhr(0:m), k2(0:m) , tong2(0:m) ,s(m)
integer kk,r,i
open (1,file="thunguyen.txt")
do i=1,1500
read(1,*) a(i)
enddo

do r= 0,m
tong=0.
do i=1,n-r
tong = tong + (a(i)*a(i+r))
enddo
k(r)  = tong/(n-r)
enddo

do r=0,m
uhr(r) = 0.5*(1+cos((3.1415*r)/m) )
enddo

do r=0,m
k2(r) = k(r)*uhr(r)
enddo

do kk = 0,m
tong2(kk)=0
```

```
do r=1,m-1
tong2(kk) = tong2(kk) + k2(r)*cos((3.1415*r*kk)/m)
enddo
s(kk) =(0.25/3.1415) * (k2(0) + 2*tong2(kk) + k2(m)*cos(3.1415*kk))
enddo

do kk=0,m
write(*,*) s(kk)
enddo
```

PRACTICE

# Methods of Data Mining 1

2022 - 2023

# Methods of Data Mining 1

| LESSON | PRACTICAL CONTENT |
|---|---|
| **1-2** | DATA ANALYSIS USING SPSS SOFTWARE (online) |
| **3** | BASIC STATISTICAL FUNCTIONS |
| **4** | REMOVING THE TREND |
| **5** | WAVE SPECTRUM CALCULATION ( **FORTRAN AND MATLAB** ) |
| **6** | CALCULATION OF UNESCO FORMULAS ( **MS EXCEL/FORTRAN** ) |

# Methods of Data Mining 1

**POINT** Preparation → Classwork → Homework

## SOME NOTES AND REGULATIONS

- **Preparation:** Understand the theory and prepare the code in advance. If a student comes to class without preparation, the instructor will ask the student to leave.
- **In class:** Discuss questions/errors with the instructor.
- **Software/data:** The relevant software/data will be sent to the students, and they are encouraged to install it before coming to class.
- **Laptop:** Students are encouraged to bring their personal laptops.
- **Attendance:** Students are expected to attend class on time. If a student is late by more than 20 minutes (without a valid reason), the instructor will ask the student to leave.

**VNUHCM-UNIVESRITY OF SCICENCE**

# FINAL EXAMNINATION

## Practice Part

## Semester I – Academic year 2022-2023

## Time: 120 minutes - Use of documents

**Question 1**: (3đ) Find the coefficients b0, b1 for the equation describing the trend of the form: y = b0 + b1*x of the data series 'Dulieu1a.txt' in dimensionless and dimensional normalized form, then de-tide.

**Question 2**: (3đ) Calculate the spectrum of the data series 'Dulieu 1b.txt' using the Hanning filter window. Choose the appropriate m value (Δt=0.25 s).

$$w(r) = 0.5\left(1 + \cos\left(\frac{\pi r}{m}\right)\right)$$

**Question 3**: (2đ) Calculate special string data and appropriate distance to find the distribution of string data "Dulieu1c.txt".

**Question 4:** (2đ) Use the UNESCO formula to convert the performance number ('Dulieu1d.txt') to the data depth number, knowing the service measurement location in Can Gio, Ho Chi Minh City.

# EXAM SUBJECT: DATA ANALYSIS METHODS IN PHYSICAL OCEANOGRAPHY

## Time: 120 minutes - Use of documents

**Student:** BUI MINH THIEN

1. **Find the coefficients b0, b1 for the equation describing the trend of the form: y = b0 + b1*x of the data series 'Dulieu1a.txt' in dimensionless and dimensional normalized form, then de-tide.**

```
program de1
integer i
parameter (n=1500)
real a(n), y1(n), tb,b0, b1, b2, b3, y2(n),y3(n)
open (1,file="Dulieu1a.txt")
open(2,file="thunguyen.txt")
open(3,file="khongthunguyen.txt")
do i=1,1500
read(1,*) a(i)
enddo
sum=0
do i=0,1500
sum=sum+a(i)
enddo
tb=sum/n
do i=1,1500
tam=tam + (a(i)-tb)**2
enddo
DLC=(tam/(n-1))**0.5
do i=1,1500
y1(i)=(a(i)-tb)/DLC
enddo
do i=1,1500
t1=sum
t2=t2 + i*a(i)
enddo
b0=b0+((4*n+2)*t1-6*t2)/(n**2-n)
b1=b1+(12*t2-(6*n+6)*t1)/(0.25*n*(n-1)*(n+1))
do i=1,1500
y2(i) = a(i) - (b0+b1*i*0.25)
enddo
```

