

An Empirical Study on Humans' Capability of Information Reconstruction



Natchaya Kijmongkolchai
Kellogg College
University of Oxford

A thesis submitted for the degree of
Master of Science in Computer Science

August 2016

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my project supervisor, Professor Min Chen, for many hours of discussion and guidance throughout the project timeline. Special thanks are due to Alfie Abdul-Rahman, a postdoctoral research associate of Professor Min Chen, whose advice and help have been so helpful. I would like to thank my supervisor Associate Professor Jonathan Barrett, who has guided me throughout the three terms of my study. I acknowledge the support of The Royal Thai Government for their financial support throughout the academic year. I would like to thank all of my friends, colleagues, and other participants in my user study for their time and efforts. I must also take this opportunity to thank my friends, especially Tuanta, Lakwadee, Kelvin, and Chayathorn who have made my stay in Oxford such a pleasant one. Special thanks are also due to Dr Sivadon and Dr Krissada for their guidance from the other side of the world. Lastly, I would like to thank my parents, my sister, and my boyfriend Warut for their encouragement and moral supports. I could never have managed without you.

Abstract

Cost-benefit analysis proposed by Chen and Golan [CG16] is a theoretic measure that allows for an objective comparison of human-centric and machine-centric processes in data intelligence. While this provides a fundamental theoretic measurement, it has not been formally assessed, for example, by an empirical study. In this project, we provide the first assessment of this cost-benefit measure through a laboratory-based empirical study. We evaluate different types of information and compare the effects they have on a reconstruction quality, in terms of accuracy and response time. Three categories of information used in this study are statistical measure, time series type, and time series pattern, representing, respectively, machine-processed information, human soft knowledge through memory recall, and human soft knowledge for pattern recognition. The results show that machine-processed information alone is not good enough for reconstruction quality. By incorporating human soft knowledge through memory recall or pattern recognition, or both, human participants yield significantly higher performance. Additionally, given any type of information yields significantly better performance than a random guess. Our study is the first empirical study that formally assessed the cost-benefit measure, which is fundamental theoretic measurement in the visualisation community.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Objectives	2
1.3	Structure	3
2	Background	5
2.1	Cognitive Psychology	5
2.1.1	Memory	6
2.1.2	Object Recognition	6
2.1.3	Pattern Recognition	7
2.1.4	Categorisation	7
2.1.5	Visual Reconstructability	8
2.2	Information Theory	8
2.2.1	Visualisation and Information Theory	9
2.2.2	Alphabets	9
2.2.3	Alphabet Transformation	10
2.2.4	Cost-Benefit Analysis	10
2.3	Time Series	12
2.3.1	Time Series Properties	12
2.3.2	Time Series Patterns	13
2.3.3	Time Series Decomposition	14
2.3.4	Time Series Type	15
2.4	Empirical Study	16
2.4.1	Research Questions and Hypotheses	17
2.4.2	Variables	17
2.4.3	Bias Caused by Nuisance Variables	18
2.4.4	Bias in Within-Subjects Design	18
2.4.5	Statistical Analysis	19

2.4.6	ANOVA	20
2.4.7	<i>t</i> -test	21
3	Methodology	23
3.1	Research Question and Hypotheses	23
3.2	Task	25
3.3	Variables in the Experiment	25
3.3.1	Control Variables	26
3.3.1.1	Visualisation Image	26
3.3.1.2	Level of Distractor Difficulty	26
3.3.1.3	Position of the ‘Next’ Button	26
3.3.2	Independent Variables	26
3.3.2.1	Statistical Measure	27
3.3.2.2	Time Series Type	27
3.3.2.3	Time Series Pattern	30
3.3.3	Dependent Variables	35
3.3.3.1	Accuracy	35
3.3.3.2	Response Time	35
3.4	Measurement Metrics	35
3.4.1	Objective Measure	35
3.4.2	Subjective Measure	35
3.5	Techniques for Analyses	36
4	User Study Design	39
4.1	Design Overview	39
4.1.1	Task Design Overview	39
4.1.2	Trial Design Overview	40
4.1.3	Stimulus Design Overview	40
4.1.4	Software Design Overview	40
4.2	Task Design	41
4.3	Trial Design	42
4.4	Stimulus Design	43
4.4.1	Data Rule Design	44
4.4.1.1	Virtual Axes	44
4.4.1.2	Reliable Data Source	45
4.4.2	Distractor Rule Design	46
4.4.3	Visualisation Image Design	48

4.4.3.1	Information Screen	48
4.4.3.2	Time Series	49
4.5	Software Design	50
4.5.1	Software Workflow	50
4.5.2	Sequence Design	51
4.5.3	Time Design	51
5	Implementation	53
5.1	Implementation Process	53
5.2	Stimulus Generation	54
5.2.1	Data Gathering	54
5.2.2	Visualisation Image Generation	54
5.2.3	Stimulus Generation Iteration	55
5.3	Software Development	56
5.3.1	Software Development Iteration	57
5.4	Experiment	58
5.4.1	Participant	58
5.4.2	Apparatus	59
5.4.3	Procedure	59
6	Result Analysis	63
6.1	Result Summary	63
6.2	Result Analyses for Statistical Measure	66
6.2.1	Accuracy	66
6.2.2	Response Time	69
6.2.3	Performance Summary	69
6.2.4	Difficulty Rating	69
6.3	Result Analyses for Time Series Type	70
6.3.1	Accuracy	70
6.3.2	Response Time	71
6.3.3	Performance Summary	73
6.3.4	Difficulty Rating	73
6.4	Result Analyses for Time Series Pattern	74
6.4.1	Accuracy	74
6.4.2	Response Time	75
6.4.3	Performance Summary	76
6.5	Further Analysis	77

6.5.1	Pairwise Combination of Time Series Type	77
6.5.1.1	Accuracy	77
6.5.1.2	Response Time	79
7	Conclusion	81
7.1	Summary	81
7.2	Evaluation	83
7.3	Future Work	84
A	Stimuli in the Study	85
B	The Experiments	101
	Bibliography	103

List of Figures

2.1	Visual Object Recognition Model [HB89].	7
2.2	General data processing workflow [CFV ⁺ 16].	11
2.3	Time series exhibiting different patterns [HA14].	14
2.4	Different time series types.	16
3.1	Selected time series type for this study.	29
3.2	Hierarchy of performance analyses.	37
4.1	Trial structure.	42
4.2	Information screen.	43
4.3	Question screen.	43
4.4	Time Series Stimuli.	45
4.5	Distractors.	47
4.6	Information screen theme	48
4.7	Time series example	49
4.8	Software workflow	50
5.1	Demographics information including age and gender	61
5.2	Familiarity rating for time series	61
6.1	Overall average accuracy.	64
6.2	Overall average response time.	65
6.3	Average accuracy of statistical measure.	66
6.4	Average accuracy of statistical measure (binary encoding).	67
6.5	Average accuracy of statistical measure (statistical measure only).	68
6.6	Average response time of statistical measure.	68
6.7	Participants' difficulty rating for statistical measures.	70
6.8	Average accuracy of time series type.	71
6.9	Average accuracy of time series type (binary encoding).	72
6.10	Average accuracy of time series type (time series type only).	72

6.11	Average response time of time series type.	73
6.12	Participants' difficulty rating for time series type.	74
6.13	Average accuracy of time series pattern.	75
6.14	Average accuracy of time series pattern (binary encoding).	75
6.15	Average accuracy of time series pattern (time series pattern only). . .	76
6.16	Average response time of time series pattern.	76
6.17	Average accuracy of pairwise type.	77
6.18	Average accuracy of pairwise type (binary encoding).	78
6.19	Average accuracy of pairwise type (time series type only).	78
6.20	Average response time of pairwise type.	79
A.1	Demographic questions	85
A.2	Question 1	86
A.3	Question 1	86
A.4	Question 2	87
A.5	Question 2	87
A.6	Question 3	88
A.7	Question 3	88
A.8	Question 4	89
A.9	Question 4	89
A.10	Question 5	90
A.11	Question 5	90
A.12	Question 6	91
A.13	Question 6	91
A.14	Question 7	92
A.15	Question 7	92
A.16	Question 8	93
A.17	Question 8	93
A.18	Question 9	94
A.19	Question 9	94
A.20	Question 10	95
A.21	Question 10	95
A.22	Question 11	96
A.23	Question 11	96
A.24	Question 12	97
A.25	Question 12	97

A.26 Question 13	98
A.27 Question 13	98
A.28 Question 14	99
A.29 Question 14	99
A.30 Question 15	100
A.31 Question 15	100
B.1 Pre-study presentation.	101
B.2 Experiment.	102

Chapter 1

Introduction

1.1 Motivation

In the era of data deluge, an enormous amount of data is readily available for collection and analysis. Our ability to capture this large volume of data has exceeded our ability to process and analyse it. Insights gained from analysing such a data are usually an essential part of an informed decision-making process. Thus, as the information overload problem continues to escalate, research studies in different areas have been conducted in an effort to address this problem. Similarly, different data processing techniques start emerging in an attempt to improve the process of understanding and analysing large amount of complex data. Such an improvement are believed to ultimately provide useful insights from the data in a timely manner.

Two general data processing approaches that are the center of research interest are *machine-centric approach* and *human-centric approach*. Research on automated data processing techniques, or machine-centric processes, has been the main focus of data processing optimisation since the start of data deluge. Scholars have been successful in establishing better statistical models and better machine learning algorithms. Both techniques are examples of fully-automated data analysis tools that are capable of extracting useful structures from large volumes of data. Although fully-automated data analysis tools are considered as a good solution for well-defined problems or known models, such a situation barely exist: most problems nowadays are complex and the model is typically unknown. Another problem with this analysis technique is the lack of communication between the automated data analysis system and human, which may lead to additional cost of search when problem arises later in the analysis process. Therefore, human intelligence is required in data processing optimisation [TC05].

Scholars have agreed on the fact that that human visual perception of patterns is the key to the sense-making process [HB89, AMST11, LS96, MGSP08, CG16]. Our ability to visually encode information and form meaningful patterns allow us to make sense of that information and gain insights in a manner that may not be possible if the data were examined in any other way. The field of visualisation integrates human-centric processes to helps saving time [CFB14] in the process of analysing the data by taking full advantage of the aforementioned visual perception power of human.

In practice, it is usually difficult to quantitatively capture knowledge derived from these human-centric processes, which results in imprecise or incomplete evaluation of such processes and pose difficulty when performing comparison with machine-centric processes. Chen and Golan [CG16] proposed an information-theoretic abstraction, a metric that enables the evaluation the cost-benefit of human-centric processes in a visualisation workflow.

To our knowledge, our research is the first empirical study to formally assess this cost-benefit measure, a fundamental theoretic measure in the field of visualisation.

1.2 Objectives

The goal of this project is **to evaluate human’s capability of information reconstruction when given different types of knowledge**.

The objectives of the project includes:

- To study concepts of object recognition and categorisation in cognitive psychology, cost-benefit analysis [CG16], time series, and empirical study methodology in general.
- To formulate the hypotheses of the empirical study, identify variables in the experiment, indicate measurement metrics, and specify result analysis techniques.
- To design the task, the stimuli, and the software for the empirical study, and implement them accordingly.
- To conduct the empirical study.
- To analyse the result of the empirical study.

Empirical study will be used to evaluate the quality of humans’ capability of reconstruction when given different kinds of information.

1.3 Structure

The dissertation contains seven chapters, including this introductory chapter describing the motivation behind the study, the objective of the study, and the structure of this dissertation.

Chapter 2 will introduce background information on cognitive psychology, information theory, time series, and empirical research methodology.

Chapter 3 will provide the details on the methodology used in this research, including the hypothesis of the study, the variables in the study, the measurement metrics, and the technique for result analysis.

Chapter 4 will discuss the user study design, including the stimulus design and the software design.

Chapter 5 will present the implementation of the study, including the stimulus creation, software development, and the experiment of the empirical study.

Chapter 6 will present the analysis of the empirical studys results, including the summary and the detailed analyses.

Chapter 7 will summarise the researchs results, projects evaluation, and suggestions for future works.

The appendices will provide all stimuli used in this study, sample interface of the software, sample pictures of the experiment, data table for the result analyses, and the project schedule.

Chapter 2

Background

The extraordinary human ability to see and recognise patterns is a gift that we should not ignore. Human has a visual power to detect and make sense of patterns from data since birth. These patterns are kept inside the memory and can be retrieved back later. Regardless of our cognitive power, it is very difficult to quantitatively capture these human cognitive abilities. Hence, concepts in information theory are needed in order to capture them.

This chapter begins by providing a background in cognitive psychology that is related to the concept of information reconstruction. The discussion then moves to the concept of information theory, which is closely related to the field of visualisation, and is used as a measurement metric for evaluating these human cognitive processes [CG16]. Then an explanation on time series, which is the type of visualisation image used in this study, follows. The chapter ends with an empirical research methodology, which is used in this project.

2.1 Cognitive Psychology

Cognitive psychology is the scientific study of how human process information. It is essentially the study of how human treat information (stimuli) that comes in and how the treatment leads to responses. Interestingly, cognitive psychology views human as an information processor, which is similar to the way computer takes in information as an input to a program and produce an output. Cognitive psychologists use this *computer metaphor* to make comparison between human minds and computers in order to understand how human process information. Some of the areas of study in cognitive psychology includes: *memory*, *object recognition*, *pattern recognition* and *categorisation*.

2.1.1 Memory

There have been arguments over the validity of the two theories on memory, Reappearance hypothesis and Reconstruction hypothesis [AB14]. *Reappearance hypothesis* claims that each experience is recorded in memory as a separate trace. *Reconstruction hypothesis*, supported by Bartlett [BB33] and Neisser [Nei67], states that one does not recall objects because traces of them exist in the mind, but after an elaborate process of reconstruction (which usually makes use of relevant stored information) [Nei67]. Anderson and Bower [AB14] have justified the validity of both hypotheses in their work. Thus, it is plausible the human memory is stored as a trace and by remembering, the trace is brought back, possibly using relevant information.

When we experience an event, a unique network of neurons is activated. Those neurons may have different functions such as facial recognition, motion detection, and so on. Yet, they are tied together as a network and such a network must be unique or else we would perceive the events as identical. Levitin [Lev14] explained the act of remembering as the process of bringing back on line the neurons that were involved in the original event. However, because memory is imperfect, the *replay* of the event usually have lower resolution than the actual event.

Remembering yields even lower quality of the replay when the original event contains nothing extraordinary from the usual routine. On the other hand, when something peculiar happens that is very different from the routine, we tend to remember it better. An explanation on this phenomenon is the fact that similar memories merge together into just a generic impression of the event. Levitin [Lev14] concluded this as the process of extracting abstract rules that tie events together, and that the brain creates a generic mixture of similar events, which can be referred to as a *category*. The next subsection describes the process of object recognition, which is based on the concept of categorisation.

2.1.2 Object Recognition

Multiple literature [HB89, AMST11, LS96, MGSP08] have pointed out that human visual recognition of objects is a gift we cannot ignore. *Visual object recognition* is a process of matching an object's description from an image with the stored representations of the characteristic of various kinds of objects [HB89].

Object recognition may have different levels of description. We recognise an object because we have *visual* knowledge of an object's shape, which describe how the object would look, even from different perspectives. We also possess *semantic* knowledge of

an object's function and associated objects, as well as *verbal* knowledge of the object's category name [HB89]. This process is illustrated in Figure 2.1

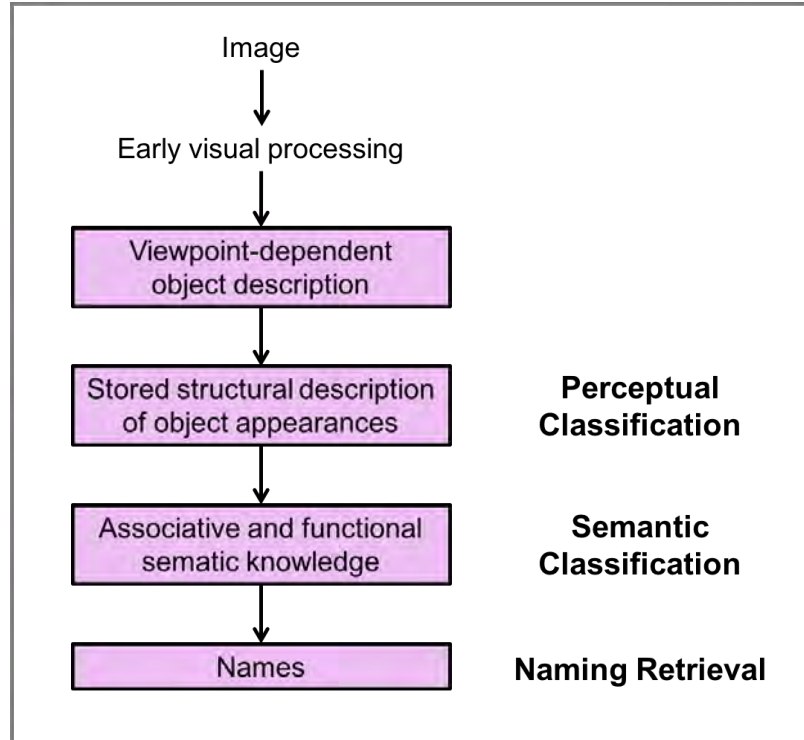


Figure 2.1: Visual Object Recognition Model [HB89].

2.1.3 Pattern Recognition

Object recognition, in a sense, is related to three-dimensional object as opposed to *pattern recognition*, which deals with two-dimensional pattern [HB89]. Two models were proposed as to explain how pattern recognition works [LN13]. *Template-based* pattern recognition is carried out by comparing a pattern of interest against a set of stored *templates* in the long-term memory. This pattern matching process involves seeing which of the templates have the most overlapping pattern with the pattern of interest. *Feature-based* pattern recognition, on the other hand, looks at the local *features* rather than the global shape. It breaks down the pattern into smaller parts and identify the pattern based on these local features.

2.1.4 Categorisation

Object recognition involves matching stored representations of characteristics of objects. These representations are in fact the *categories* that the brain creates. As

the brain has a limited capacity of remembering information, it automatically groups like objects together into a category [Lev14]. The term *categorise*, hence, means to consider an object of interest as equivalent to other things in the category and different from things that are not in the category. By categorising, we eliminate things that do not belong in the category, using the available visual, semantic, and verbal descriptions in the elimination process.

2.1.5 Visual Reconstructability

Jänicke *et. al.* [JWC⁺11] defined the act of visualise as a process that extracts information from the data and constructs a visual representation of that information. They further explained that there are three stages involved in visualising: (i) visualisation image is created, (ii) the image is viewed by human (iii) a *mental image* about the data is created in human mind. *Reconstruction* is a process of using the given features in the data to recreate a visualisation image.

In the process of making sense of the visualisation image, several human-centric processes are involved in order to support a viewer in reconstructing the data. Jänicke *et. al.* [CJ10] provided some examples of perceptual and cognitive processes including, “feature extraction and object identification, the formation of an abstract representation in the memory, and semantic reasoning based on one’s knowledge and experience [JWC⁺11]”. This is consistent with the visual object recognition model illustrated in Figure 2.1.

Levitin [Lev14] also used the concept of mental image to explain the process of categorisation. As he pointed out, “We bring to the mind a mental image of what we’re looking for, and neurons in the visual cortex help us to image in our mind’s eye what the object look like.” The semantic reasoning based on knowledge and experience is used to eliminate all things that are unlikely or irrelevant to the thing we are looking for in order to facilitate the search of relevant knowledge. The resulting mental image can be in the form of lower-resolution or incomplete version of the original experience. Levitin’s suggestion on human’s cognitive ability of categorisation explains the process of feature extraction and object identification [Lev14].

2.2 Information Theory

Information theory, a discipline introduced by Shannon [Sha01], integrates different *machine-centric components* such as statistics, algorithms, and communication

theory to study the limitations in the process of data compression and data transmission [CT12]. One of the most important concepts in the field of information theory is *Shannon’s entropy*.

2.2.1 Visualisation and Information Theory

Visualisation shares a wide range of conceptual connections with the field of information theory. Chen and Jänicke [CJ10] have shown that information theory, which contains a large number of existing concepts, can be used to explain different phenomena in visualisation. Yet, one aspect of visualisation that differs from information theory is the involvement of *human-centric components*. Chen and Golan use the term ‘soft’ knowledge in their recent paper [CG16] to describe knowledge derived from the human-centered processes, such as intuition, belief, and value judgment. They emphasised the importance of incorporating ‘soft’ knowledge as part of an information-theoretic measure for visualisation processes.

In practice, it is usually difficult to quantitatively capture ‘soft’ knowledge, and this results in imprecise or incomplete estimation of the cost-benefit ratio, which is a cost function for optimising a visualisation process. Chen and Golan [CG16] proposed an information-theoretic abstraction of data spaces as alphabets, and an abstraction of data processing as alphabet transformation, which are used as an *information-theoretic metric* for evaluating the cost-benefit of processes in a data analysis and visualisation workflow.

2.2.2 Alphabets

In information theory, let Z be a variable, the set of all valid values $\mathbb{Z} = \{z_1, z_2, \dots, z_m\}$ is called an *alphabet*, and each of its member z_i is called a *letter*. Each letter has an associated probability $p(z_i)$, which is used to calculate the *entropy* $H(Z)$, a concept introduced by Shannon [Sha01]

$$H(Z) = -\sum_{i=1}^m p(z_i) \log_2 p(z_i).$$

Shannon’s entropy can be interpreted as the average uncertainty of the variable Z . This uncertainty identifies the amount of information contained in a variable. Thus, Shannon’s entropy implies that information and uncertainty are two equivalent concepts. In an information-theoretic abstraction [CG16], an alphabet is an abstraction of a *data space*.