```
do i=1,1500
t3=t3+y1(i)
t4=t4+i*y1(i)
enddo
b2=b2+((4*n+2)*t3-6*t4)/(n**2-n)
b3=b3+(12*t4-(6*n+6)*t3)/(0.25*n*(n-1)*(n+1))
do i=1,1500
y3(i) = y1(i)- (b2+b3*i*0.25)
enddo
do i=1,n
!thunguyen
write(2,*) y2(i)
!khongthunguyen
write(3,*) y3(i)
enddo
end
```

**Comment:**

The above code has no syntax errors and the results of running the data file match the results of the given exercise. (3đ)

2. **Calculate the spectrum of the data series 'Dulieu 1b.txt' using the Hanning filter window. Choose the appropriate m value (Δt=0.25 s).**

$$w(r) = 0.5\left(1 + \cos\left(\frac{\pi r}{m}\right)\right)$$

```
program bai2
parameter (n=2048, m=1000)
real a(0:n), tong, k(0:m), k2(0:m), tong2(0:m), s(m)
integer kk, r,i
open(1, file ="dulieu1b.txt")
do i=1,1500
read(1,*) a(i)
enddo
do r=0,m
tong=0.
do i=1,n-r
tong=tong + (a(i)*a(i+r))
enddo
k(r)=tong/(n-r)
enddo
do r=0,m
k2(r)=k(r)*0.5*(1+cos((3.14*r)/m))
```

```
enddo
do kk=0,m
tong2(kk)=0
do r=1,m-1
tong2(kk) = tong2(kk) + k2(r)*cos((3.1415*r*kk)/m)
enddo
s(kk) = (0.25/3.1415)*(k2(0) + 2*tong2(kk)+k2(m)*cos(3.1415*kk))
enddo
do kk=0,m
write(*,*) s(kk)
enddo
end
```

**Comment:**

The above code has no syntax errors, but the live spectrum data results are not good. (1.5đ)

3. **Calculate special string data and appropriate distance to find the distribution of string data "Dulieu1c.txt".**

```
program bai3
integer i,j,h
parameter (r=20,h=300)
real max, min, aver,s,tam,dlc,ps,sum,z
real a(h), p(h), x(h)
real f(h)
open(unit=3, file="ketquab3.txt", status="unknown")
open (2, file="Dulieu1c.txt")
do i=1,h
read(2,*) a(i)
enddo
max=a(1)
do i=1,h
if(max<a(i)) then
        max=a(i)
endif
enddo
write(3,*) "MAX", max
min=a(1)
do i=1,h
if(min>a(i)) then
        min=a(i)
endif
enddo
write(3,*) "MIN", min
do i=1,h
```

```
s=s+a(i)
enddo
aver = s/h
write(3,*) "Trung binh", aver
do i=1, h-1
do j=i+1, h
if (a(i)<a(j)) then
tam =a(i)
a(i) = a(j)
a(j)=tam
endif
enddo
enddo
do i=1,h/3
t=t+a(i)
enddo
aver2=t/(h/3)
write(3,*) "trung binh 1/3 song lon nhat", aver2
do i=1,r+1
x(i)=0.006 + (i-1)*0.00575
enddo
do k=0,r+1
        p(k) = 0
do i=1,h
        if(a(i)>=x(k).and.a(i)<x(k+1)) then
        p(k)=p(k)+1
        endif
enddo
enddo
        p(20)=p(20)+1.
        write(3,*) "ket qua cua 20 khoang chia"
do i=1,r
write(3,*) p(i)
enddo
sum =0
do i=1,h
sum=sum+a(i)
enddo
tb=sum/h
do j=1,h
z=z+(a(i)-tb)**2
enddo
do j=1,h
```

```
PS = z/h
enddo
DLC = sqrt(ps)
do i=1,h
f(i) = (1./(DLC*sqrt(2*3.1415)))*exp((-((a(i)-tb)**2))/(2*PS))
enddo
write(3,*) "phan bo du lieu"
do i=1,h
write(3,*) f(i)
enddo
end
```

**Comment:**

The above code has no syntax errors, however, the data distribution results are not good. (2đ)

4. **Use the UNESCO formula to convert the performance number ('Dulieu1d.txt') to the data depth number, knowing the service measurement location in Can Gio, Ho Chi Minh City.**