2.2.3 Alphabet Transformation

In the present where an enormous amount data is flooding in at every second, the alphabet of a raw data may contain numerous letters. Chen and Golan [CG16] uses an example of a share price time series to illustrate this. Consider a time series that records a share price every 5 seconds for an hour. The time series would consist of $60/5 \times 60 = 720$ data points. Assuming that a share price unit is USD \$0.01, and we use a resolution of 32-bit unsigned integers, each data point can take the value between $[0, 2^{32} - 1]$. If the probability of different time series were uniformly distributed, the entropy of this alphabet would be $23040 = 720 \times \log_2(232)$ bits. This is the *maximal entropy* of this alphabet.

In reality, the actual entropy is much lower than 23040 because some of the possible instances of a share price time series are very unlikely. For example, a share price time series with very high volatility between each data point is very rare. Sometimes, a single time series is insufficient in the decision-making process. If we need r time series in such a process, the maximal entropy would increase to $23040r$.

Chen and Golan [CG16] described different mechanisms to *reduce the maximal entropy* by *eliminating the highly improbable letters from the alphabets*. From the share price example, we would like to eliminate the unlikely time series such as that with extremely high volatility. This can be done by carefully considering and making decisions based on the information derived from machine-centric processes and human-centric processes.

Machine-centric information is obtained by computationally processing the data using mathematical or statistical measures. Examples include minimum value, maximum value, average value, standard deviation, cross-correlation index, *etc.* *Human-centric information* is gained from ‘soft’ knowledge such as past experience, known theories, intuition, belief, and value judgment. A viewer can identify certain features as a new set of variables from the observed data. This is called feature recognition, and it is a form of categorisation in cognitive science [Lev14] as described in Section 2.1.4. In an information-theoretic abstraction, data processing as *alphabet transformation*.

2.2.4 Cost-Benefit Analysis

A visualisation process can be regarded as an abstraction of a *general data processing workflow*, which consists of multiple intermediate processing steps as illustrated

in Figure 2.2 [CFV⁺16]. The workflow transforms data into decisions, in the form of observation, judgment, or hypothesis.

The abstractions of data spaces and data processing into alphabets and alphabet transformation enables cost-benefit analysis while taking into account ‘soft’ information introduced by human.

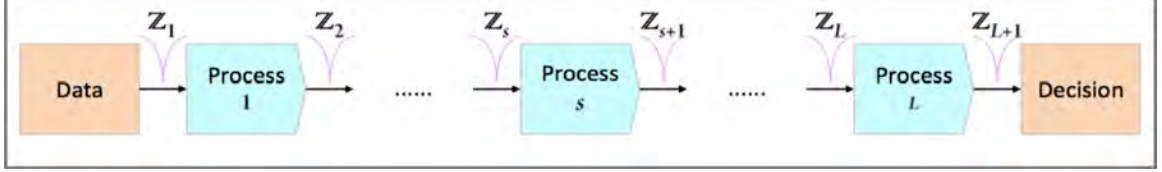


Figure 2.2: General data processing workflow [CFV⁺16].

Suppose \mathbb{Z}_s and \mathbb{Z}_{s+1} be two sets of alphabets where \mathbb{Z}_{s+1} is the alphabet that results from the process s as shown in Figure 2.2. We have a *forward mapping function* F_s

$$F_s : \mathbb{Z}_s \rightarrow \mathbb{Z}_{s+1}.$$

Now if we consider the reverse function of F_s , namely G_s that guess the original data value, we have a *backward mapping function* G_s , where \mathbb{Z}'_s is the impression of the original data value. In fact, this function G_s is a *reconstruction function* [CG16, JWC⁺11]

$$G_s : \mathbb{Z}_{s+1} \rightarrow \mathbb{Z}'_s.$$

An example of data processing workflow that includes a forward and a backward mapping functions can be summarised as follow. First, a forward mapping function F_s extracts a mean value from a time series. Here, the mean value is an output of an analytical process on time series data. Now consider its reverse function, G_s . An analyst gains an impression about the time series based on the given mean value. This is equivalent to a mapping from the given mean value to the impression of the time series. We call the process of backward mapping a *reconstruction process*.

Alphabet Compression Ratio (ACR) measures the level of reduction of data space at each data processing stage along the data processing workflow.

$$\Psi_{ACR}(F_s) = \frac{H(\mathbb{Z}_{s+1})}{H(\mathbb{Z}_s)}.$$

Potential Distortion Ratio (PDR) makes use of *Kullback-Leibler* divergence (relative entropy) to calculate the deviation of the impression \mathbb{Z}'_s from \mathbb{Z}_s

$$\Psi_{PDR}(F_s) = \frac{D_{KL}(Z'_s \parallel Z_s)}{H(Z_s)}.$$

The two metrics, *ACR* and *PDR*, can be used to evaluate the quality of the information reconstruction.

2.3 Time Series

A *time series* is a sequence of numerical data points in successive order, where the data points are recorded at regular intervals. A time series is used to represent time-oriented data [AMM⁺07], such as financial data, medical data, and meteorological data. It is denoted mathematically as $\{y_t\}_{t=1}^T$, where y_t is a time series, T is the time span, and t is the time that the data point is collected, ranging from 1 to T . A time series can be represented using a *Cartesian coordinate system* or a *Polar coordinate system*. However, a recent work by Adnan *et.al.* [AJB16] suggested that time series that are based on Cartesian coordinate system are generally more perceptible than Polar coordinate system. Hence, the Cartesian coordinate system will be used to represent time series throughout this study, and the rest of this section will describe different aspect of *only* this type of representation.

A time series that is based on Cartesian coordinate system can be thought of as a line chart, in which each data point is drawn chronologically. This kind of visual representation can reveal several patterns in the data over a period of time. In the following subsections, basic properties of time series, typical patterns that are observable in time series, time series decomposition, and some of the types of time series will be introduced.

2.3.1 Time Series Properties

Certain mathematical measures can be used to give a basic description of time series. These include frequency, time span, mean, variance, and covariance [KC⁺14].

Frequency is the time span between y_t and y_{t+1} . Each data point can be collected yearly, quarterly, monthly, weekly, daily, hourly, and at a greater frequency. Yearly frequency of data refers to collecting one data point each year. Quarterly frequency data means collecting one data point each quarter, and so on.

Time span is the period of time that all the data points are collected. It is in fact the number of observation times the frequency, and is denoted as T .

Mean is denoted as $t = E(y_t)$. The function $E(x)$ is the expected mean of a discrete random variable x . For each observation in a time series, a mean is defined. Thus, with T observations in a time series, there are T means defined.

Variance is written as $var(y_t) = E[(y_t - \mu_t)^2]$, where μ is the mean. It is also defined for each observation in a time series.

Covariance is defined as $cov(y_t, y_t - s) = E[(y_t - \mu_t)(y_t - s, \mu_t - s)]$. It is defined for each time t and for each time difference s . In a time series with T observations, there are $T^2 - T$ covariances defined. Note that half of these will be symmetrically equal.

2.3.2 Time Series Patterns

The most typical patterns that are found in time series data include autocorrelated pattern, trend, seasonal pattern, cyclic pattern, and irregular pattern [MJK15, HA14, KC⁺14, Bri01]. A pattern can occur on its own or in combination with other patterns as well.

Positively autocorrelated pattern can be detected when a value above a long-run average tends to be followed by other values above the average. Similarly, a value below the long-run average tends to be followed by other values below the average. In general, autocorrelation refers to the correlation of a time series with the previous and the future values.

Trend refers to a slope in a time series. This is usually in a long-term period. A trend can occur in a linear manner where the slope is constant, or in a quadratic manner where the trend is exponential.

Seasonal pattern is observable when a seasonal variations influence a time series to behave in a certain way, mostly rise and fall. Seasonal variations can take effect, for instance, at certain month of a year, a day of a month, a day of a week, or an hour of a day. Seasonality occurs at a fixed or known period.

Cyclic pattern exists when data exhibit fluctuation, by rising and falling, at a longer period than seasonal pattern. The period in cyclic pattern, unlike in seasonal pattern, is not fixed.

Irregular pattern is not very predictable and does not show strong trend, seasonality, or cyclic behaviour. Irregular pattern often occurs as random fluctuations, which is commonly seen in financial data. The unpredictability of this pattern makes it difficult to develop an accurate forecasting model.

From Figure 2.3, the monthly sales (top left) exhibit seasonality within each year and cyclical pattern in 6-10 years. Also, positive correlation can be observed between

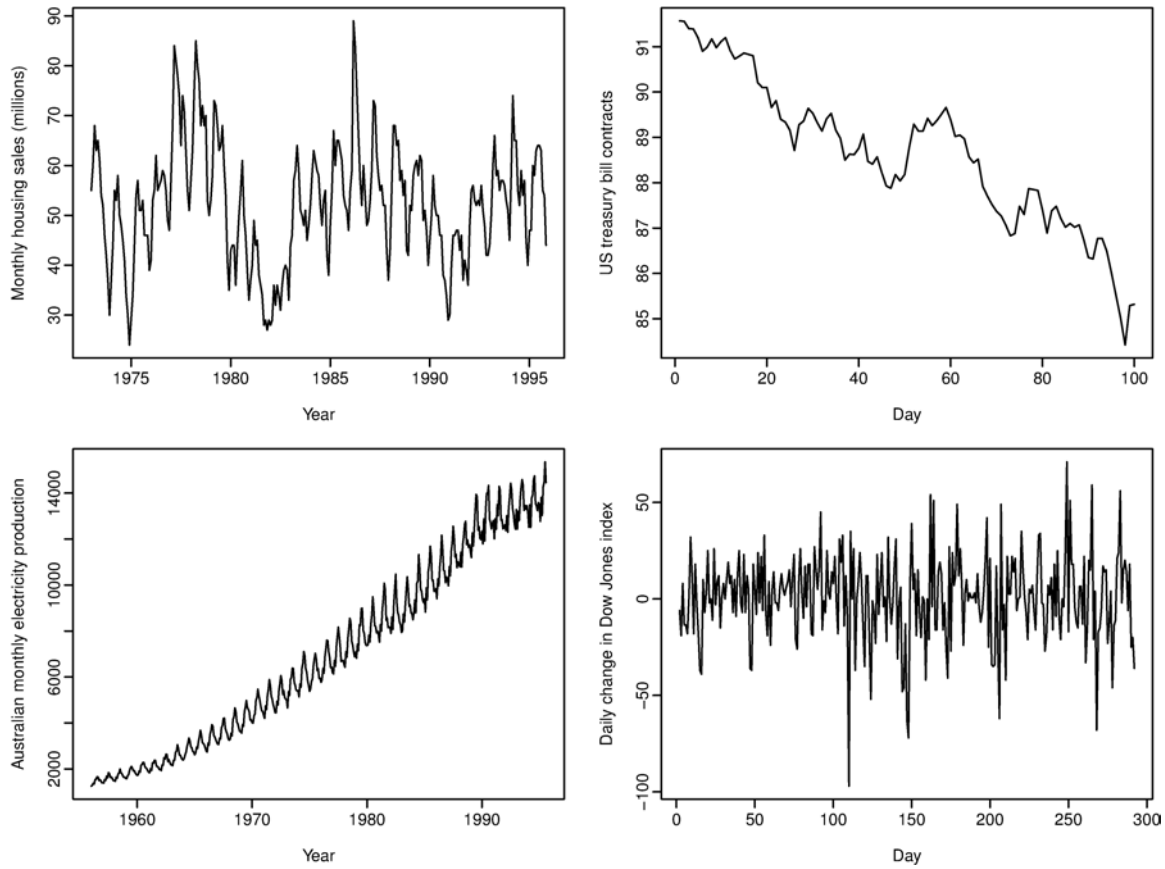


Figure 2.3: Time series exhibiting different patterns [HA14].

1980 to 1984. The US treasury bill contract (top right) exhibit downward trend. The Australian monthly electricity production (bottom left) exhibits an upward trend, with seasonality within each year. The daily change in Dow Jones index (bottom right) shows irregular pattern with no trend, seasonality, or cyclical pattern [HA14].

2.3.3 Time Series Decomposition

Some literature [HA14, MJK15, KC⁺14] noted the three basic components of a general time series as consisting of the aforementioned patterns: the trend, the seasonal pattern, and the irregular pattern. Mathematically, a time series can be written as

$$y_t = T_t + S_t + I_t,$$

where T_t is to a trend, S_t is the seasonal pattern, and I_t is the irregular pattern. This decomposition is useful for time series analysis and forecast.

2.3.4 Time Series Type

Time series have been used to visualise data in many different disciplines such as meteorology, oceanography, seismography, medicine, finance, and engineering.

Surface temperature time series exhibits both daily and yearly seasonal pattern. In a day, surface temperature is highest during daytime and lowest at night due to the presence and absence of the sun, respectively. Within a year, the hottest (highest temperature) and the coldest (lowest temperature) month vary according to the location of the city. The unit of surface temperature time series include Fahrenheit ($^{\circ}\text{F}$) and Celsius ($^{\circ}\text{C}$). An example is shown in Figure 2.4(a).

Rainfall time series also exhibits yearly seasonal pattern. Within a year, rainfall level varies across cities and possibly exhibits irregular pattern. The data unit is millimetres (mm). An example is shown in Figure 2.4(b).

Tidal height time series, such as that in Figure 2.4(c), are used by oceanographers to make predictions. In most parts of the world, there are two high tides and two low tides within a day. This characteristic can be seen as a seasonal pattern within a time series. However, there are some variations in some parts of the world where there are more or less high tides and low tides in each day.

Seismogram is a type of time series that is used to display the ground displacement (in microns or centimetres) caused by an earthquake. Time series can also be used to visualise the ground velocity (in nanometres per second) during earthquake. Figure 2.4(d) depicts a seismogram.

Electrocardiogram (ECG) is a test that measures electrical activity of the heart-beat as shown in Figure 2.4(e). It has a distinct cyclic pattern.

Electroencephalogram (EEG) is a test that measures electrical activity of the brain. There are multiple variants of EEG as shown in Figure 2.4(f). The pattern is somewhat random unlike ECG.

Photoplethysmogram (PPG) measures the oxygen saturation level as shown in Figure 2.4(e). Cyclic pattern can be seen in this time series type.

Respiration time series measures the volume of air inhaled and exhaled with every breath as shown in Figure 2.4(e). The volume of air is called a tidal volume and is written as V_t . The unit of V_t is millilitres (mL). Cyclic pattern can be observed in respiration time series.

Stock price, **exchange rate**, and **gross domestic product** (GDP) are visualised using time series. An example of each type is shown in Figures 2.4(g), (h), and (i), respectively. Stock price and exchange rate time series patterns are mostly

irregular. Changes in these time series are typically subject to accidents in nature or politics such as floods, earthquakes, or strikes.

The amount of electricity produced by different fuel types is also visualised using time series as shown in Figures 2.4(j) and (k). It is measured in megawatt hour (Mwh). Different fuel types have different electricity yield, and thus different patterns. For example, the amount of electricity produced by wind power exhibit a seasonal pattern with an increasing trend as shown in Figure 2.4(j). On the other hand, electricity produced by nuclear power in Figure 2.4(k) displays irregular pattern.

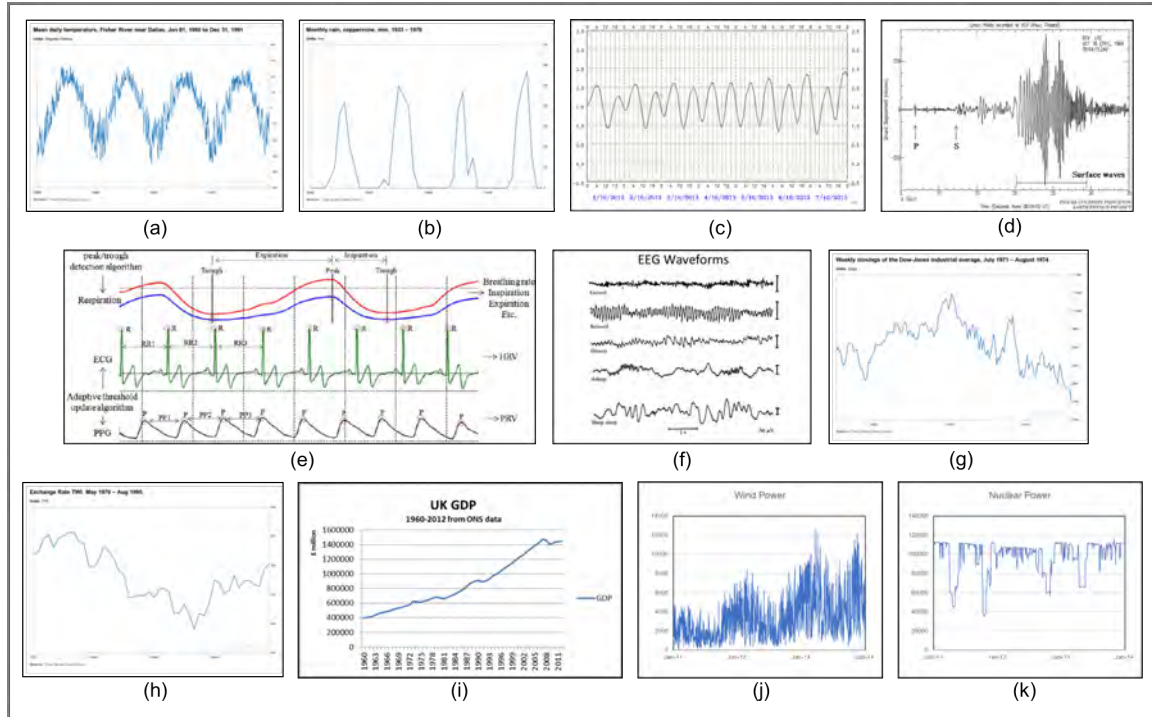


Figure 2.4: Different time series types.

2.4 Empirical Study

Empirical studies involve the process of learning through systematic observations or experience [Goo09]. The goal of the study is to evaluate hypotheses about the causal relationship between variables [Ler11]. In order to make such an evaluation, research questions and hypotheses must first be carefully formulated. Then different variables and possible biases in the experiment must be defined. This depends on the chosen study design principle. After that, the experiment is conducted and the data

is collected. As the effects of independent variables, element of interest, is reflected through dependent variables, dependent variables will be measured and compared to see such effects. Statistical analyses are then used to understand the experiment's outcomes and test the validity of the hypotheses. Finally, conclusions are drawn against the hypotheses.

The following subsections introduce each process in detail, including formulating research questions and hypotheses, identifying different variables in the study, minimising biases and confounding effects, and performing statistical analyses.

2.4.1 Research Questions and Hypotheses

The first step towards designing an empirical study is formulating research questions and hypotheses. A *hypothesis* is a logical prediction about the study's outcome or an empirical result, which must be statistically testable [Goo09, Ler11]. Hypothesis formulation helps deciding on the best mechanism to be used under the scope and the context of the study. Leroy [Ler11] suggested that the goal of the study must be defined at this stage. Goodwin [Goo09] used the term *empirical question (or research question)* to refer to a question that can be answered by observing the data. He suggested that the question must be asked precisely in order to allow specific predictions to be made. Altogether, by defining a clear goal and asking a precise empirical question, a hypothesis can be developed.

2.4.2 Variables

Several kinds of variables are considered in an empirical study. *Independent variables* are the factors of interest that are controlled and investigated by the researcher [Goo09]. Since an empirical study involves making a comparison among different outcomes under different levels of independent variable, it is suggested that the independent variable must have at least two levels (conditions) for the ease of comparison [Goo09]. *Dependent variables* are the factors that are measured during the study, which represent the outcome and allow the investigator to draw conclusions from the hypothesis [Ler11]. A hypothesis is thus a prediction of a causal effect of the independent variable on the dependent variable [Ler11]. Or in other words, it is a prediction of the effect of X on Y [Goo09], where X is the independent variable and Y is the dependent variable.

2.4.3 Bias Caused by Nuisance Variables

Another kind of variable that must be taken into account when designing an empirical study is *nuisance variables*, variables that are not of interest but must be controlled to avoid adding any variations to the outcome [Ler11]. When the variation is unsystematic, it is called *noise*. Otherwise, it is called *bias*.

If a bias is caused by a confound, the independent variable and a bias are *confounded variables*. A *confound* is an uncontrolled extraneous variable that changes at the same time an independent variable changes [Goo09]. This makes it impossible to distinguish whether the effect is resulted from independent variable, confounded variable, or both. As a result, a confound should be detected and held constant during the study.

Similarly, nuisance variables and biases should be recognised and controlled in order to maximise the validity of the experiment [Ler11]. Different kinds of bias exists in the study, and this varies according to the study design principle.

2.4.4 Bias in Within-Subjects Design

The study design principle used in this study is the *within-subjects design* where the same participants are observed across all levels or conditions of the independent variable [Goo09, Ler11]. There are repeated measures conducted for each of them, hence the alternative name *repeated measures design* [Ler11].

Not only does this design principle requires less participants, it also reduces bias and nuisance variables due to individual differences. As opposed to between-subject design, this method risks certain biases such as an *order effect*, which refers to the situation when the experience or change in circumstances after a participant has completed the first part of the study affects the performance in the subsequent parts of the study [Goo09]. There are several types of order effects including a progressive effect and a carryover effect. A *progressive effect* refers to the steady change in a participant's performance each time a study is conducted. A positive progressive effect is an outcome of learning, while a negative effect is due to tiredness, inattention, or boredom. A *carryover effect* occurs when the outcome of certain trial sequence is different from other trial sequences. Different *counterbalancing* techniques are often used to eliminate these effects such as complete counterbalancing, partial counterbalancing, and reverse counterbalancing [Goo09].

2.4.5 Statistical Analysis

After the study is conducted and the data is collected, two kinds of statistics are calculated. *Descriptive statistics* describe the results from the sample of participants, including the mean, maximum, minimum, standard deviation, and median [Goo09]. The *mean* is a descriptive statistics that is most commonly used to measure the *central tendency*. It expresses the average value of a set of N values. The *standard deviation* is a most commonly used descriptive statistics to measure the *variability*, showing how spread out a data set is from its mean.

Inferential statistics allow investigator to draw conclusion about the data, which can be applied to the wider population [Goo09]. For each hypothesis, two mutually exclusive statistical hypotheses are stated [Ler11]. The *null hypothesis* (H_0) states that there is no significant difference amongst different levels of independent variable. In other words, the independent variable has no effect on the dependent variable. The *alternative hypothesis* (H_1) states that there is at least one level of independent variable that is significantly different from the others. The study can either reject or fail to reject the null hypothesis. Rejecting the null hypothesis means that the difference in the outcome does not occur by chance, and that the effects of the independent variables exist and can be generalised. By rejecting the null hypothesis, the alternative hypothesis is accepted as the two hypotheses are mutually exclusive. On the other hand, failing to reject the null hypothesis can be implied that the differences observed from the experiment are likely due to chance.

In order to systematically decide whether to reject the null hypothesis or not, the **significant level** (α) is stated in advanced as the *pre-defined* probability of rejecting the null hypothesis, given that the null hypothesis is true [Ler11]. The most common significant level used is .05 [Goo09], and this level is used in this study. The *p-value* is the *actual* calculated probability of rejecting the null hypothesis, given that the null hypothesis is true. In order to reject the null hypothesis, *p-value* must be less than the chosen significant level ($p\text{-value} < .05$). As Rice (2007) put out, The smaller the *p-value*, the stronger the evidence against the null hypothesis [Ric06]. This is because *p-value* can be thought of as the risk of concluding that the independent variable has an effect on the dependent variable, when there is no such effect. In summary, if the *p-value* is less than the specified significant value, we reject the null hypothesis. This implies that there is a *statistically significant* difference amongst each group of independent variables. Otherwise, we fail to reject the null hypothesis.

Two types of error may occur when performing hypothesis testing. The first type of error, namely *Type I error*, is the result of rejecting the null hypothesis when there

is no real effect from the independent variable. In fact, the probability of making this type of error is equal to the significant level. Another type of error is called *Type II error*, which is caused when researchers fail to reject the null hypothesis when there is real effects from the independent variable [Goo09]. These two types of error should be minimised during the design phase of the experiment.

In a within-subjects design, an inferential analysis will be more sensitive to small differences between means than will be the case for between-subjects design [Goo09]. The statistical tests that are most commonly used with the within-subjects design are the paired samples *t*-test and repeated measures ANOVA [Ler11], both of which will be explained in details in the following subsections.

2.4.6 ANOVA

Analysis of Variance (ANOVA) is a statistical model that is used to analyse the differences among the group means. The group mean is calculated for each experimental condition, which is equivalent to each level of an independent variable [Ler11]. When there are two or more experimental conditions, there are more than two means that are needed to be compared. Although *t*-test can be used to test every combination of the experimental conditions, the chances of making a Type I error increases.

In this study, one-way repeated-measures ANOVA analysis is used as it is the most appropriate model for comparing user performance among different types of information [Sel15]. There is an important assumption in repeated-measures ANOVA that needs to be taken into consideration: the *sphericity* [Sel15]. Sphericity refers to a condition where the variances of the differences between all possible combinations of the groups are equal. Violation of sphericity may increase the chance of making Type I error. Therefore, sphericity must be tested when repeated-measures ANOVA is used. In this study, *Mauchly's Test of Sphericity* will be used.

In Mauchly's Test for Sphericity, the null hypothesis states that there is no violation on the sphericity assumption. If the *p*-value derived from the test is significant ($p < .05$), we do not reject the null hypothesis as the sphericity assumption is not violated. Otherwise, systematic corrections need to be made. *Greenhouse-Geisser* and *Huynh-Feldt* Corrections both estimates epsilon (ϵ), the measure of departure from sphericity [Gir92]. Rutherford [Rut01] commented that Huynh-Feldt introduced a different epsilon measure in order to compensate for the underestimate bias in Greenhouse-Geisser epsilon. As epsilon value varies between 0 and 1, Girden [Gir92] recommends that when epsilon > 0.75 , Huynh-Feldt Correction should be used. Otherwise, if epsilon < 0.75 , Greenhouse-Geisser Correction is preferred.

The F -test is used to test whether the mean difference between groups is higher than that the mean within each group or not [Ler11]. It will indicate a significant difference when more variation exists between the groups than variation within each group. The F -measure is defined as

$$F = \frac{MS_{between}}{MS_{within}},$$

where $MS_{between}$ is the mean square value between groups and MS_{within} denotes the mean square value within groups [Ler11]. The *mean square* (MS) is calculated as

$$MS = \frac{SS}{df},$$

where SS denotes the sum of squares (SS) and df is the degrees of freedom [Ler11]. When reporting the result of ANOVA analyses, this notation is used:

$$F(df_{between}, df_{within}) = F - value, p = p - value$$

2.4.7 t -test

In order to analyse the difference between the *two* group means, t -test is used. Since ANOVA result does not provide details on the comparison between each group, t -test analysis is required when the difference among the group means from ANOVA analysis is significant ($p < .005$). *Bonferroni correction* is then used to reduce the false discovery rate (Type I error).

Chapter 3

Methodology

In this empirical study, three hypotheses will be formulated for each category of information, namely statistical measures, time series type, and time series pattern. Then the evaluation of these hypotheses will be performed by designing and conducting the experiment with relevant independent and dependent variables. Next, descriptive and inferential statistics will be used to analyse the result from the experiment. The analyses will provide statistical evidence to support the formulated hypotheses. Besides the objective measures through statistics, the subjective measure in the form of user rating will also be taken into consideration.

In this chapter, we will first describe the three hypotheses, followed by different types of variables involved in the experiment. Then the measurement metrics will be discussed, followed by the techniques for the analyses. The study design will later be explained in Chapter 4.

3.1 Research Question and Hypotheses

The research questions in our study are ‘*Do different kinds of information affect the user performance in reconstructability*’ and ‘*Does machine-processed information alone yield higher performance in reconstructability than integrating it with human soft information?*’

To answer these question, we formulate the hypotheses for each category of information, and later evaluate them during the analysis phase. The type of information that will be assessed in this study falls into one of the three categories: statistical measure, time series type, and time series pattern.

We summarise here how the concepts in cognitive psychology, information theory, and time series are connected in our study:

- The task in this study is to identify, out of eight optional time series, the time series that satisfies the given three categories of information mentioned earlier.
- In cost-benefit analysis of a data processing workflow, the process of using the given information to find the impression of the original time series is the backward mapping, or the reconstruction function (Section 2.2.4).
- The given information falls into one of the three categories: statistical measure (Section 2.3.1), time series type (Section 2.3.4), or time series pattern (Section 2.3.2).
- Statistical indicator is a machine-processed information.
- The use of time series type information to identify the time series involves the memory (Section 2.1.1)
- By using time series pattern information to identify the time series, pattern recognition is signalled (2.1.3).
- Statistical indicator can be easily evaluated as it is a machine-processed information. However, time series type and time series pattern cannot be evaluated using a forward mapping. Thus, a backward mapping (reconstruction function) is used (2.2.4).

We propose three hypotheses for each category of information respectively:

H1: Min/Max \succ Average \succ Standard deviation \succ Random guess

H2: Temperature \succ ECG \succ Stock market \succ random guess

H3: Global (simple) \succ Local (complex) \succ random guess

The symbol \succ is used to represent that the category on the left hand side of the operator yields higher performance than the one on the right hand side. The performance is measured in terms of accuracy and response time, which will be explained in Section 3.3.3.

In the hypothesis for the statistical measure type (**H1**), we state that the minimum value (min) and maximum value (max) will be more useful for narrowing down the eight optional answers than the average value (mean) and the standard deviation. This is because the minimum and maximum values are directly observable on a time series. Moreover, we believe that the average value will be more useful than the

standard deviation because the standard deviation is not easily measured by only visually observing the time series. Additionally, we state that all of these information, when given, yields higher performance than random guess.

In the hypothesis for the time series type (**H2**), we state that if information on the type of time series (temperature, electrocardiogram (ECG), and stock market) is given, the performance will be higher than random guess. However, the participants may have different knowledge on time series type. For example, medical students may find ECG time series easier to recognise than temperature and stock market time series, while economists may instinctively recognise more patterns in stock market time series than people from other disciplines. Hence, the symbol \succ is used instead of an equal symbol ($=$) in order to make the hypothesis definition consistent. We additionally state that given either one of the three types of time series information will yield higher performance than a random guess.

In the hypothesis for the time series pattern (**H3**), we believe that the global pattern will be easier to discover than the local pattern. Hence the alternative names simple pattern and complex pattern respectively. We will define the definition of *global* and *local* later in Section 3.3.2.3. We also state that the pattern information will yield higher performance than no information (random guess).

3.2 Task

The task in this study is to identify a time series (out of eight candidates) that satisfies given information, which falls into three categories and given at two stages. The three categories of information are the three independent variables that will be explained in Section 3.3.2. More detailed description of the task will be provided in Chapter 4.

3.3 Variables in the Experiment

In order to minimise the confounding effects explained in Section 2.4, it is important to define the variables of interest and control other variables. In this study, the category of information is the independent variables while the quality of reconstruction process is the dependent variable. The following subsections will detail different types of variable involved in this study.

3.3.1 Control Variables

Control variables are variables that are not of interest, but must be controlled to avoid adding any variations to the outcome. Some control variables may also be nuisance variables, and this may have an effect on the outcome of the experiment if not properly controlled.

3.3.1.1 Visualisation Image

The first thing that is held constant in all visualisation images (stimuli) is the *total number of data points*, which is set to 209. Since each data point represents a value that is collected at a certain point of time, the time axis range would be between [0, 208].

Next, all of the data points should be *in the range* of [0, 1,000]. In addition, the format of the image should remain constant in all stimuli. We will explain the mechanism used to ensure that these requirements are met later in Section 4.4.1.

3.3.1.2 Level of Distractor Difficulty

Eight optional answers will be provided for each question. One of them is the correct answer, while the other seven are the distractors.

We control these distractors to have *seven different distracting criteria*. The definition of each criteria will be explained in Section 4.4.2.

3.3.1.3 Position of the ‘Next’ Button

The ‘Next’ button, which is used to transport the participant from pages within a question, and from one question to another, is placed at a constant position on the screen. This is in order to reduce the confounding effect, where the participants need to move the mouse to different positions to proceed.

3.3.2 Independent Variables

In this study, the category of information presented to the human is the independent variables. There are three categories of information including *statistical measure*, *time series type*, and *time series pattern*.

3.3.2.1 Statistical Measure

Several statistical measures can be used to describe time series characteristic. We have introduced some of the measures in Section 2.3.1, which include frequency, time span, mean, variance, and covariance. Some of these indicators are defined for each observation in a time series. Some others are regarded as control variables in our study, such as frequency and time span. Both will be explained in more detail in Section 4.4.1. However, we need indicators that describe the general characteristic of the overall time series. Hence, we introduced these statistical measures in our study:

Minimum and maximum values refer to the lowest and the highest value in a time series, respectively. Since the two values are intuitive and does not require much effort in learning, we select this statistical value in our study. Please note that either the minimum value or the maximum value will be given to the participant, but not both. Yet, we group them into the same statistical measure type.

Average (mean) value is used to measure the central tendency. The mean is calculated by adding up all the data points in a time series and divide by the number of data points. We decide to use the mean value in our study as it is not too difficult to understand.

Standard deviation measures the variability of the data points by showing how spread out the data is from its mean. We decide to use standard deviation in our study because we believe that it is not intuitive to specify the right time series based on its standard deviation.

Other types of statistical measures such as **mode**, **median**, **percentile**, **autocovariance**, **coefficient of variation (CV)**, **skewness**, and **kurtosis** are not used for various reasons. Some of the measures require that the data is sorted (median and percentile). Although we plan to have a pre-study presentation before the study for about 15 minutes that will include the introduction to each type of statistical measure used in this study, some of the measures (autocovariance, coefficient of variation, skewness, and kurtosis) may be too complex for the participants to grasp within a very short period of time. Table 3.1 shows the summary of statistical measure selection for the study.

These four statistical measures represent a *machine-processed* information.

3.3.2.2 Time Series Type

There are several variants of the time series type, some of which may display certain characteristics or features described in Section 2.3.4.

Statistical Measure	Selected for the Study
Minimum and maximum value	Yes
Average (mean)	Yes
Mode	No, too similar to mean
Median	No, required sorted data
Standard deviation	Yes
Autocovariance	No, too complex
Coefficient of variation (CV)	No, too complex
Skewness	No, too complex
Kurtosis	No, too complex

Table 3.1: Statistical measure selection for the study.

For the purpose of the study, four types of time series are carefully selected from those described in Section 2.3. We will now briefly describe their characteristics and the supporting reasons for selecting or not selecting them.

Surface temperature time series exhibits both daily and yearly seasonal pattern. Within a year, the hottest (highest temperature) and the coldest (lowest temperature) month vary according to the location of the city. We believe that this pattern is intuitive and easily recognisable to most people. Thus, this type of time series is selected.

Rainfall time series also exhibits yearly seasonal pattern. Within a year, rainfall level varies across cities. However, sometimes they exhibit irregular pattern. As a result, we do not include this type of time series in this study as it is less intuitive.

Tidal height has seasonal pattern. In most parts of the world, there are two high tides and two low tides within a day. However, there are some variations in some parts of the world where there are more or less high tides and low tides in each day. Consequently, tidal height is not selected for this study.

Seismogram contains certain pattern, which is known to people in the field. However, for lay participants, seismogram may seem unpredictable. Thus, it is not selected to be used in this study.

Electrocardiogram (ECG) patterns are extremely unique. The pattern is recognisable, given that one has seen it before. Also, there are multiple variants of ECG, which are used by doctor when diagnosing a patient. Thus, this type of time series is selected.

Electroencephalogram (EEG), on the other hand, exhibits more irregularity, making it quite difficult to recognise. As a result, it is not selected.

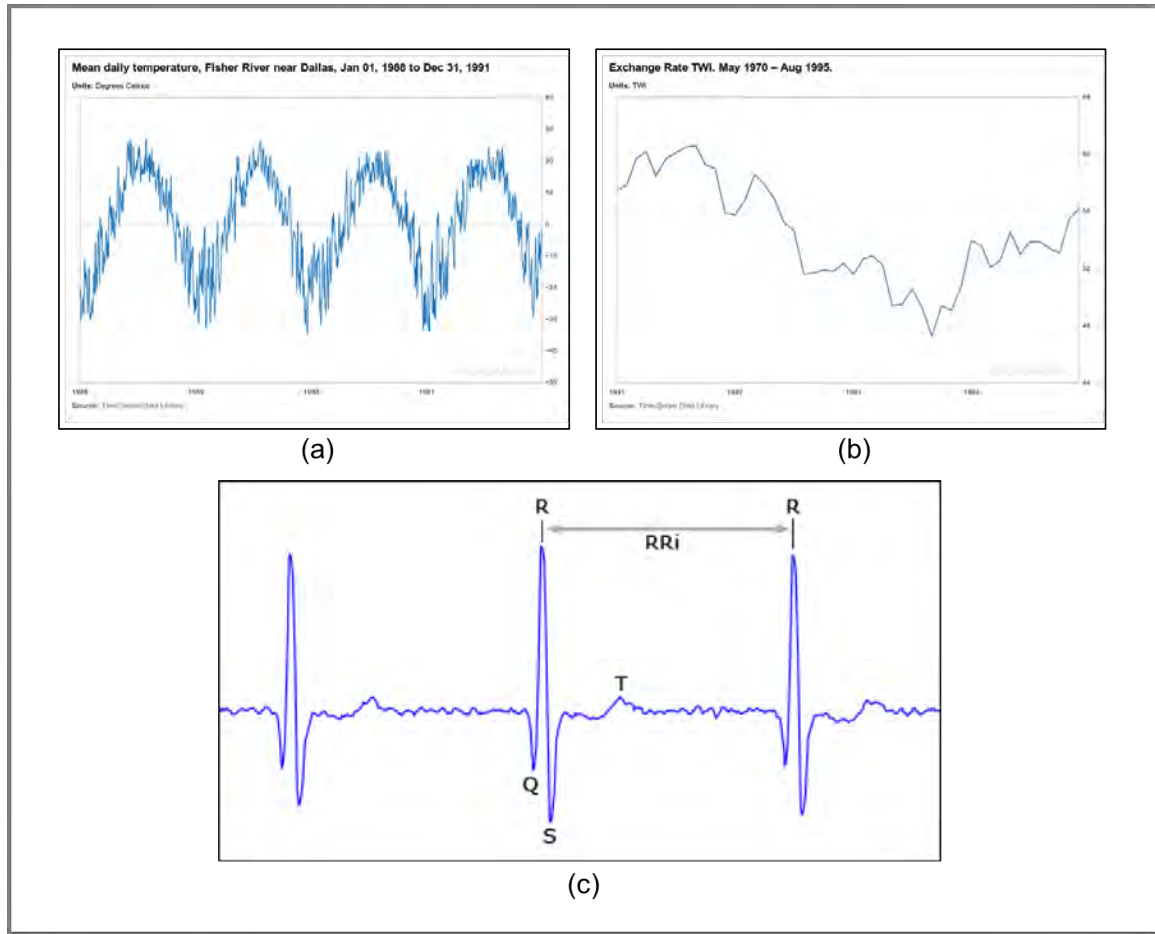


Figure 3.1: Selected time series type for this study.

Photoplethysmogram (PPG) and **respiration** time series, although exhibit cyclic pattern, are too similar to other types of time series that have already been selected. Consequently, they are not selected for this study.

Stock price mostly displays irregular patterns is selected for the study. Despite its general irregularity, stock price does in fact exhibit various distinguishable patterns. These patterns have been widely accepted and used in analysis and forecasting by economists.

Exchange rate is not selected as there are less available data than stock price. In addition, different participants with different countries may have more knowledge on an exchange rate of some currencies than the others. This will eventually lead to confounding effect in the study.

Gross domestic product (GDP) is not selected with similar reason to exchange rate.

Electricity production time series is selected. Different types of fuel used for

electricity production display different pattern in a time series.

In conclusion, the four types of time series selected for this study are surface temperature, electrocardiogram (ECG), stock price, and the amount of electricity produced by different fuel types. These are shown in Figures 3.1(a), (b), (c), and (d), respectively.

Although four types of time series are chosen, only three types will be used in the main experiment. We decide that electricity production is reserved exclusively for the training session. We also intend to do a pairwise comparison between all types of time series. Hence, there are $3! = 6$ possible pairwise combinations of the time series type in the actual experiment and three pairwise combinations for the training session. Details on the training session will be provided in Section 4.3. For simplicity, we use H to represent electrocardiogram time series, S to represent stock market time series, T to represent temperature time series, and E to represent electricity production time series.

We use the notation $Type_1Type_2$ to denote the pairwise combination of time series type. $Type_1$ refers to the correct time series type while $Type_2$ refers to the distractor time series type. This implies the distinction of $Type_1Type_2$ and $Type_2Type_1$. Therefore, pairwise combinations of time series type in the real experiment include: HS , HT , SH , ST , TH , and TS . For training session, EH , ES , and ET are used.

These time series types represent *human soft knowledge* related to memory recall.

3.3.2.3 Time Series Pattern

We have described general time series patterns in Section 2.3.2. These include positively autocorrelated pattern, trend, seasonal pattern, cyclic pattern, and irregular pattern. These patterns are what we call **simple patterns** as they occur *globally* throughout the time span of a time series. Another pattern type of time series we will use is **complex patterns**, which refer the *local* patterns that occur at only certain point of time in a time series. We define the general rule for pattern to based on the global-local criteria. Some exceptions are made as some local patterns may be easier to identify than global ones.

Additionally, we define a set of pattern *sub-type*, which reveal different properties of the data as illustrated in Table 3.2 and capture these patterns (or combination of patterns) in the real data. The formulation of the sub-type is based on Few’s work [Few06] on visual pattern recognition.

For electrocardiogram (ECG) time series, there are a total of 12 patterns. Half of them are simple patterns whereas the other half are complex patterns as shown in

Sub-type	Pattern
1	The relative magnitude of separate sets of values
2	The general nature of change
3	The rate of change of the comparative rate of change
4	Seasonal or cyclic events
5	The stability or volatility of a set of values

Table 3.2: Pattern sub-type.

Table 3.3. The sub-type column in the table specify which properties from Table 3.2 that the pattern should reveal.

Normal ECG contains waves that are regularly spaced and the baseline does not wander up and down.

Wandering baseline, caused by several factors such as patient’s motion, deep breathing, and loosely connected electrodes, shows a series of waves that have different level of baseline. Hence, the name wandering baseline.

Asystole is more commonly known as ‘flat line’. It refers to a cardiac arrest rhythm or simply as an absence of the heartbeat.

Premature Ventricular Contraction (PVC) is a type of ventricular ectopic beat, a beat that occurs at an abnormal position. Normally each beat in a normal ECG has similar height and width. However, in an ECG containing a PVC beat, we can observe a beat that is taller than other beats in the same ECG.

Muscle tremor is a kind of noise in ECG. It is displayed in the ECG as irregular spiky interference at the baseline. However, the R waves (peaks) of the ECG are still observable.

Bigeminy is an abnormal heart rhythm in which each normal beat is followed by an abnormal one.

Atrial Flutter (AF) is an abnormal heart condition, which makes a very distinct sawtooth pattern on an ECG. The patterns occur at the baseline of an ECG.

Atrial Fibrillation (AFib) is another abnormal heart rhythm that shows fine wavy patterns in the baseline of an ECG.

Ventricular Tachycardia (VT) shows a regular bizarre-looking rhythm. The beats are tightly spaced since the heart is beating really fast.

Ventricular Fibrillation (VFib) is a serious abnormal heart condition, which makes a highly irregular pattern on an ECG, and looks much different from a normal ECG.

Type	Sub-type	Pattern name
Simple	2	Normal ECG
	2	Wandering baseline (up)
	2	Wandering baseline (down)
	2, 3	Asystole (front)
	2, 3	Asystole (end)
	1, 4	Premature Ventricular Contraction (PVC)
Complex	5	Muscle tremor
	3, 4	Bigeminy
	5	Atrial Flutter (AF)
	5	Atrial Fibrillation (AFib)
	5	Ventricular Tachycardia (VT)
	5	Ventricular Fibrillation (VFib)

Table 3.3: Electrocardiogram (ECG) time series pattern.