```
program bai4
real p(50), d(50)
open(1,file="Dulieu1d.txt")
open(unit=2, file="dosau.txt", status="unknown")
do i=1,50
read(1,*) p(i)
enddo
c1=9.72659
c2=-2.2512E-5
c3=2.279E-10
c4=-1.82E-15
g10=9.780318*(1.0+(5.2788E-3)*sin(10.)**2 + (2.35E-5)*sin(10.)**4)
gm=2.184E-6
do i=1,50
d(i) = (c1*p(i) + c2*p(i)**2 + c3*p(i)**3 + c4*p(i)**4)/(g10+0.5*gm*p(i))
enddo
do i=1,50
write(2,*) d(i)
enddo
end
```

**Comment:**

The above code has no syntax errors, the data results are good. (2đ)

VNUHCM-UNIVESRITY OF SCICENCE

# FINAL EXAMNINATION

## Semester II – Academic year 2022-2023

**ARCHIVE CODE**
(written by ET&QA Office)

| | | | |
|---|---|---|---|
| **Course name:** | **Methods of Data Mining 1** | **Course code:** | **OMH10013** |
| Time: | **50 minutes** | Date: | |

Note: *Students are* [ ☐ *allowed* / ☒ *not allowed*] *to use materials during the examination*

**Question 1: (2.0 points)**

Compare and explain the significance of the statistical method and spectral method.

**Question 2: (3.0 points)**

What are the procedures (steps) for sampling in data mining according to Bendat?

In your opinion, which procedure is the most important? Why?

**Question 3: (2.0 points)**

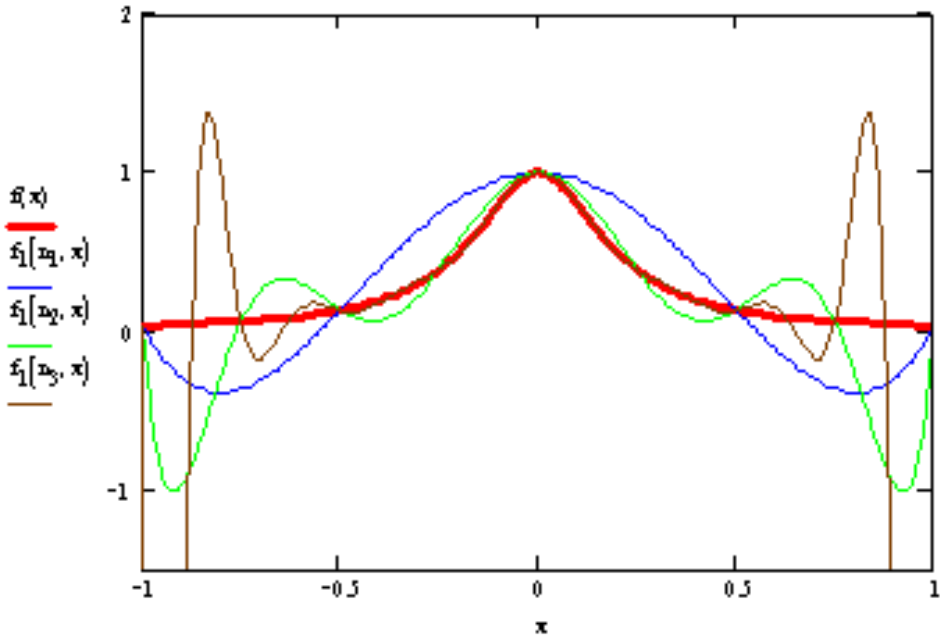
State the central limit theorem (CLT).

Explain the significance of this theorem in practical measurements?

**Question 4: (3.0 points)**

Please explain the meaning of Figure 1.

Analyze the pros and cons of these methods.




**----------end----------**

VNUHCM-UNIVESRITY OF SCICENCE
# FINAL EXAMNINATION
## Semester II – Academic year 2022-2023

**ARCHIVE CODE**
(written by ET&QA Office)

**Course name:** **Methods of Data Mining 1** **Course code:** **OMH10013**

Time: **50 minutes** Date:

Note: *Students are* [ ☐ *allowed /* ☒ *not allowed*] *to use materials during the examination*

**Question 1: (2.0 points)**

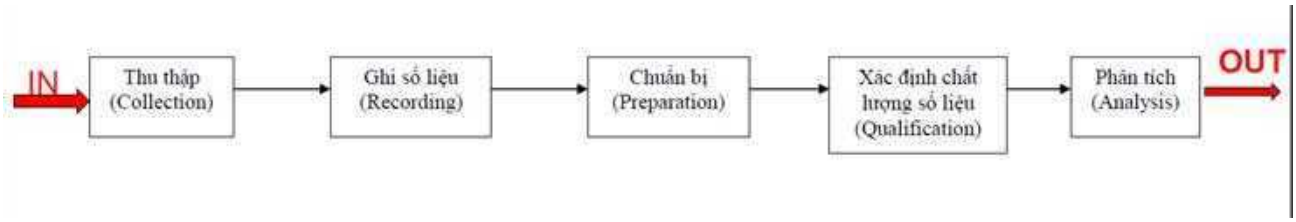
Compare and explain the significance of the statistical method and spectral method.

| Statistical methods | spectral method |
|---|---|
| characterize data in the time domain by calculating statistical parameters like mean, variance, standard deviation, etc. | characterize data in the frequency domain by determining the spectral density distribution. |
| help summarize and describe the overall properties of a data set. | reveal periodicities, cycles, and frequency patterns that may not be observable from statistics alone. |
| can be applied to any data set | best suited for cyclic or periodic data |
| measures of the total energy or variation in a data set | shows how that energy is distributed across different frequencies |
| lose time information | retains it by displaying energy vs. frequency. |
| | |

Together statistical and spectral analysis provide complementary time and frequency domain perspectives that give deeper insight into the data's true structure and behavior.

**Question 2: (3.0 points)**

What are the procedures (steps) for sampling in data mining according to Bendat?




In your opinion, which procedure is the most important? Why?

*Answer any step and explain acceptable reasons*

**Question 3: (2.0 points)**

State the central limit theorem (CLT).

*The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the actual distribution of the population from which the samples are taken.*

*In other words, as the sample size increases, the sampling distribution of the sample mean approaches a normal distribution, even if the original population is non-normal. The sample means vary randomly around the population mean with a normal distribution.*

Explain the significance of this theorem in practical measurements?

*Know how to collect the samples significant enough*

*It indicates that the arithmetic mean of a sufficiently large number of independent random variables will be approximately normally distributed. This is true regardless of their underlying distributions.*

*Many measurement processes produce random error that is from multiple independent sources. The central limit theorem states these errors will tend to follow a normal distribution when aggregated.*

**Question 4: (3.0 points)**

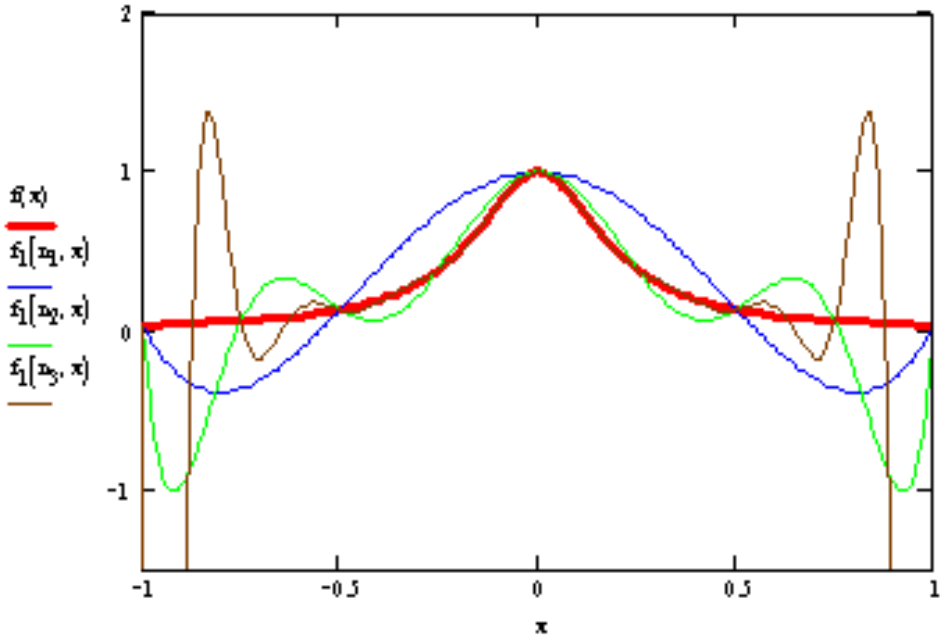
Please explain the meaning of Figure 1.

*Higher order, more accurate especially in the center*

Analyze the pros and cons of these methods.

***Lower order:*** *for unknown problems, safter but less accuracy*

***Higher order****: focus the exact point, neglected the surrounding (less acurracy), careful for application*




**----------end----------**

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

PHÒNG THI: .....

Học kỳ: I.. Năm học: 2022 - 2023

Tên học phần: Các phương pháp khai thác dữ liệu I

Mã học phần/Mã lớp: QM.H1.0015

Họ và tên sinh viên: Bùi Minh Thiện

Mã số sinh viên: 21D120039 Số thứ tự (theo danh sách dự thi) .....

Họ tên và chữ ký của cán bộ coi thi:

1) .....

2) .....

Số phách:

| Điểm số | Điểm chữ | Chữ ký của cán bộ chấm thi | Số tờ | |
|---|---|---|---|---|
| 8.75 | | Họ và tên: ..... | | Số phách |

Câu 1 >

* Phương pháp thống kê: Là phương pháp trình bày số liệu dưới dạng bảng, gồm nhiều thành phần và thuộc tính khác nhau.