In stock market time series, there are a total of 12 patterns. Half of them are simple patterns whereas the other half are complex patterns as shown in Table 3.4.

Slowly trending up pattern on a stock market time series refers to a stock market that has a steady growth.

Slowly trending down pattern on a stock market time series refers to a stock market that has is trending downwards.

Sharp rise in a stock market refers to a relatively high growth rate compared to the stock market that grows slowly and steadily.

Sharp drop refers to a pattern of a stock market time series, in which the rate of change in the price is relatively high and is trending downwards.

Stable pattern refers to a stock market time series with low volatility.

Data missing pattern may be caused by the fact that the market closes for holidays or simply human error. This is shown in a time series as a period of straight line.

January effect is is a seasonal increase in stock prices during the month of January that follows the drop in price that typically happens in December when investors, engaging in tax-loss harvesting to offset realized capital gains, prompt a sell-off.

Super Bowl was found to have an effect on the stock price, as shown by Chang *et.al.* [CJK09]. When a firm's Super Bowl commercial was highly liked, it enjoyed a stock price boost on the Monday following the Super Bowl.

Outlier value can sometimes be found in stock market time series. It creates a spike that stands out in the time series.

Bull trap is a false signal indicating that a declining trend in a stock or index has reversed and is heading upwards when, in fact, the security will continue to decline. First, price falls from the all-time high, bounces but fails at a lower high creating a ‘bull trap’ and then falls as fear and capitulation become the driving psychological forces.

High volatility stock means that there is high fluctuations in the price.

Price that stays the same for some period may be seen in a very old stock price time series, which contains multiple short flat lines. Some believed that it was a lot more difficult to keep a daily historical stock price in the past than today.

Type	Sub-type	Pattern name
Simple	2	Slowly trending up
	2	Slowly trending down
	2, 5	Sharp rise
	2, 5	Sharp drop
	5	Stable
	2, 5	Data missing
Complex	2, 4	January effect
	2, 4	Superbowl
	1	Outlier value
	2, 3	Bull trap
	5	High volatility
	2, 5	Price stay the same for some period

Table 3.4: Stock market time series pattern.

In surface temperature time series, there are a total of 12 patterns. Half of them are simple patterns whereas the other half are complex patterns as shown in Table 3.5.

Northern Hemisphere temperature is lowest in the beginning and at the end of the year and highest around mid-year. A one-year temperature time series resembles a bell curve. The pattern then repeats in a cycle for each year.

Southern Hemisphere countries have summer during the months of December, January, and February. Its one-year time series resembles a V (or U) letter or a cosine wave.

Biggest range temperature means that the difference between the hottest day and the coldest day is really high.

Straight-line temperature time series usually belongs to those countries that are near the equator as the difference between winter and summer is barely noticeable.

Indian Summer refers to the climate in which September is typically the average warmest month of cities. Its one-year temperature time series looks somewhat asymmetry. Its peak leans slightly toward the right making it looks similar to an obtuse triangle, a triangle that has an angle greater than 90° .

Summer in April can be experienced in countries (or cities) that are close to the equator. They typically have a tropical climate. The coldest month is around December.

Erratic temperature data is observable in a time series. Weather station in some regions may to be more problematic than the others and tend to record erratic data.

Missing value in a temperature data is typically fixed by setting that value to some bizarre default value so that it is recognisable that the actual data is missing. However, if we include those ‘default value’ will be included in the calculation, a bizarre-looking hole is created in the time series.

Long missing gaps may sometimes be seen in a temperature time series in some countries. It look like a true straight line at a certain period of the year. This may be the result of leaving a broken data-capturing machine unfixed for a period of time.

Hot at certain year (and **cold at certain year**) refers to a pattern in a temperature time series when the highest (and the lowest) point in typical cycle is more extreme than usual.

Fluctuate winter refers to a temperature pattern where there is a high fluctuation along the baseline.

Type	Sub-type	Pattern name
Simple	2, 4	Northern Hemisphere
	2, 4	Southern Hemisphere
	5	Biggest range
	5	Straight line
	2, 4	Indian Summer
	1, 4	Summer in April
Complex	5	Machine error
	1	Missing a data point (hidden)
	5	Missing multiple data points
	1, 2	Hotter at certain year
	1, 2	Colder at certain year
	4, 5	Fluctuate winter

Table 3.5: Surface temperature time series pattern.

These time series patterns represent *human soft knowledge* related to pattern recognition.

3.3.3 Dependent Variables

The dependent variable in this study is the quality of reconstruction process. This can be determined using accuracy and response time.

3.3.3.1 Accuracy

Accuracy is an indicator of effectiveness. It will be measured using the defined binary encoding system. In brief, the encoding will distinguish whether the participants choose the correct answer or not and the type of error they make. The design of the binary encoding will be described in Section 4.4.2.

3.3.3.2 Response Time

Response time is an indicator of efficiency. It will be measured in milliseconds. The response time include the time span from when the participant reads the question on the question page until the participant select an answer from the optional answers.

3.4 Measurement Metrics

3.4.1 Objective Measure

To analyse user performance, the dependent variables, the accuracy and the response time, are used. Accuracy is an indicator of effectiveness, while response time indicates the efficiency as discussed in Section 3.3.3.

3.4.2 Subjective Measure

In this study, we will use the ease of identifying a time series as the subjective measurement. The participants will be asked to rate two things: the difficulty of identify the right time series when given different types of time series, and the difficulty of identify the right time series when given different types of statistical measures. The scale we use consists of five levels: “*Very difficult*”, “*Slightly difficult*”, “*Neither difficult nor easy*”, “*Slightly easy*”, and “*Very easy*”.

3.5 Techniques for Analyses

In this study, there are three independent variables (statistical measure, time series type, and time series pattern), and two dependent variables (accuracy and response time). In order to accurately evaluate the three hypotheses provided in Section 3.1, we will analyse user performance in using different kinds of information. The hierarchy of performance analysis is illustrated in Figure 3.2. Within each single performance analysis, accuracy and response time will be analysed separately.

In order to ensure a fair comparison, we will apply a set of formal analyses as follows. First, we will use *one-way repeated-measures ANOVA analysis* described in Section 2.4.6, to test whether there is a significant effect of the information type on the performance. The null hypothesis for our analysis is that there is no difference of effectiveness among each type of information in the three categories, which can be written as $H_0 = \mu_i = \mu_j = \dots = \mu_k$ where $\mu_i, \mu_j, \dots, \mu_k$ are the performance mean for each of the information type in a category. For example, μ_i, μ_j , and μ_k are the performance mean for minimum and maximum value, mean value, and standard deviation in the statistical measure category.

Second, if ANOVA reports that a significant effect exists, we will further perform *t-test analysis with Bonferroni correction* to examine the source of the main effect and provide statistical evidence for the pairwise comparison results. The null hypotheses are $H_0 = \mu_i, H_0 = \mu_j, \dots, H_0 = \mu_k$ for each of the information type in the same category.

The significant level of these two analyses will be set to $\alpha = .05$, as described in Section 2.4.5. In this study, we use R to perform both analyses.

We will report the study's result in pairwise in comparison among each type of information in the same category by using '[' to enclose the information types with no significant difference performance ($p > .05$), and use '>' to show that the information on the left hand side of the operator is significantly more useful than the technique on the right hand side ($p < .05$).

Furthermore, the value in '(' after each information type will indicate the average performance of that information type. For statistical measure and time series type, an average accuracy range is between [0, 12] since there are three types of statistical measures and time series type. The average accuracy range for time series pattern is [0, 18] because there are only two generalised pattern. The average response time will be shown in seconds.

We will provide a bar chart displaying the average performance for each type of information in terms of accuracy and response time separately. In each accuracy chart, we will add one more column, namely the *random guess*, in which the mean probability of answering the question correctly will be reported assuming normal distribution among the optional answers. For example, out of eight optional answers, the probability of selecting each answer is 12.5%. This is in order to evaluate whether the hypotheses, which states that any kind of information, when given to human, will be more effective in the process of reconstruction than a complete random guess (no information), is satisfied or not.

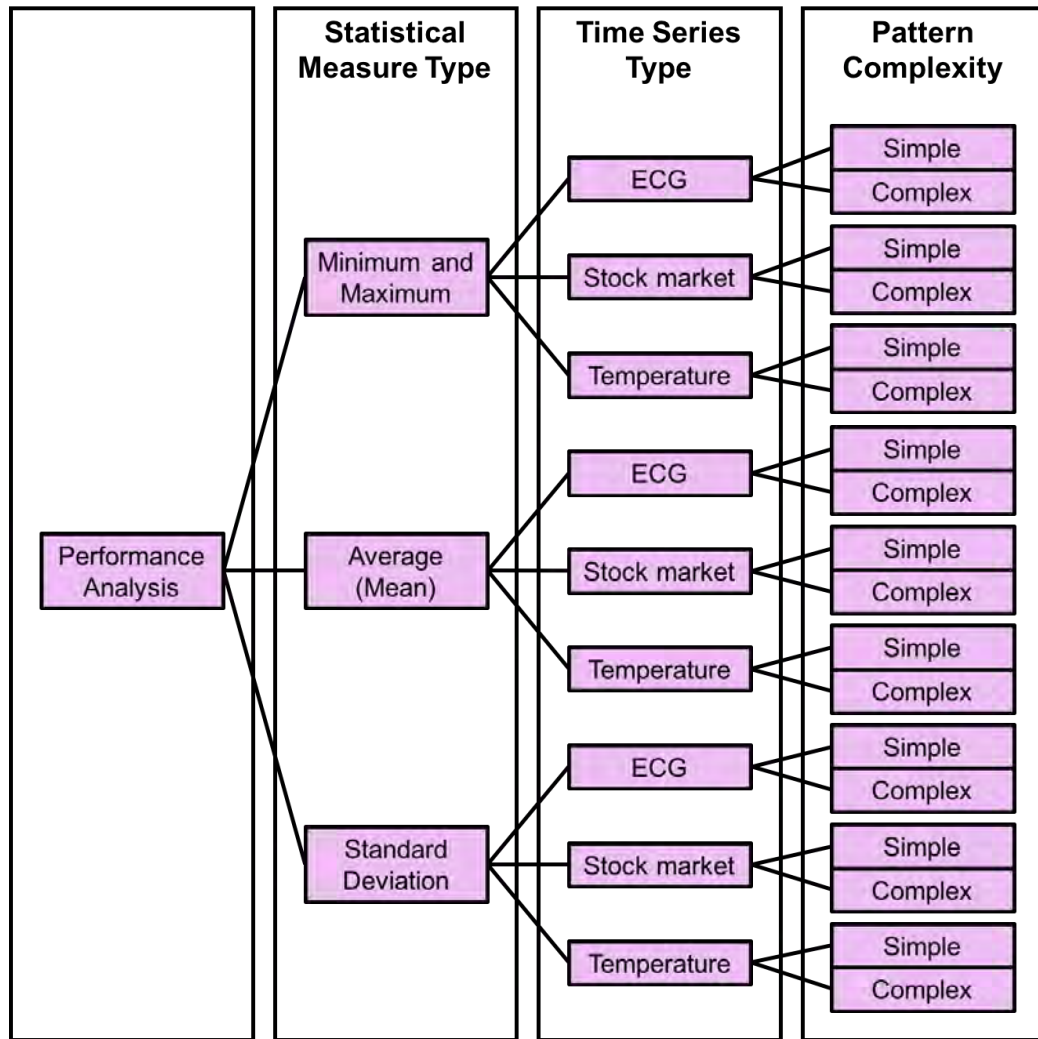


Figure 3.2: Hierarchy of performance analyses.

Chapter 4

User Study Design

User study design is a crucial part of an empirical study. In this phase, it is important to minimise biases and confounding effects, which may affect the outcome of the study. This is because the cost of solving the problems that arise at a later stage, in terms of effort, is more expensive. The user study design consists of four main elements, task design, trial design, stimulus design, and software design. In this chapter, we will first provide an overview of these design elements, then we will explain these four design elements in detail.

4.1 Design Overview

Confounding effects are commonly unavoidable when conducting an empirical study. Although such effects cannot be eliminated completely, a careful design of a user study should be able to reduce them to the level that they will not remarkably affect the participants' performance.

4.1.1 Task Design Overview

The task in this study is to *identify a time series (out of 8 candidates) that satisfies given information, which falls into three categories and given at two stages*. The three categories are the three independent variables in this study, namely statistical measure, time series type, and time series pattern. The two stages refers to the sequence that the information is displayed. In the first stage, time series type and time series pattern are given. In the second stage, statistical measure and its value are provided along with the 8 optional answers.

To collect user performance in each trial, the *multiple choice* question format will be used as it is assumed that most participants are familiar with this format, and

thus require little or no learning effort. The trial design is explained in the following subsection.

4.1.2 Trial Design Overview

A *trial*, in this context, refers to a set of screens presented to the participants during the experiment. Each trial is composed of three elements, *a masking screen*, *an information screen*, and *a question screen*. An information screen and a question screen are collectively referred to as stimuli pairs. In each trial, these two stimuli will be given in a separate stage.

4.1.3 Stimulus Design Overview

In order to minimise the confounding effect, we have to ensure that there are an *adequate number of stimuli* for each information category. Since a piece of information in the three category is always given in the same set of stimuli, we design an appropriate number of stimuli accordingly in order to ensure that we conduct a fair comparison.

For the first category, *statistical measure*, there are 3 types of information namely minimum or maximum value, mean value, and standard deviation, which were selected in Section 3.3.2.1. In the second category, *time series type*, we selected electrocardiogram, stock market, and temperature time series. In the last category, *time series pattern*, there are two generalised types, namely simple pattern and complex pattern, which were explained in Section 3.3.2.3.

In order to test all the pairwise combinations with all types of statistical measures and all type of pattern, we have: 3 types of statistical measure \times 6 pairwise combinations of time series type \times 2 types of time series pattern = 36 stimuli pairs. Note that we refer to stimuli as *pairs* because a single stimuli is given in one trial at *two stages*. There will be three additional stimuli pairs, all of which are for electricity production type and are solely for the training session. In summary, there will be 36 stimuli pairs for the main experiment and three stimuli pairs for the training session.

4.1.4 Software Design Overview

The software is designed to collect accurate user performance on different kinds of information from the three categories.

In each trial, the mask screen is first shown briefly for 2000 milliseconds to the participants. At this stage, no interaction with the software is allowed. After 2000

milliseconds, an information screen is shown on the screen containing information from the two categories. As soon as the information screen appears, we start the *reading timer*. The ‘Next’ button is placed at the bottom of this screen to enable the participants to proceed to the question screen when they have finished reading the information on this information screen. After the ‘Next’ button is clicked, the reading timer stops while the *answering timer* starts at the same time a question screen appears, containing the information on the last category. In this screen, 8 optional answers are placed under the question. The ‘Next’ button is placed at the bottom of the page to enable the participants to proceed to next trial when they have selected one of the optional answers. The answering timer is stopped when the participant click on the ‘Next’ button.

The answering time will be used in the analysis of the *response time*, while the reading time will be used to further analyse the result. The order of trials in the software will be pseudo-randomised in order to minimise the sequence effects, as described in Section 2.4.4.

4.2 Task Design

The task in this study is to *identify a time series (out of 8 candidates) that satisfies given information, which falls into three categories and given at two stages*.

In this study, three pieces of information are given in each trial: statistical measure, time series type, and time series pattern. *Statistical measure* information and its associated value that will be given is either minimum/maximum value, mean value, or standard deviation. *Time series type* will be indicated as either electrocardiogram, stock price, or temperature time series. *Time series pattern* is not directly indicated as simple or complex. Rather, it is given as a textual description of ‘what to look for’.

There are two stages at which the three information is given to the participant. The stage refers to the sequence that the information is displayed. In the first stage, time series type and time series pattern are given in the form of an article from newspaper or magazine. In the second stage, statistical measure and its value are provided along with the 8 optional answers.

The task can be performed by carefully reading the passage on the *information screen* then locate the time series type and the time series pattern first. Later, the participant may try to remember these two pieces of information before proceeding to the *question screen*. When the question screen is shown, the participant has to

read the question to find what type of statistical measure the question is asking for together with its associated value. Finally, they can analyse these three pieces of information and eliminate the time series that seem ‘unlikely’ from the 8 optional answers.

4.3 Trial Design

This trial structure is illustrated in Figure 4.1.

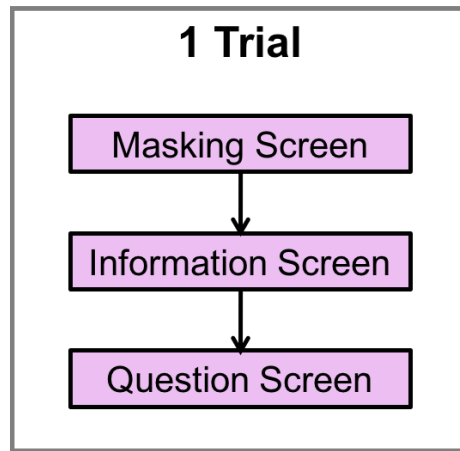


Figure 4.1: Trial structure.

A *masking screen* is meant to signal the participants that the new trial has started. It is also used to keep the participants’ performance consistent by reducing confounding effects such as remembering the pattern of the time series.

An *information screen* provides the *two* type of information to the participants, namely the time series type and the time series pattern. The information provided is in the form of an article from newspapers or magazine. This is the *first stage* in the sequence of information given to the participant. An example of an information screen is illustrated in Figure 4.2.

A *question screen* provides an instruction for answering the question. It also provide the last type of information, namely the statistical measure type and its value. The question screen is the *second stage* in the sequence of information that is provided to the participants. The instruction and the statistical measure are stated in the question at the top of the screen. Below the question, eight optional answers (time series) are displayed.

Among the eight optional answers, one of them is the correct answer and the rest are distractors, all of which has different distraction criteria using a predefined

binary encoding system. The encoding system will be explained in Section 4.4.2. An example of a question screen is shown in Figure 4.3

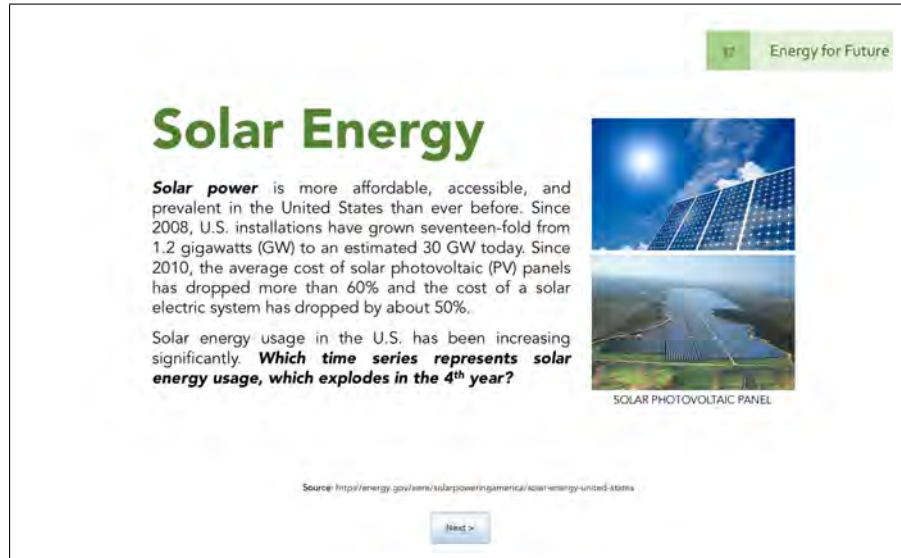


Figure 4.2: Information screen.

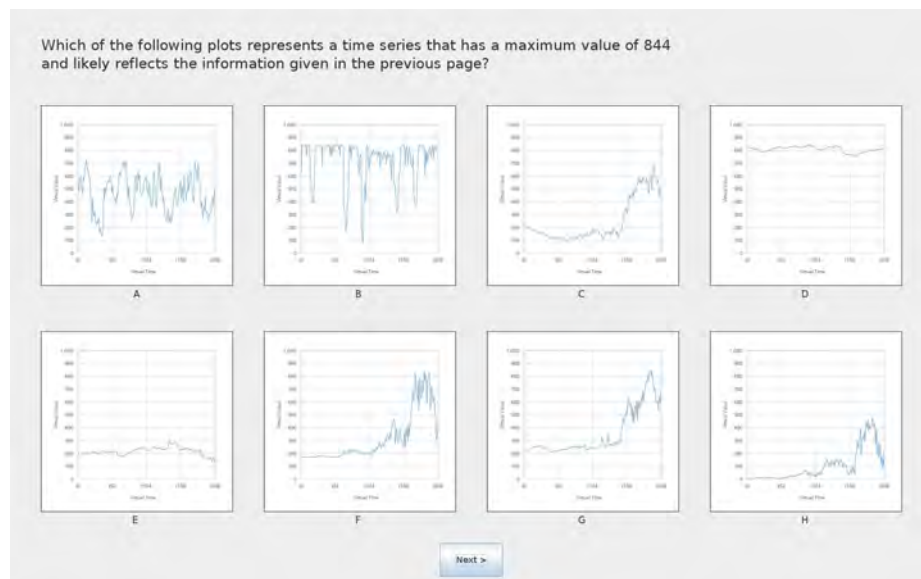


Figure 4.3: Question screen.

4.4 Stimulus Design

There are two processes involved in designing the stimuli, data rule design and visualisation image design.

4.4.1 Data Rule Design

In each stimulus, there are a total of *209 data points*, where each data point in the stimuli is a bivariate data (x and y). A number of rules are applied when the data are captured and transformed in order to minimise the confounding effects.

4.4.1.1 Virtual Axes

Firstly, since the three types of time series (+ 1 type for trial session) utilise different data unit over different time span, we have *normalised* and *standardised* the data in such a way that the original data range and time range are not observable. This is in order to prevent participants from detecting the type of a time series by observing the x - and y -axis instead of using the information we provided.

We replaced the original x -axis and y -axis by the virtual time and the virtual value axes, respectively. The *virtual time* is in the range of 0 to 209 as we have set the number of data point to be precisely 209. The label on the x -axis shows five virtual time points including $t_0, t_{52}, t_{104}, t_{156}$, and t_{208} . For the *virtual value* we define the range to be between 0 to 1,000. The label on the y -axis shows eleven virtual value points from 0 to 1,000 with the difference of 100 between each point.

For electrocardiogram (ECG) time series, the frequency (time span from t_i to t_{i+1}) is every 0.004 second. Since the frequency is too condense, we take the average value of the three data points and define the frequency to be every 0.012 seconds. The time span for ECG time series is thus $0.012 \times (209 - 1) = 2.496$ seconds. Since the original data range of ECG sometimes is around $[-5, 5]$, we normalise the data to change the data range to $[0, 1,000]$.

For stock market, surface temperature, and electricity production time series, the time span is set to 4 years. The frequency is set to weekly. Hence, $(52 \times 4) + 1$ leap day make 209 data points. However, a weekly data maybe more difficult to find than daily data. Consequently, if only daily data is available, we use seven-day average value.

Stock market data unit is in U.S. dollars (\$). When the stock price exceeds our specified virtual value range, we normalised it by scaling down the data. Similarly, when the price difference between each data point is as low as tenth decimal places, we normalised it by scaling up the data.

The data unit we select for surface temperature time series is Fahrenheit ($^{\circ}\text{F}$). On the Fahrenheit scale, the freezing point of water is 32°F and the boiling point is

212°F (at standard atmospheric pressure). Since the temperature can be lower than the freezing point, we normalised the data to make it non-negative.

Finally, the amount of electricity produced by different fuel types is measured in megawatt hour (Mwh). As different fuel types have different electricity yield, the data range differ across the fuel types. Hence, we need to normalise them to fit within the specified virtual data range.

By doing so, we fulfilled the first rule of not having negative data point values, and having the value lies within a specified range. An example of an electricity production, electrocardiogram, stock market, and temperature time series are illustrated in Figures 4.4(a), (b), (c), and (d) respectively.

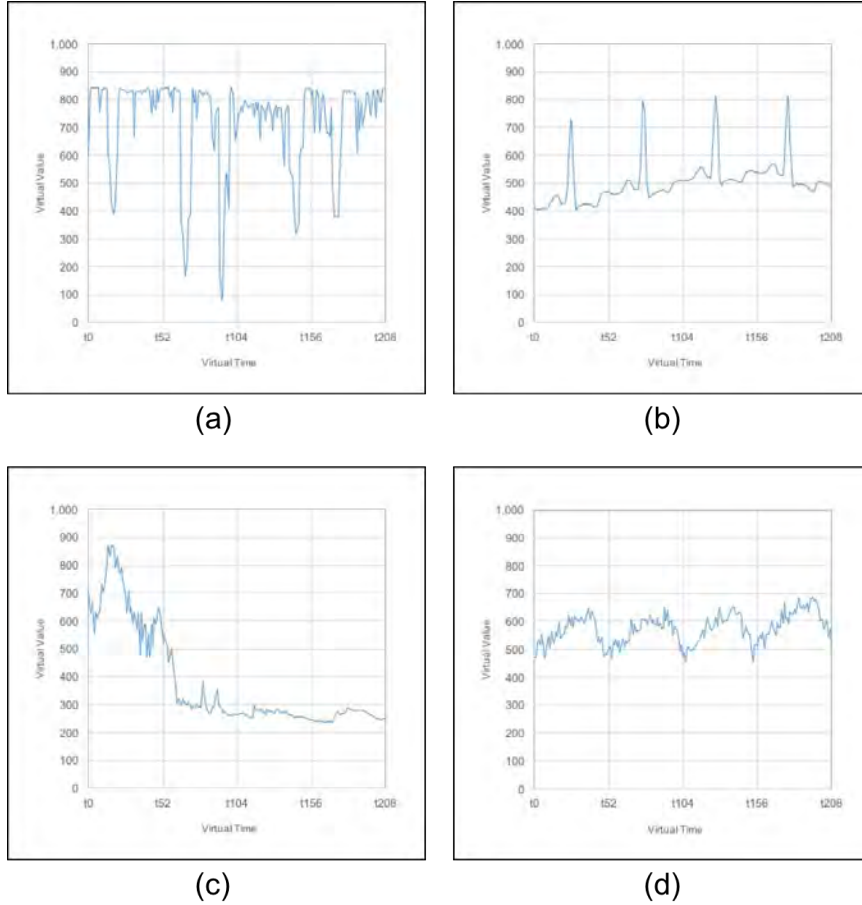


Figure 4.4: Time Series Stimuli.

4.4.1.2 Reliable Data Source

Since ‘soft’ knowledge includes known theories and past experience, the *time series data* that are used in this study must be generated from the actual data. This is to

ensure that the participants are allowed to actually use ‘soft’ knowledge to determine the answer.

For electrocardiogram (ECG) time series, we capture the data from Physio.net [GAG⁺00], which is a collection of ECG databases containing thousands of actual ECG time series. The site enables the user to visualise the data as a graph in different time span before exporting the data into a comma separated values (CSV) file. We explore different types of ECG from different databases and select the ones that are appropriate to the study.

Surface temperature time series are generated from the data from the Environmental Protection Agency Average Daily Temperature Archive of the University of Dayton [Knu]. The archive contains temperature data for multiple cities from around the world. It contains the data of all states in the United States. Unfortunately, the site does not provide the feature to visualise the data before downloading it as a comma separated values (CSV) file. Thus, we select some of the cities and go through them one-by-one to search for appropriate ones to be used in the study.

We take stock market data from Yahoo Finance [Yah]. The site provides interactive chart that visualises the historical stock price before downloading it as a comma separated values (CSV) file. We explore multiple stock price of different companies and select those that are suitable for the study.

The amount of electricity production data is taken from ISO New England, which is the independent, not-for-profit corporation responsible for keeping electricity flowing across the six New England states. The site provide a data file without any visualisation feature. Similar to surface temperature data, we go through each type of fuel one-by-one and look for the data that is suitable for the study.

In addition, the *information contained in the information screen* is also ensured to be from reliable sources. For example, information about stock market is taken mainly from Investopedia, which is the world’s leading source of financial content on the web.

4.4.2 Distractor Rule Design

In each trial, we design different distraction criteria as a **binary encoding system**. Each of the eight optional answers are paired with a three-digit binary number. Each digit represents each *information category*. Suppose i , j , and k represent the number at the three positions of a three-digit binary number from left to right respectively, this number is written as ijk .

- i represents the time series type correctness.
- j represents the time series pattern correctness.
- k represents the statistical measure correctness.

i can take a value of either 0 or 1. While 1 indicates that the time series has the type that satisfy the question criteria (correct type), 0 suggests that the time series is of the wrong type. This definition is similar for j and k . Hence, we have eight possible three-digit binary number as illustrated in Table 4.1

Binary Code	Type	Pattern	Statistics
111	Correct	Correct	Correct
110	Correct	Correct	Incorrect
101	Correct	Incorrect	Correct
100	Correct	Incorrect	Incorrect
011	Incorrect	Correct	Correct
010	Incorrect	Correct	Incorrect
001	Incorrect	Incorrect	Correct
000	Incorrect	Incorrect	Incorrect

Table 4.1: Binary encoding optional answers.

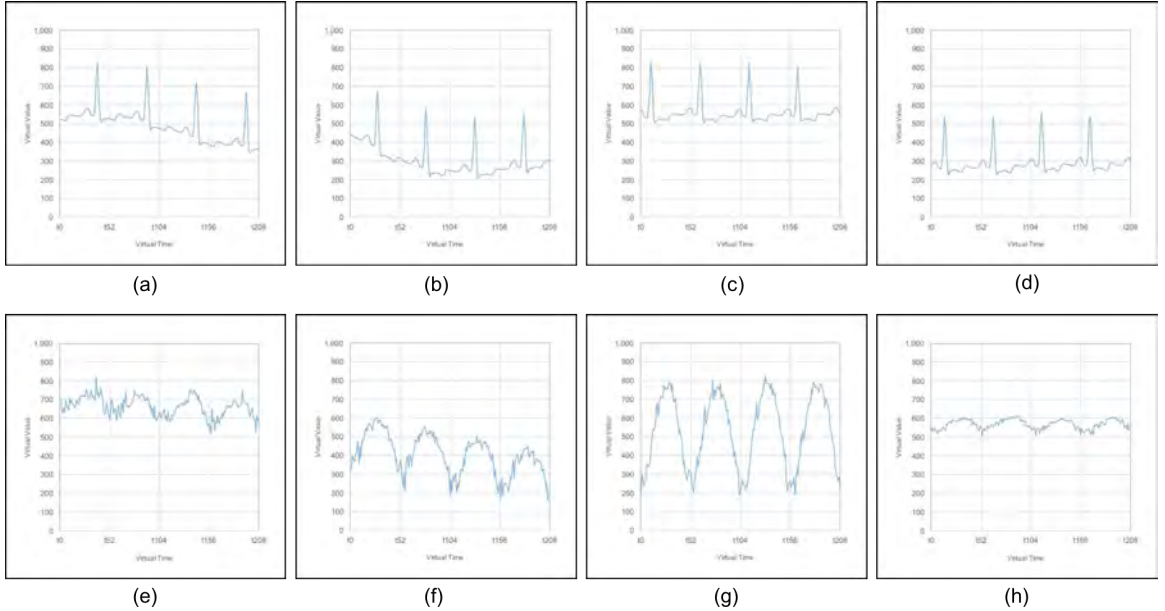


Figure 4.5: Distractors.

An example of the eight optional answers design is illustrated in Figure 4.5. For this particular case, the information given to human include: *maximum value* =

824, *time series type = ECG*, and *pattern = wandering baseline*. Figure 4.5(a) is the correct choice (111). Figure 4.5(b) has the wrong statistical measure, but the type and pattern is correct (110). Figure 4.5(c) is an ECG with correct maximum value, but missing the specified pattern (101). Figure 4.5(d) only has the type that is correct (100). Figure 4.5(e) is the wrong type of time series (011). Figure 4.5(f) is the wrong type and does not satisfy the maximum criteria (010). For Figure 4.5(g), only statistical measure is satisfied. Finally, Figure 4.5 does not satisfy any of the three given information.

4.4.3 Visualisation Image Design

There are two types visualisation image in each stimuli pairs, the information screen, and the eight optional answers, which are time series.

4.4.3.1 Information Screen

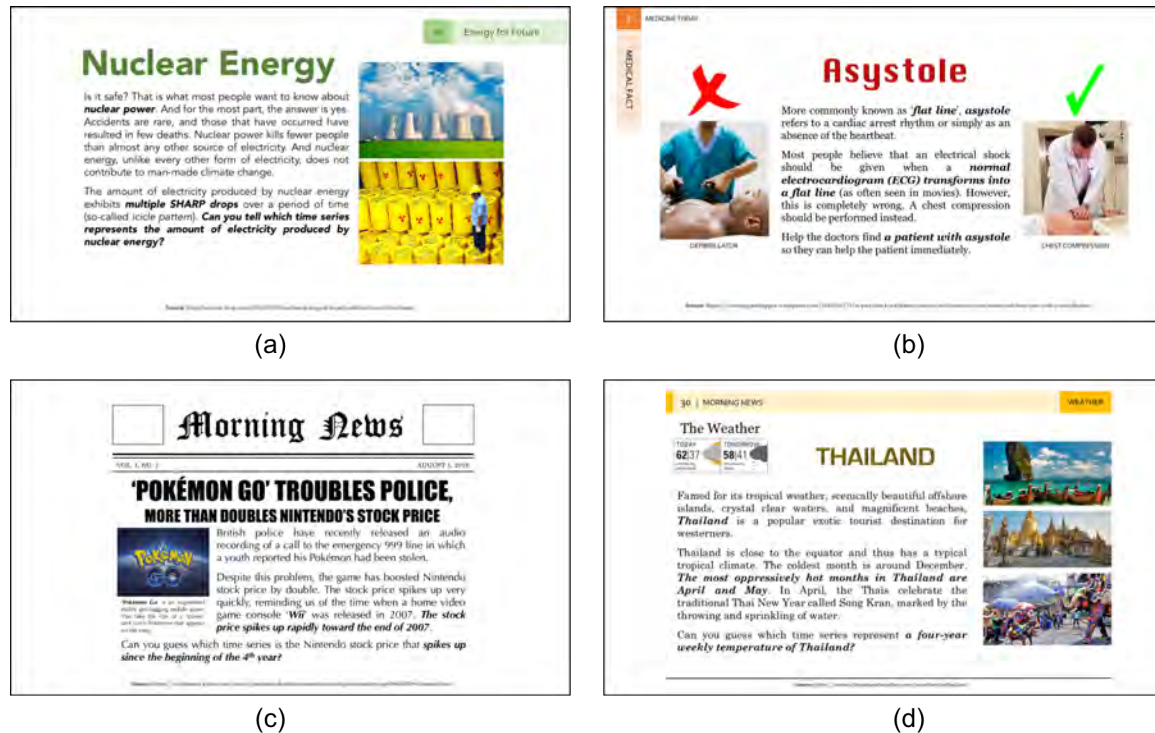


Figure 4.6: Information screen theme

The information screen are designed using *three distinct themes*, one for each type of time series. For information screen describing an electrocardiogram and its pattern, we use a medical magazine style. For information screen describing a stock market

and its pattern, we use a newspaper article style. For information screen describing a temperature and its pattern, we use a travel magazine style.

Among the three themes, the we use different set of colours, fonts, format, and style. However, we ensure that information screen of the same time series type have the same theme. An example of each theme is provided in Figure 4.6. Electricity production, ECG, stock market, and temperature time series use a consistent theme as shown in Figures 4.6(a), (b), (c), and (d) respectively.

4.4.3.2 Time Series

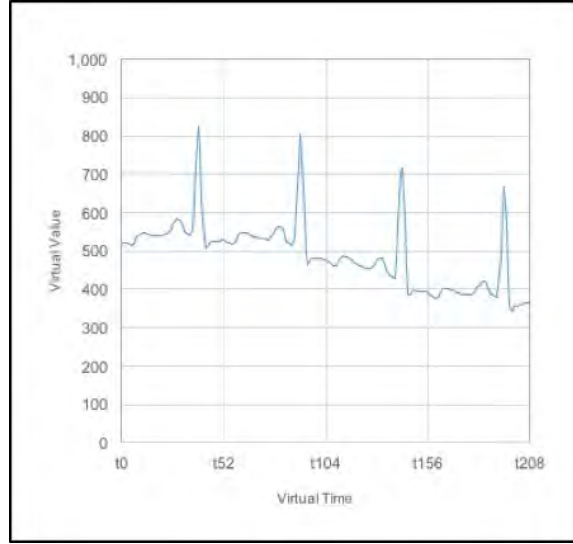


Figure 4.7: Time series example

All of the time series in this experiment are designed using some general rules of colour, axis scale and tick marks, font, and size.

Colour. Since only one time series is shown in each Cartesian coordinate plot, a single colour of a time series suffices. The colour chosen for a time series is blue, the grid lines are in light grey, the fonts are all black, and the background is white.

Axis Scale and Tick Marks. We design the axis scale to be consistent with the data rule in Section 4.4.1. The data axis have a fixed scale from $[0, 1,000]$, with tick marks at every 100 values. The time axis have a fixed scale from $[0, 208]$, with tick marks at every 52 values. They are labelled as $t_0, t_{52}, t_{104}, t_{156}$, and t_{208} .

Font. All text in every time series use *Arial* font to ensure consistency.

Size. All time series visualisation image has a fixed size of 400×375 pixels.

An example of a time series that is designed based on these criteria is displayed in Figure 4.7

4.5 Software Design

The software in the experiment is used to collect the participants' performance of reconstructability based on the different types of given information. In the following subsections, we will present a workflow of the software and explain each of the workflow component. Then we will explain the design of the sequence as well as the time scheme.

4.5.1 Software Workflow

The workflow of the software consists of four main parts as illustrated in Figure 4.8. These include demographic questions and familiarity rating, training session, study session, and acknowledgement of the end of the study.

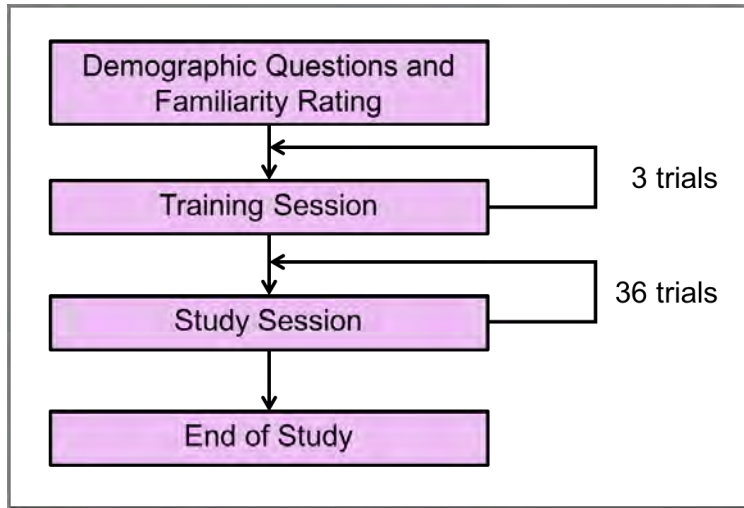


Figure 4.8: Software workflow

In the first part of the software, the participants are asked to provide some *demographic information* including the allotted user ID, gender, age group, and occupation. They are also asked to rate the *familiarity with the time series* from the five-level scale including “*Never heard of it before the study*”, “*Heard of it, but do not understand it*”, “*Moderately familiar*”, “*Very familiar*”, and “*Highly knowledgeable*”.

The second part of the software is the *training session*, which consists of three trials. Each trial is designed to familiarise the participants with the instruction on how to perform the task accurately. To facilitate learning in each trial, the participants are given feedback in the form of correct answers. This allows the participants to compare their answers against the correct answers in order to improve their performance in

the subsequent trials. The three trials in the training session provide training on retrieving a time series that satisfy the maximum or minimum value, the average (mean) value, and the standard deviation, respectively. The type of time series used in these training trials is the amount of electricity produced by solar, wind, and nuclear power, respectively. This type of time series is not used in the actual study session.

The third part of the software is the *study session*. In this session, the total number of 36 trials are presented to each participant. However, unlike training session, no feedback is provided.

In the last part of the software, participants are acknowledged that they have reached *the end of the study*. They are also notified to complete a paper-based subjective questionnaire before leaving the experiment venue.

4.5.2 Sequence Design

As mentioned in Section 2.4.4, stimuli (or stimuli pairs) may create positive and negative effects on the stimuli (or stimuli pairs) displayed later. The order of the optional choices may also affect the time required to select the correct answer since the participants need to move the mouse in different amount to different position.

Therefore, we reduce this order effect by using *pseudo-randomisation* so that participants would not be able to guess the order of the stimuli pairs, or to reason the position of the correct answer. Within each trial, the pseudo-randomisation scheme uses complete counterbalancing technique to ensure that all eight optional answers have equal chance to be shown in each position. Additionally, the scheme also ensures that among all the trials: i) no more than two equivalent type of statistical measurement is shown consecutively and ii) no more than two equivalent combination of time series type is compared consecutively. For example, if the trial for time series type pairs *HS* are already shown consecutively once, the next trial cannot be *HS*.

4.5.3 Time Design

We do not have a time limit for each trial in the study and the participants are reminded of this in the pre-study presentation.

Chapter 5

Implementation

The implementation process consists of three main components, stimulus generation, software development, and the experiment itself.

Stimulus generation and software development were iteratively performed using Agile software development methodology, which assures the software quality. Then we conducted a controlled experiment using the developed software to collect user performance on the three types of information that influence human capability of information reconstruction.

There are a total of 1,656 trials included in our study (excluding trials in the training session): $46 \text{ participants} \times 3 \text{ statistical measures} \times 6 \text{ pairwise combinations of time series type} \times 2 \text{ general time series pattern}$.

We will begin this chapter by describing the implementation process used for stimuli generation and software development in the experiment. Later, we will explain the stimulus generation process and software development process in details. We will finish the chapter with the details of the experiment.

5.1 Implementation Process

In our project, *Agile software development* methodology is used to generate the stimuli and develop the software. Agile methods are suitable for time-critical projects as it is based on iterative, incremental, and evolutionary process. Furthermore, it is also suitable for projects that require requirement change along the development life cycle. Each iteration is equivalent to a *sprint*, which is a set period of time during which specific work has to be completed and made ready for review. Activities in each iteration includes *planning*, *requirement analysis*, *design*, *implementation*, and *testing*. At the end of each iteration cycle, feedback is sought from our research team members to evaluate the stimuli and the software. The feedback is then used

as a requirement for the next iteration. In this project, we define an iteration as a one-week period.

Since the stimuli and the software can be implemented independently, we do them concurrently. The following subsections describe the iterations involved in stimulus generation and software development separately.

5.2 Stimulus Generation

Stimulus generation consists of two main processes, data gathering and visualisation image generation.

5.2.1 Data Gathering

This study intends to study how well human can use different information in the process of reconstruction. As the reconstruction process involves cognitive processes (pattern recognition, remembering, and so on), it is necessary that the data used in this study is *gathered* from real examples instead of *randomly generated* by the experimenter.

For time series data, we did a research on each of the time series type to find interesting patterns that fits into the criteria defined in Section 3.3.2.3 to be used in this study. After the patterns were defined, we searched for information to be put on an information screen from reliable sources. As the pattern name was not sufficient for participants to identify the correct time series, we provided additional information on the pattern using simple analogy such as “*looks like a bell curve*” and “*looks like a true straight line*”. We also tried to keep the information screen interesting to avoid boredom, which may result in confounding effects in the experiment result.

5.2.2 Visualisation Image Generation

We use *Microsoft Excel* to generate time series images and *Microsoft PowerPoint* to generate information screen images.