Phương pháp phổ: Là phương pháp trình bày số liệu dưới dạng biểu đồ, mô tả chữ liệu theo miền tần số.

* So sánh

- Về tính chính xác: phương pháp thống kê thể hiện rõ ràng hơn thông qua số liệu cụ thể, độ chênh lệch và các sai số được kiểm soát. Còn phương pháp phổ khó thể hiện độ chính xác của số liệu

0.75

- Về tính trực quan: Phương pháp phổ thông qua đồ thị sẽ thể hiện các xu hướng thay đổi, khoảng cách chênh lệch một cách trực quan hơn, giúp người đọc có thể dễ dàng hình dung hơn so với dữ liệu số trong thống kê.

* Ý nghĩa

- Phương pháp phổ mang lại hiệu quả cao trong diễn đạt, là minh họa cụ thể cho các xu hướng, diễn biến của đối tượng và dễ dàng truyền tải ý nghĩa của nghiên cứu khoa học đến với mọi người

- Phương pháp thống kê cho phép kiểm tra, kiểm soát các đối tượng với độ chính xác cao. Là nguồn "dữ liệu gốc" thuận lợi cho việc lưu trữ và truyền tải trong thời gian dài.

9.25

- Cả hai phương pháp đều giúp hiểu sâu hơn về cấu trúc và đặc tính của dữ liệu. Thống kê mang tính tóm tắt các đặc tính liên quan còn phân tích phổ thể hiện dao động dữ liệu dưới dạng tần số.

THÍ SINH KHÔNG ĐƯỢC VIẾT VÀO PHẦN CÓ GẠCH CHÉO NÀY

Câu 2>

a) Quy trình lấy mẫu của Bendat

Thu thập số liệu → Ghi số liệu → Chuẩn bị → Xác định chất lượng

Phân tích ← số liệu

1.5

b) Theo em, quá trình xác định chất lượng số liệu là quan trọng nhất. Vì dữ liệu thu thập được có được phụ vụ trong các nghiên cứu hiện tại và trong tương lai hay không là do quá trình này quyết định.

1.5

Trong quá trình này, số liệu sẽ được so sánh và đối chiếu từ các nguồn khác hoặc quan trắc thực tế. Nếu các số liệu có sự chênh lệch nhất định mà không có sự biện luận hợp lý cũng sẽ bị loại bỏ.

Câu 3>

a) Định lý giới hạn trung tâm: Tổng của một lượng lớn các biến ngẫu nhiên độc lập với giá trị trung bình và phương sai hữu hạn có phân bố chuẩn.

1

b) Trong thực tế, định lý giới hạn trung tâm được ứng dụng trong lựa chọn mẫu (số mẫu lớn) phù hợp trong đo đạc, khảo sát để số liệu có ý nghĩa.

0.5

Câu 4>

- Thông qua Hình I, ta thấy khi đạo hàm bậc càng cao, giá trị ở vị trí trung tâm sẽ chính xác hơn (phổ dữ liệu sẽ hẹp hơn) so với các đạo hàm bậc thấp. Tuy nhiên, dao động nhiễu ở xung quanh sẽ nhiều hơn và khó xác định hơn.

1

- Ưu và nhược điểm.

Về độ an toàn, đạo hàm bậc thấp sẽ ổn định hơn (ít nhiễu.) nhưng về độ chính xác thì không cao. Đạo hàm bậc cao sẽ chỉ

1.5

THÍ SINH KHÔNG ĐƯỢC VIẾT VÀO PHẦN CÓ GẠCH CHÉO NÀY

tập trung đị chính xác ở vị trí trung tâm, các giá trị xung quanh càng xa trung tâm thì nhiễu càng cao.