For time series generation, we first gathered data from reliable sources and saved it in *Microsoft Excel* as a table. We then used the built-in chart generation functionality to generate the stimuli and modified the chart design according to the standard defined in Section 4.4.3. The images were then saved as a static visualisation image in JPEG format. JPEG is named after its developer, *Joint Photographic Experts Group*. These JPEG images were used in the software program.

For information screen images generation, we first designed the theme for each type of time series as mentioned in Section 4.4.2.1. Then relevant textual information on the pattern was added into the image. The images were designed using *Microsoft PowerPoint* and saved as static visualisation image in *Portable Network Graphics* (PNG) format.

5.2.3 Stimulus Generation Iteration

Stimulus generation took approximately eight weeks for completion.

Iteration 1	<ul style="list-style-type: none"> • Research on time series patterns for each time series type from different sources. • Define data sources for each time series type that are reliable for subsequent data gathering.
Iteration 2	<ul style="list-style-type: none"> • Define data rules. • Design visualisation image for time series. • Retrieve ECG data from reliable sources. • Generate ECG time series.
Iteration 3	<ul style="list-style-type: none"> • Redesign visualisation image for time series according to feedback. • Retrieve stock market and temperature data from reliable sources. • Generate stock market and temperature data time series.
Iteration 4	<ul style="list-style-type: none"> • Modify time axis in the visualisation image. • Design visualisation image for information screen. • Retrieve ECG information to put on information screen. • Generate ECG information screens.

Iteration 5	<ul style="list-style-type: none"> • Add grid lines in visualisation image. • Redesign information screen. • Retrieve more time series data. • Generate missing time series.
Iteration 6	<ul style="list-style-type: none"> • Redesign information screen. • Retrieve more time series data. • Generate missing time series. • Retrieve stock market information to put on information screen. • Generate stock market information screen.
Iteration 7	<ul style="list-style-type: none"> • Retrieve more time series data. • Generate missing time series. • Retrieve temperature information to put on information screen. • Generate temperature information screen.
Iteration 8	<ul style="list-style-type: none"> • Review all stimuli. • Modify stimuli pairs according to feedback. • Correct spelling mistakes.

Table 5.1: Stimulus Generation Iteration

5.3 Software Development

We implement the software in *Java*, using *Swing* componenets to provide the graphical user interface (GUI) for the Java programs.

5.3.1 Software Development Iteration

Software took approximately eight weeks for completion.

Iteration 1	<ul style="list-style-type: none">• Design software workflow and demographic question screen.
Iteration 2	<ul style="list-style-type: none">• Redesign software workflow according to the feedback.• Design information screen and question screen.• Implement demographic question screen.
Iteration 3	<ul style="list-style-type: none">• Implement information screen and question screen.
Iteration 4	<ul style="list-style-type: none">• Add masking screen for each trial.• Implement reading timer and answering timer.
Iteration 5	<ul style="list-style-type: none">• Implement input reading function to read input data from a comma separated values (CSV) file.• Implement output generation function to keep the result.
Iteration 6	<ul style="list-style-type: none">• Add introductory screens.• Modify demographic question screen format.• Implement training session.• Test program flow.
Iteration 7	<ul style="list-style-type: none">• Improve software to enhance user interface and usability and ensure all needed results are recorded.• Implement question feedback for training session.
Iteration 8	<ul style="list-style-type: none">• Test the program on demo account.• Test the program on demo account running concurrently on 15 machines.• Prepare pre-study presentation.

Table 5.2: Software Development Iteration

5.4 Experiment

The experiment was conducted in six sessions, two for the pilot study, and another four for the real experiments. All sessions took place in a computer laboratory (Room 379) at the Department of Computer Science, University of Oxford.

In the first pilot study, there were two participants, one male and one female, who performed the experiment to evaluate the study design and implementation. The participants suggested that they misunderstood the time period of the electrocardiogram (ECG) time series to be four years. We concluded that we did not sufficiently explain about the time period in the presentation. Consequently, we made some changes to the presentation.

In the second pilot study, there was one female participant. The feedback from this study suggested that the length of the pre-study presentation was a little lengthy. However, we did not make any changes to the presentation length because participants in the real study may come from different backgrounds. Other than that, the pilot study went well with no potential problems.

In the real experiment, each session consists of 10-13 participants. We set the limit of the participants to 15 for each session in order to provide them with an adequate support. The following subsections will provide the details of the participants, apparatus, and procedure in the experiment.

5.4.1 Participant

A total of 47 participants took part in the experiment in return of a 10 Amazon voucher. One participant did not finish the experiment, and thus his result was not used in the calculation. This leaves a total of 46 participants in our study.

Among these, there were 30 males and 16 females. All of the participants were recruited from the University of Oxford and related communities from various disciplines such as Computer Science, Engineering, Mathematics, Materials, Oceanography, Anthropology, Applied Linguistics, Economics, and Public Policy. A diverse background of participant population ensured unbiased analyses results. From these 46 participants, 29 participants are university students, 10 participants are university staffs, and 4 participants specified their occupations as others. Out of 46 participants, 25 participants belong to 20-29 age group, 11 participants belong to 30-39 age group, 6 participants belong to 40-49 age group, 3 participants belong to 50-59 age group, and 1 participant belong to 60-69 age group. Figure 5.1 illustrates the participants'

demographics information whereas Figure 5.2 illustrates the participants' familiarity rating.

5.4.2 Apparatus

The stimuli were generated using Microsoft Excel. The custom software program were written in Java. The software was run in fullscreen mode, which disallows users from quitting the software and allows full concentration on the software as no background distractions are presented. The stimuli generation and software development are described in Sections 5.2 and 5.3. A pre-study presentation was given on a 10×12 inch projector before each session of the experiment started. The experiment was run on computers with 3.7 GB of RAM, 3.30 GHz quad-core Intel core i5-3550 processors running on Fedora, a Linux based operating system, with GNOME version 3.4.2. Each computer had 24-inch Dell's LCD at 1920×1200 (16:10) resolution and sRGB colour mode display. We adjusted the monitors to the same level of brightness and contrast. Each participant was required to interact with the software using the mouse on the desk. A total of 10-13 computers were used during each session.

5.4.3 Procedure

The time taken to complete this study was approximately 45-80 minutes in total, including 15 minutes of pre-study presentation. The time spent on the experiment varied according to the time that the participants took to perform each trial.

Prior to the experiment, the experimenter gave a brief introductory presentation to the participants. The presentation included the explanation on the time series type and the statistical measures used in the study, as well as the screenshots of the actual software that the participant would be interacting with. This was aimed to help the participants understand the software, accustom themselves with all elements in each trial, and ensure they understand the task correctly. The participants were acknowledged that they can take as much time as they need in each trial.

After the pre-study session, the information sheets and the consent forms were distributed for each participants to read and sign. This was to acknowledge the participants about the study in general and ensured they agreed to take part in the study. Then the participants were allotted a unique user ID to fill in the demographic form. This form requires the participant to enter their user ID, gender, age group, occupation, and familiarity rating in the software program.

After submitting the demographic form, the experiment began. They were required to undertake a training session consisting of three trials. After that, the participants were informed in the software program that the main experiment (study session) would begin.

The main experiment consists of 36 trials. In each trial the white noise was first shown to indicate that the new question has started. This was shown for two seconds before the information page appears. The participants could take as much time as they need to read this information page, before clicking the ‘Next’ button at the bottom of the page to proceed to the question page. Then the participants were required to select one of the optional answers in order to proceed to the next question.

When all the trials had been completed, participants were presented with a screen indicating completion of the experiment. They were required to complete a subjective questionnaire on the ease of identifying a time series when different kinds of information were given. Once they had done, they were acknowledged for their time and service by a 10 Amazon voucher.

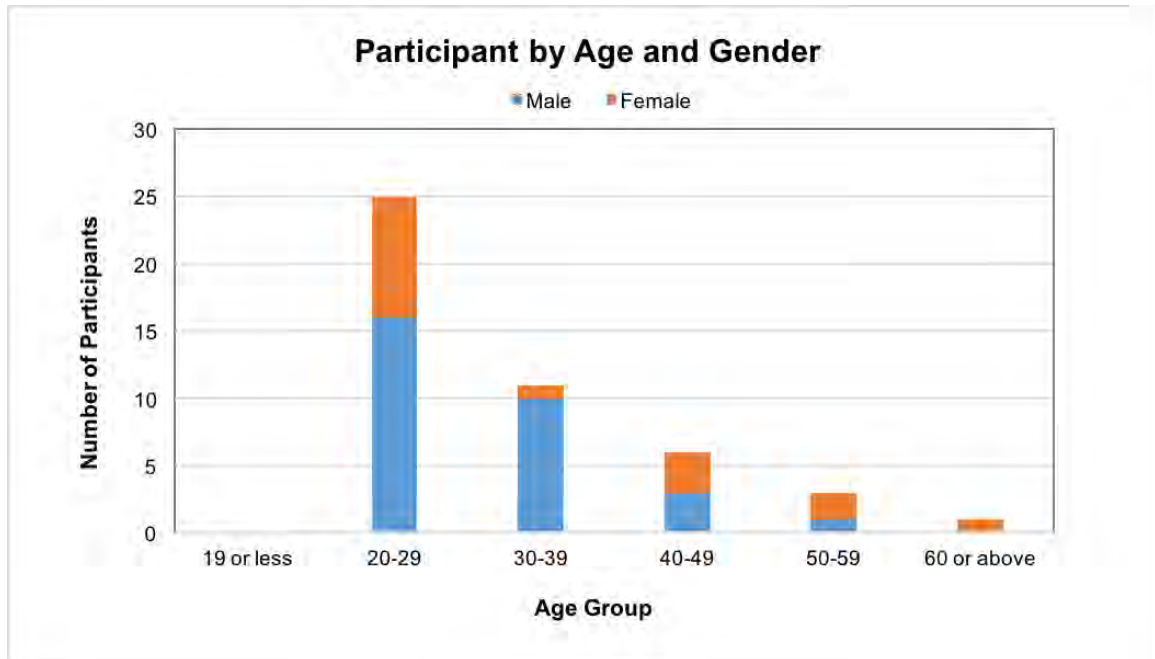


Figure 5.1: Demographics information including age and gender

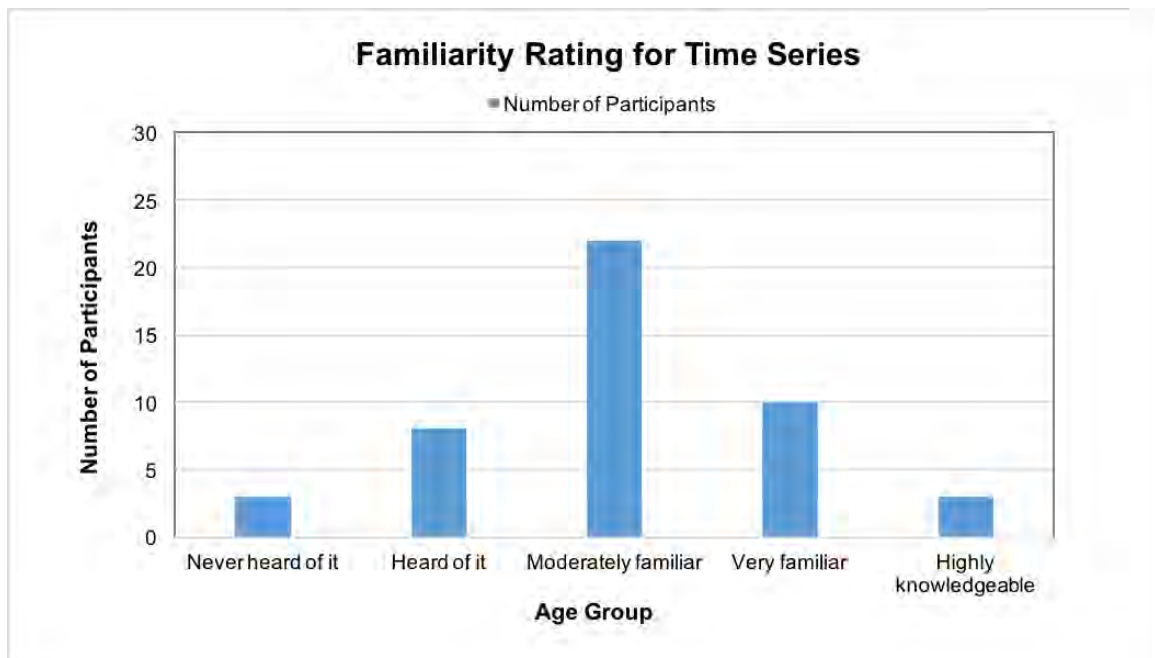


Figure 5.2: Familiarity rating for time series

Chapter 6

Result Analysis

We perform descriptive and inferential analysis on the results of the experiment. As part of descriptive analysis, mean and standard deviation of the result are calculated. Inferential analysis tests the formulated hypotheses and provides statistical evidence for the conclusion. We also analyse subjective rating for each type of statistical measures and time series along with other relevant statistics.

In this chapter, we first provide the summary of the result against the formulated hypotheses. Then we provide detailed analyses on the result of each information category. Finally, we provide analyses on information reading time and choice selection time.

6.1 Result Summary

The result analyses **confirm the three hypothesis**, formulated in Section 3.1. The results for the three categories of information in terms of accuracy are:

R1: Min/Max \succ Average \succ Standard deviation \succ Random guess

R2: Temperature \succ ECG \succ Stock market \succ random guess

R3: Global (simple) \succ Local (complex) \succ random guess

For simplicity, we use the uppercase letters H , S , and T to refer to ECG, stock market, and temperature time series respectively. The letter pair refers to the pairwise combination of time series type, which is explained in Section 3.3.2.2.

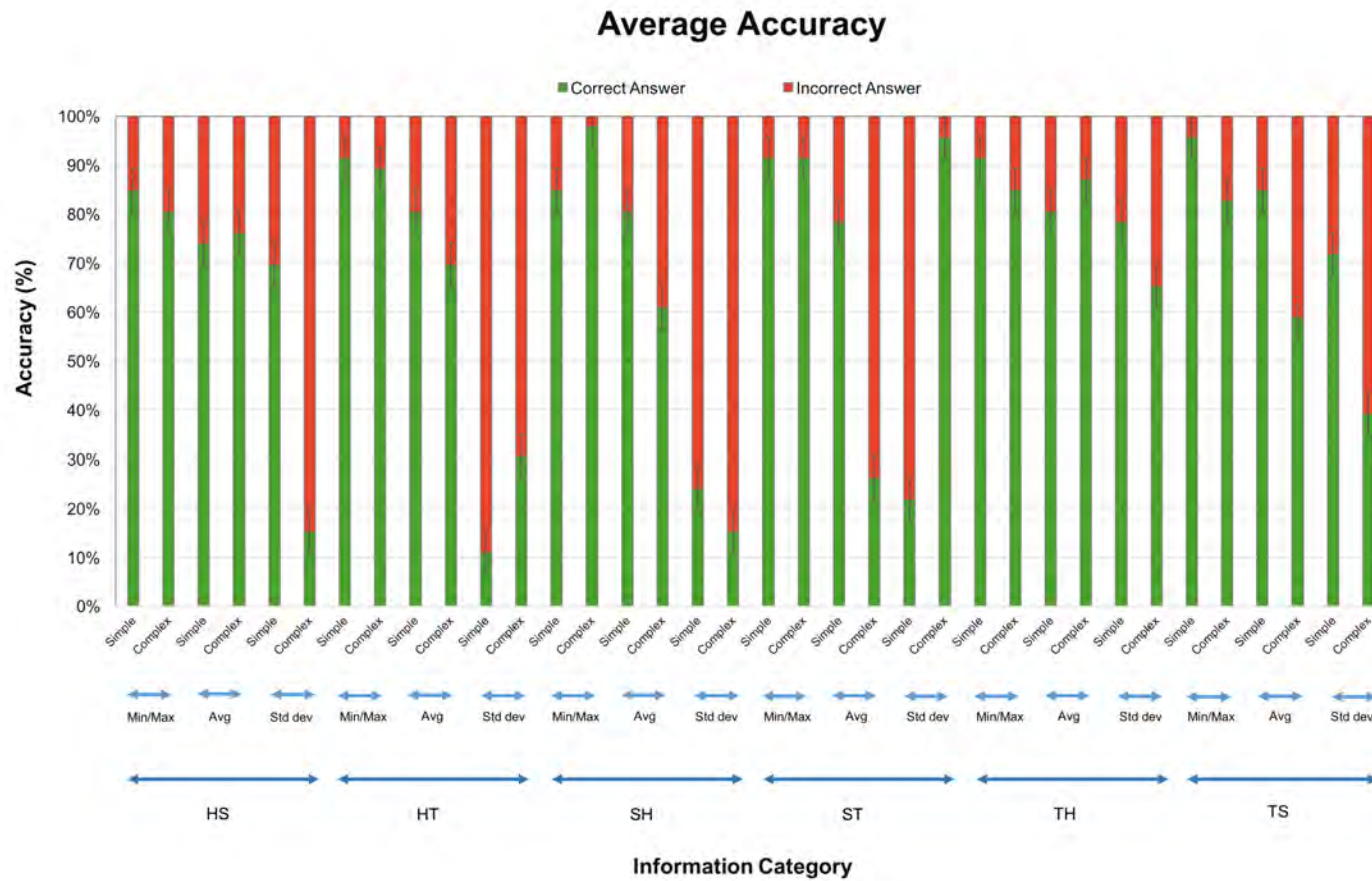


Figure 6.1: Overall average accuracy.

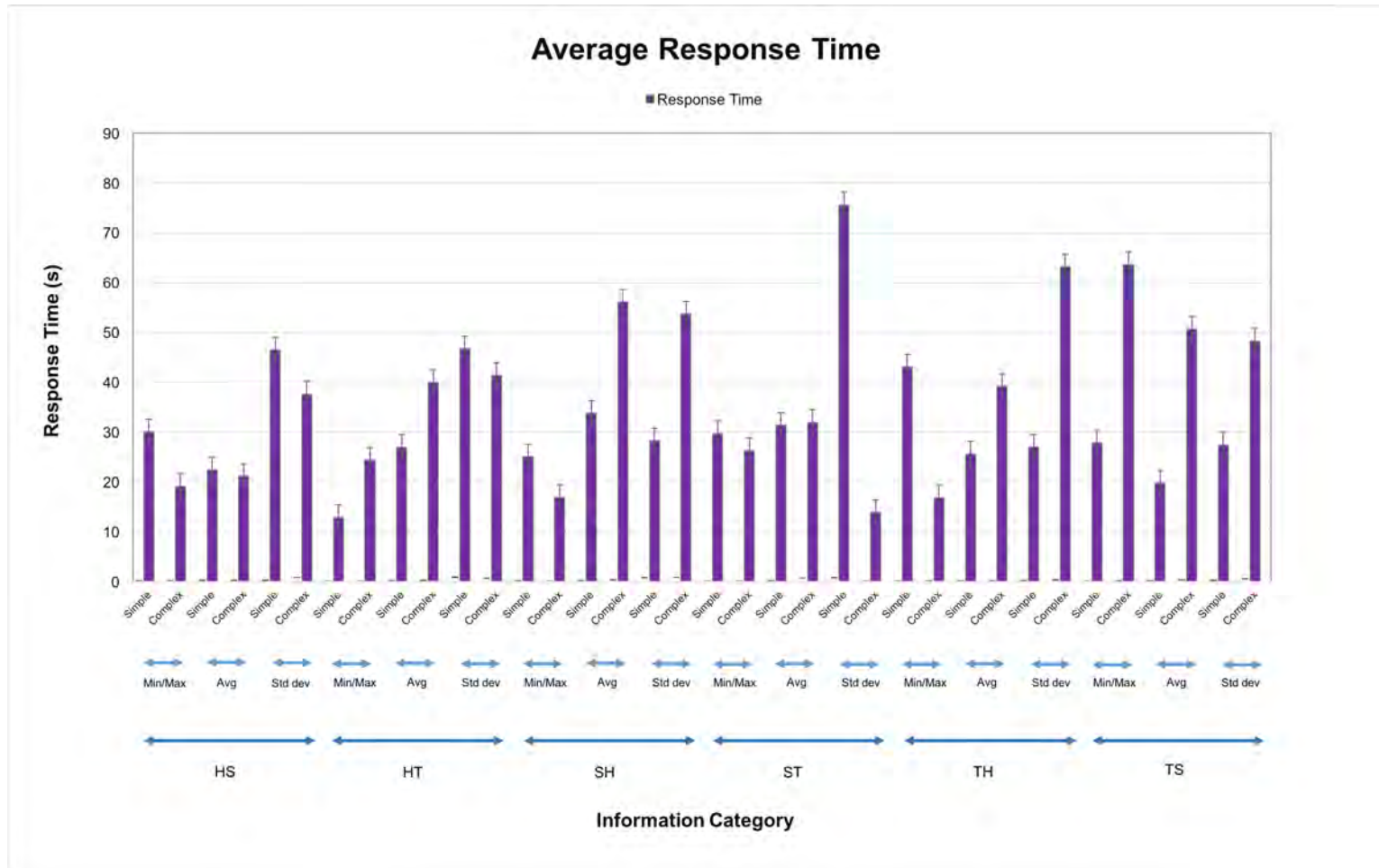


Figure 6.2: Overall average response time.

For statistical measure information (**R1**), minimum and maximum yields the overall highest accuracy, with relatively less response time, followed by the average value and standard deviation respectively.

For time series type information (**R2**), electrocardiogram (ECG) and stock market shows no significant difference in terms of accuracy. Temperature time series, however, display obvious higher accuracy than the two time series type. In terms of response time, ECG holds advantage over the other two types. This follows by stock market and temperature respectively.

For time series pattern information (**R3**), simple pattern displays advantages over complex pattern in terms of both accuracy and response time. Figure 6.1 illustrates the average accuracy for each information categories while Figure 6.2 demonstrates the average response time for each information category.

6.2 Result Analyses for Statistical Measure

The first category of information that we analyse is the statistical measure.

6.2.1 Accuracy

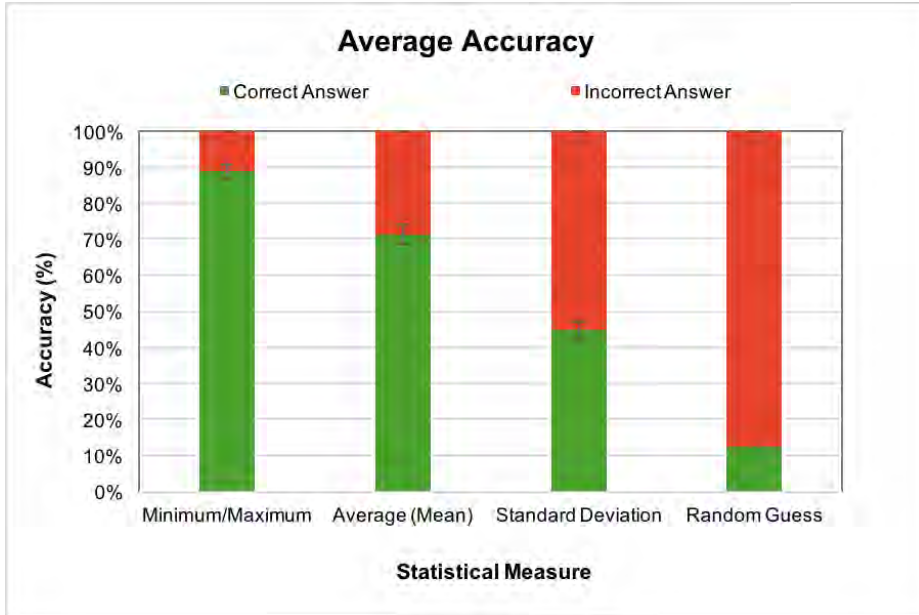


Figure 6.3: Average accuracy of statistical measure.

Figure 6.3 illustrates the accuracy for each type of statistical measure. The right-most column, *no information* is calculated based on the probability of answering the

question correctly out of eight optional answers when no statistical measure is given (12.5%). No information is equivalent to a *random guess*.

Mauchly's Test of Sphericity verifies that the assumption of sphericity had been met ($p = .423$). ANOVA analysis reports that there is a significant main effect of the statistical measures in accuracy ($F(2, 90) = 191.04, p < .001$).

Further t -test analysis reveals that all types of statistical measures are significantly different from each other. Minimum and maximum values yield the overall highest accuracy, followed by average value and standard deviation respectively (all $p < .001$).

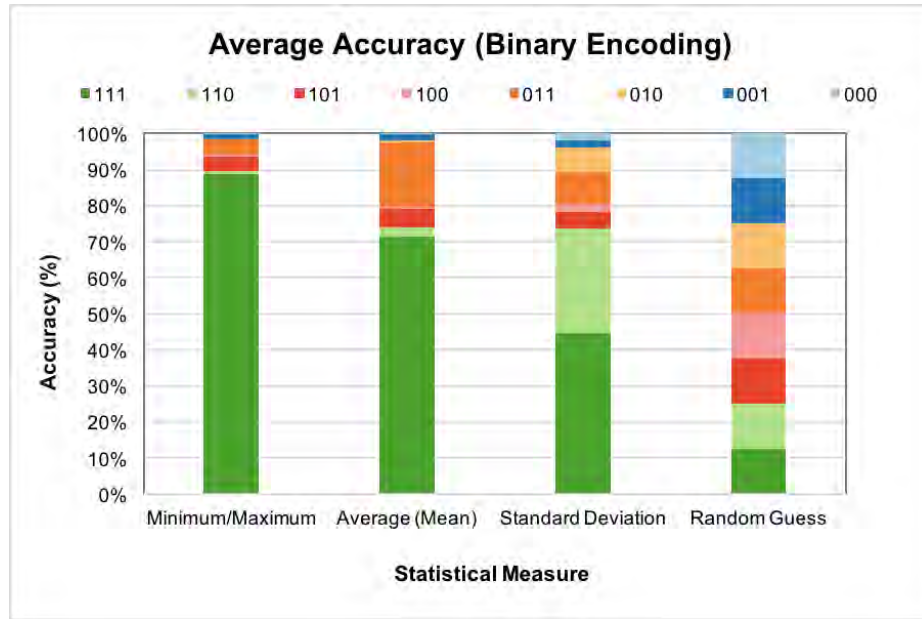


Figure 6.4: Average accuracy of statistical measure (binary encoding).

Further analyses on the average accuracy is illustrated in Figure 6.4. As opposed to the Boolean correctness (true or false) used in 6.3, we analyse the average accuracy of each statistical measures based on the binary encoding of each optional answer that we defined in Section 4.4.2, which will be restate briefly again. This further reveals the types of error that the participants made.

For each error type, which is a three-digit binary number, the correctness of time series type, time series pattern, and statistical measure are corresponds to each digit from left to right. Thus, if the participant selected the option that corresponds with 110, it means that they have chosen the time series with correct type and pattern, but the statistics is incorrect. It is clear that the participants have chosen more 110 option when they were given a standard deviation than other types of statistical measure.

Figure 6.5 illustrates the Boolean correctness of identifying the time series with the correct value of statistical measure. Thus, if the participant chooses any options

that ends with 1 (001, 011, 101, and 111), we consider that the participants can correctly identify the statistical measure. This is because the last digit represents the correctness of the statistical measure.

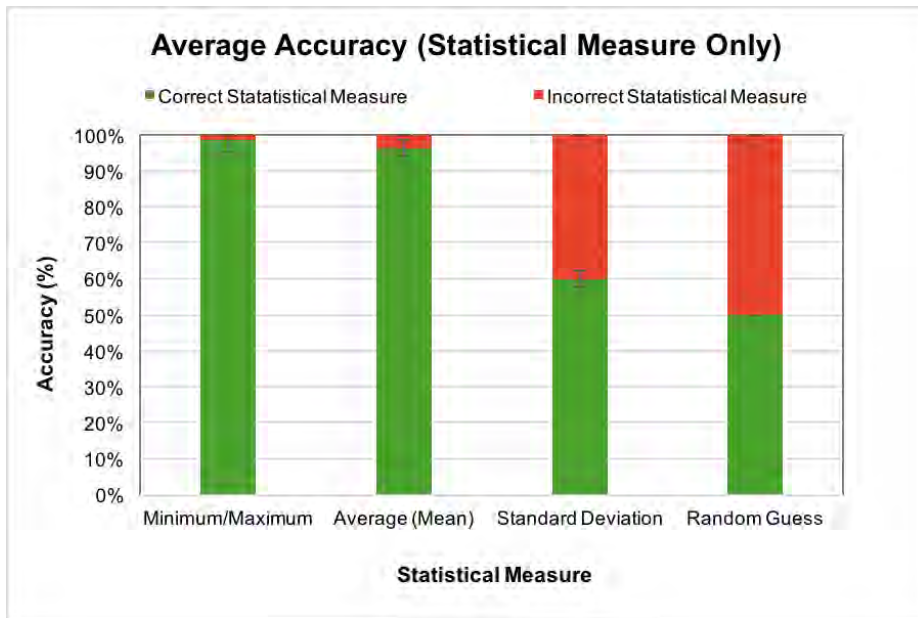


Figure 6.5: Average accuracy of statistical measure (statistical measure only).

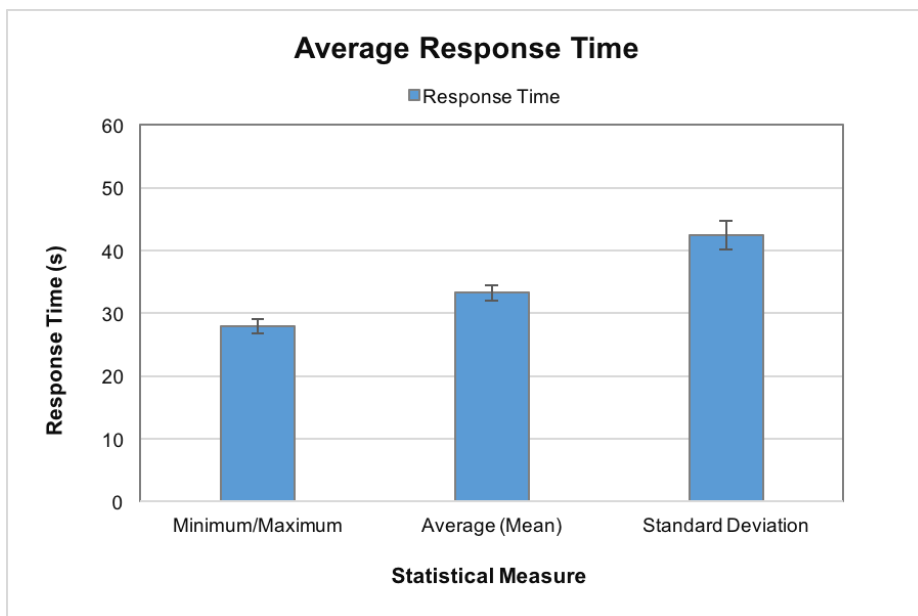


Figure 6.6: Average response time of statistical measure.

6.2.2 Response Time

The average response time for each statistical measure is shown in Figure 6.6.

Mauchly's Test of Sphericity reports that the assumption of sphericity had been violated ($p < .001$). As a result, the ANOVA analysis with Huynh-Feldt Corrections ($\epsilon = 0.777$) have been used. The result suggests that there is a significant main effect of the statistical measure in accuracy ($F(1.553, 69.917) = 34.217, p < .001$)

Further t -test analysis establishes that all types of information are significantly different from each other. Minimum and maximum values yield the overall fastest response time, followed by average value, and standard deviation respectively (all $p < .001$).

6.2.3 Performance Summary

Among the three statistical measures, minimum and maximum value is clearly the better information for identifying a time series. Standard deviation, on the other hand, yields the lowest performance among all types, both in terms of accuracy and response time.

The relative performance relationships among the three types of statistical measure in accuracy and response time is demonstrated as follow:

Accuracy

Min/Max (10.652) \succ Average (8.565) \succ Standard deviation (5.370)

Response time

Min/Max (27.958) \succ Average (33.235) \succ Standard deviation (42.443)

6.2.4 Difficulty Rating

Participants' subjective rating on the difficulty of each statistical measure is illustrated in Figure 6.7. This is consistent with the result from ANOVA and t -test analyses, which suggests that minimum and maximum values are most effective for narrowing down the eight options of time series. Almost 90% of the participants agreed that it was easy to narrow down the optional answers when they are given minimum or maximum value. For mean value, approximately half of the participants find it easy. Less than 5% of the participants report that standard deviation is an

effective choice of statistical measure. However, regardless of an extremely low confidence, we would like to emphasise the fact that the average accuracy of choosing the time series that has a corresponding standard deviation is still higher than a random guess.

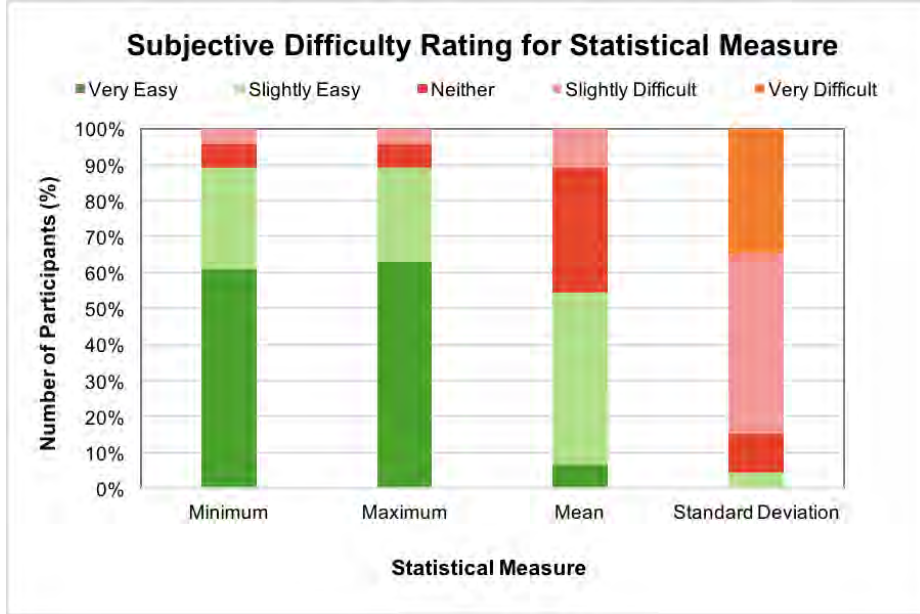


Figure 6.7: Participants' difficulty rating for statistical measures.

6.3 Result Analyses for Time Series Type

The second category of information that we analyse is the time series type. For this category of information, we analyse both the general type and the pairwise combination of the type separately. There are three general time series type in this study that need to be analysed: electrocardiogram (ECG), stock market, and temperature. We will analyse the general time series type in this section and continue further analysis on the pairwise combination in Section 6.5.1.

6.3.1 Accuracy

Mauchly's Test of Sphericity verifies that the assumption of sphericity had not been violated ($p = .991$). ANOVA analysis reports that there is a significant main effect of the time series type in accuracy ($F(2, 90) = 14.147, p < .001$).

Further t -test analysis suggests that temperature time series is the source of the main effect, yielding the highest accuracy relative to the other two time series type

(all $p < .001$). However, there is no significant difference between ECG and stock market time series ($p = 1$).

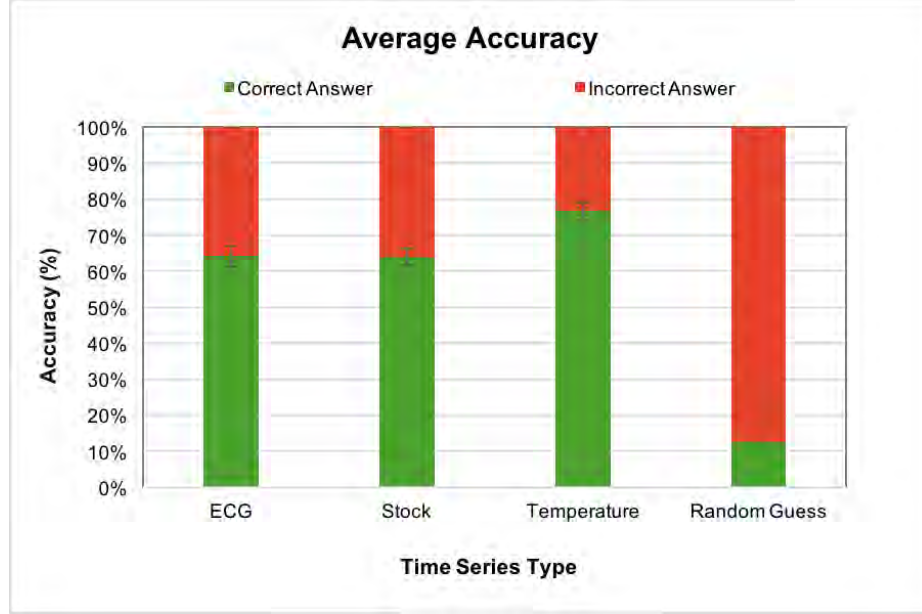


Figure 6.8: Average accuracy of time series type.

Further analyses on the average accuracy is illustrated in Figure 6.9. We analyse the average accuracy of each general time series type based on the binary encoding of each optional answer. This helps to understand the types of error that the participants made. We can observe that the participants equally likely made 110 (incorrect statistics) and 011 (incorrect time series type) errors for each time series type. It is clear that the participants made less 110 error for temperature time series than the other two types.

Figure 6.10 illustrates the Boolean correctness of identifying the time series of the specified type. Hence, if the participant chooses any options that begins with 1 (100, 101, 110, and 111), we consider that the participants can correctly identify the time series type. This is because the first digit represents the correctness of the time series type.

6.3.2 Response Time

Mauchly's Test of Sphericity verifies that the assumption of sphericity had not been violated ($p = .208$). ANOVA analysis reports that there is a significant main effect of the time series type in response time ($F(2, 90) = 16.048, p < .001$).

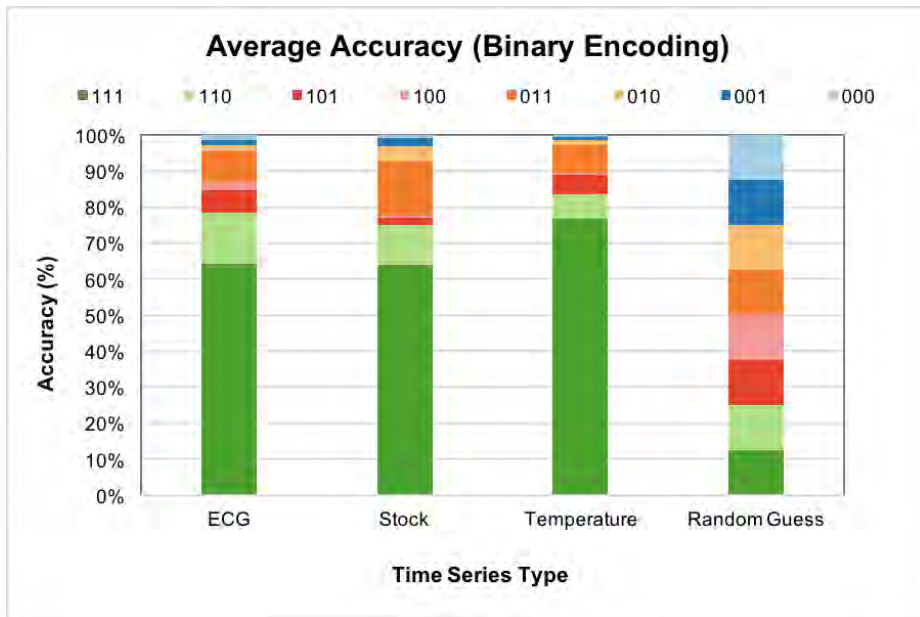


Figure 6.9: Average accuracy of time series type (binary encoding).

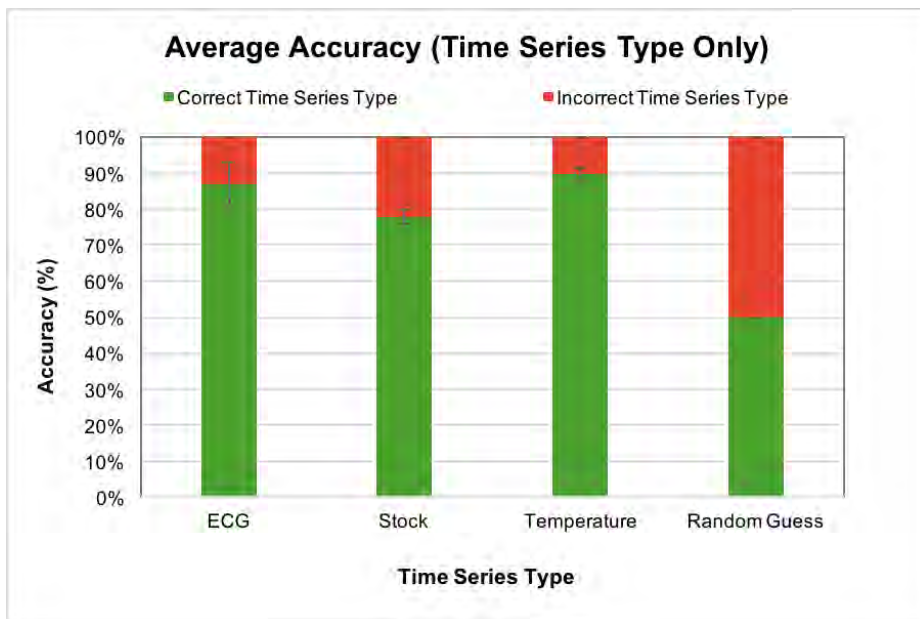


Figure 6.10: Average accuracy of time series type (time series type only).

Further *t*-test analysis suggests that ECG yields the fastest average response time relative to the other two types (all $p < .001$). However, there is no significant difference in response time between stock market and temperature time series ($p = 0.074$).

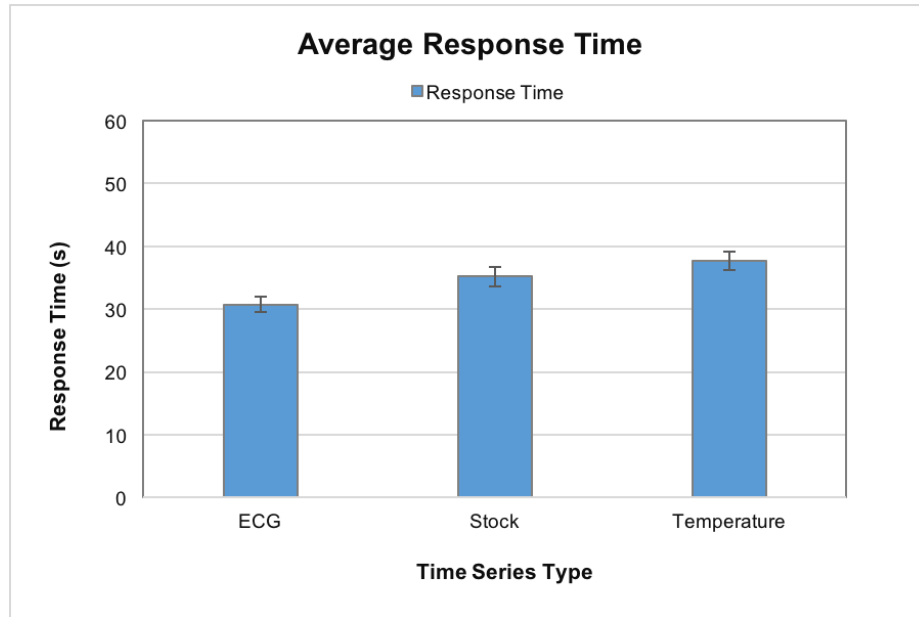


Figure 6.11: Average response time of time series type.

6.3.3 Performance Summary

All three types of time series is clearly better than a random guess in identifying a time series. Temperature time series yields highest accuracy, but also highest response time.

The relative performance relationships among the three types of time series in accuracy and response time is demonstrated as follow:

Accuracy

Temperature (9.196) \succ ECG (7.717) \succ Stock market (7.674)

Response time

ECG (30.754) \succ Stock market (35.201) \succ Temperature (37.680)

6.3.4 Difficulty Rating

Participants' subjective rating on the difficulty of identifying each time series is illustrated in Figure 6.12. The chart suggests that about 50% of the participants find ECG and temperature easy. While the other 50% report that it is difficult to identify ECG and temperature time series correctly. Less than 10% of the participants suggest

that temperature time series is difficult to spot.

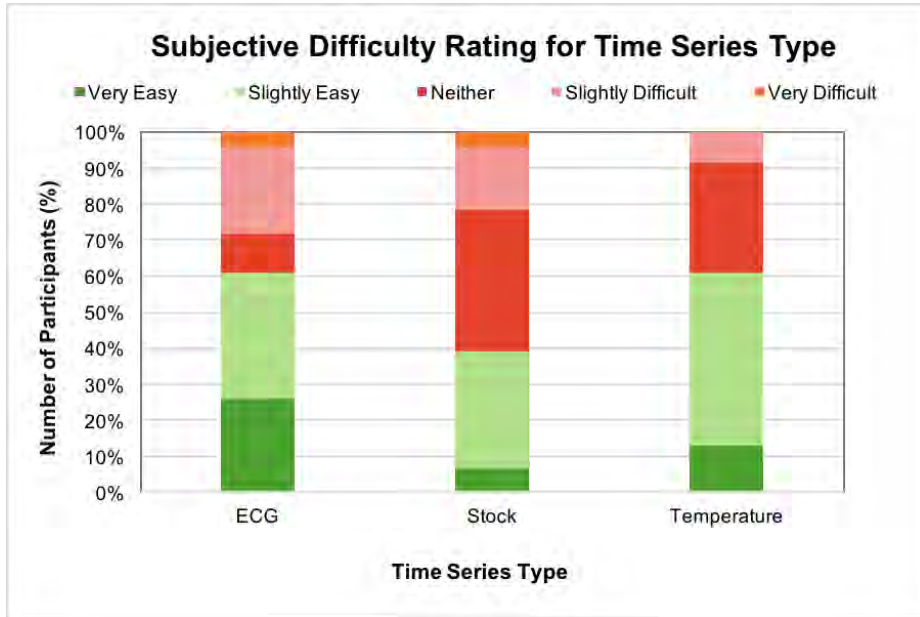


Figure 6.12: Participants' difficulty rating for time series type.

6.4 Result Analyses for Time Series Pattern

The third category of information that we analyse is the time series pattern.

6.4.1 Accuracy

Since there are only two conditions for this independent variable, Mauchly's Test of Sphericity is not necessary. ANOVA analysis reports that there is a significant main effect of the time series pattern in accuracy ($F(1, 45) = 15.526, p < .001$).

Further analyses on the average accuracy is illustrated in Figure 6.14. We analyse the average accuracy of each time series pattern based on the binary encoding of each optional answer.

Figure 6.15 illustrates the Boolean correctness of identifying the time series that has a corresponding general pattern. Hence, if the participant chooses any options that have 1 in the middle (010, 011, 110, 111), we consider that the participants can correctly identify the time series type. This is because the second digit represents the correctness of the time series pattern.

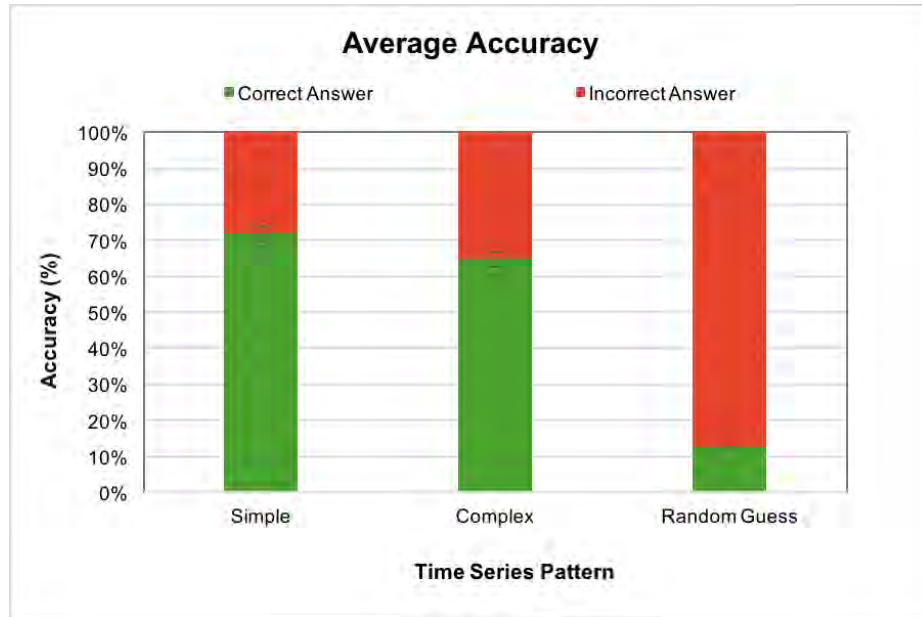


Figure 6.13: Average accuracy of time series pattern.

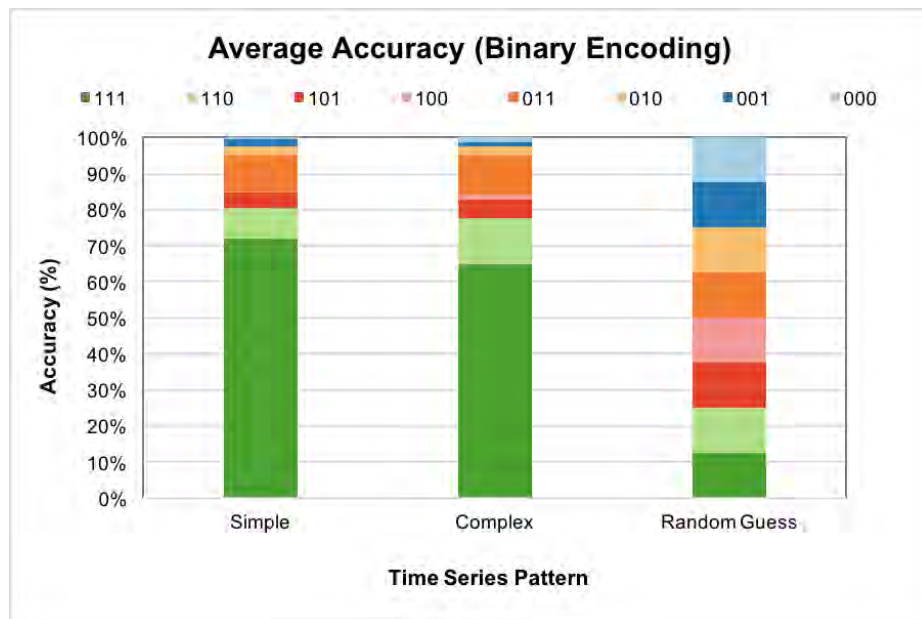


Figure 6.14: Average accuracy of time series pattern (binary encoding).

6.4.2 Response Time

Since there are only two conditions for this independent variable, Mauchly's Test of Sphericity is not necessary. ANOVA analysis reports that there is a significant main effect of the time series pattern in response time ($F(1, 45) = 18.038, p < .001$).

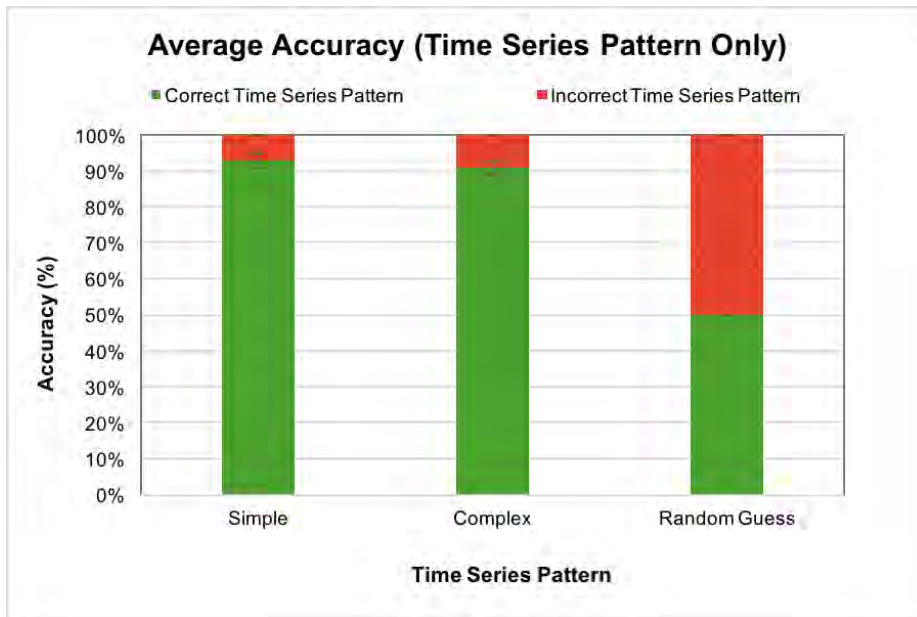


Figure 6.15: Average accuracy of time series pattern (time series pattern only).

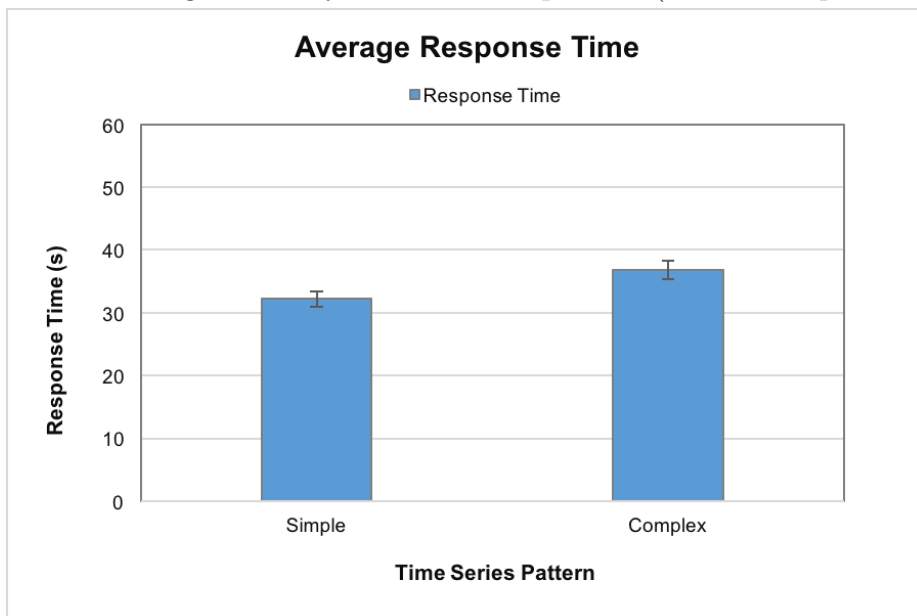


Figure 6.16: Average response time of time series pattern.

6.4.3 Performance Summary

The simple pattern (local) yields higher accuracy and lower response time than the complex pattern (global).

We did not perform subjective rating on the pattern type as the type criteria was not explicitly given to the participants.

Accuracy

Simple (12.935) \succ Complex (11.652)

Response time

Simple (32.211) \succ Complex (36.879)

6.5 Further Analysis

6.5.1 Pairwise Combination of Time Series Type

There are a total of six possible pairwise combination of the three general time series type. These include HS , HT , SH , ST , TH , TS .

6.5.1.1 Accuracy

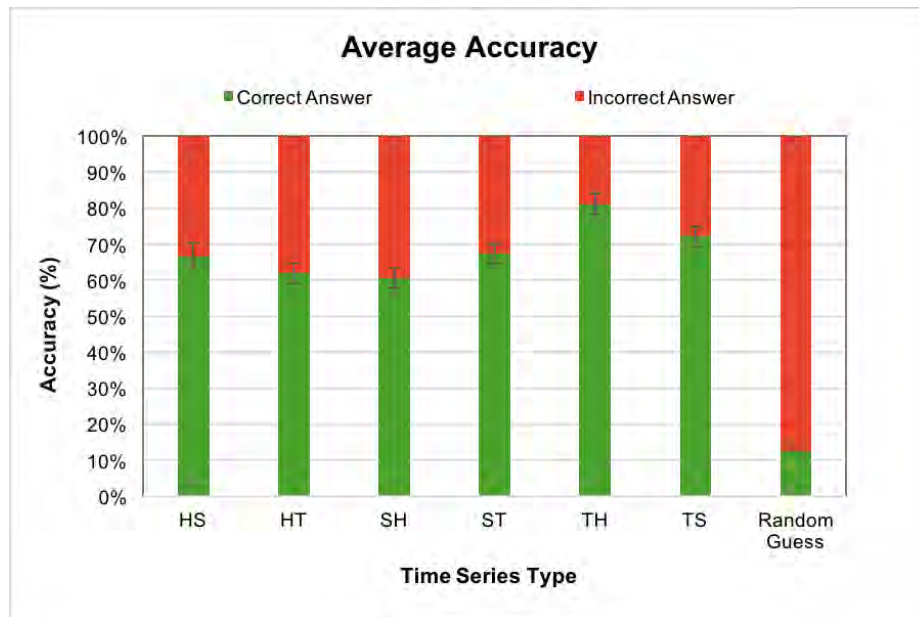


Figure 6.17: Average accuracy of pairwise type.

Mauchly's Test of Sphericity reports that the assumption of sphericity had been violated ($p = .027$). Consequently, the ANOVA analysis with Huynh-Feldt Corrections ($\epsilon = .887$) have been used. The result suggests that there is a significant main effect of the time series type in accuracy ($F(4.437, 199.666) = 10.552, p < .001$)

Further t -test analysis establishes that the combination TH and TS are the causes of the main effect, yielding higher accuracy than the other combinations. The most significant effects can be seen when comparing TH with HT and SH (both $p < .001$) follow by HS ($p = .001$), ST ($p = .002$), and TS ($p = .032$). TS yields slightly higher accuracy than HT ($p = .043$) and SH ($p = .049$). However, there is no significant difference in accuracy between TS and HS , TS and ST , HS and HT , HS and SH , HS and ST , HT and SH , and HT and ST (all $p = 1$).

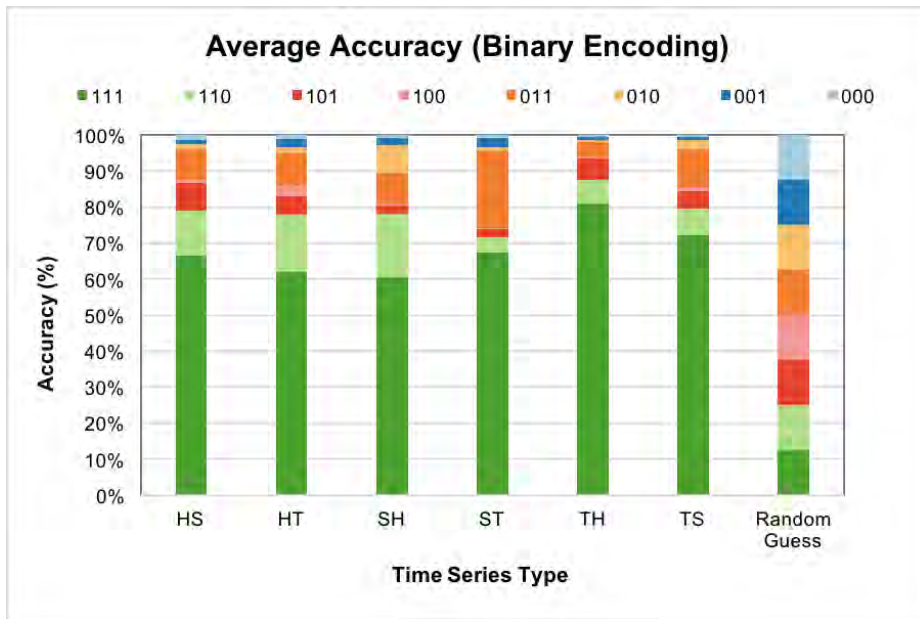


Figure 6.18: Average accuracy of pairwise type (binary encoding).

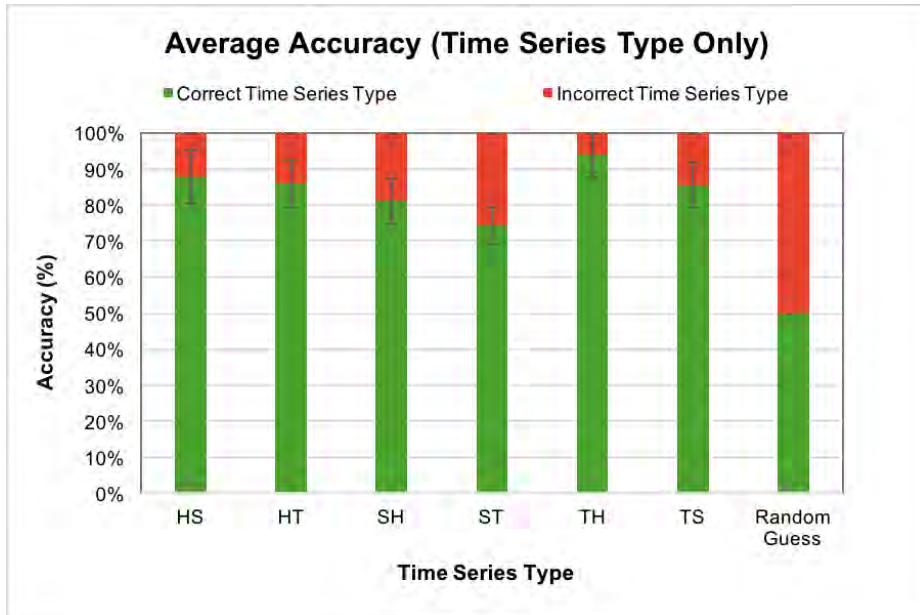


Figure 6.19: Average accuracy of pairwise type (time series type only).

Further analyses on the average accuracy is illustrated in Figure 6.18. We analyse the average accuracy of each time series type combination based on the binary encoding of each optional answer. This helps to understand the types of error that the participants made. We observe that 011 error (incorrect time series) is highest in *ST*. This suggests that distinguishing stock market from temperature time series may be difficult.

Figure 6.19 illustrates the Boolean correctness of identifying the time series of the specified type. Hence, if the participant chooses any options that begins with 1 (100, 101, 110, and 111), we consider that the participants can correctly identify the time series type. This is because the first digit represents the correctness of the time series type.

6.5.1.2 Response Time

Mauchly's Test of Sphericity establishes that the assumption of sphericity had not been violated ($p = .072$). ANOVA analysis reports that there is a significant main effect of the time series type in response time ($F(5, 255) = 7.633, p < .001$).

Further *t*-test analysis suggests that the combination *TS* yields faster response time than the other combinations. The only significant effects can be seen is when comparing *TS* to *HS* ($p < .001$).

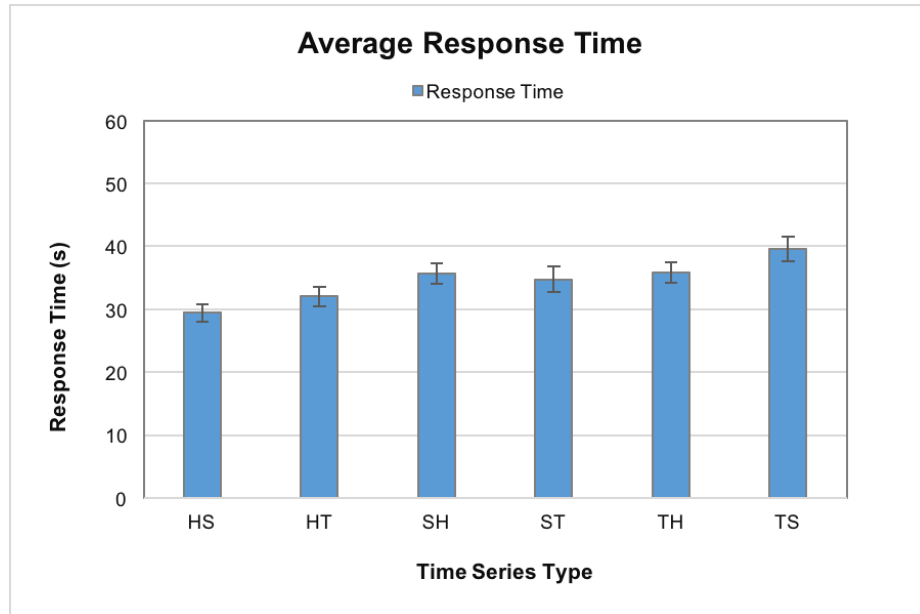


Figure 6.20: Average response time of pairwise type.

Chapter 7

Conclusion

The goal of this project is to evaluate the effects of different categories of information on the process of reconstruction.

We have achieved our goal and have done the project according to the planned schedule. The completed works include:

1. Study cognitive psychology concepts of object recognition and categorisation, cost-benefit analysis [CG16], time series, and foundation of empirical study in general.
2. Design the empirical study by formulating the hypotheses, identifying variables, indicating measurement metrics, and specifying result analyses' techniques.
3. Design the task, the stimuli, and the software for the experiment, and develop them accordingly.
4. Conduct and empirical study using the developed software to collect user performance in the experiment.
5. Perform descriptive and inferential analyses to analyse the study's results and provide conclusion for the study.

7.1 Summary

Cost-benefit analysis [CG16] is fundamental theoretic measure that allows for an objective comparison between machine-centric and human-centric processes. To our knowledge, our project is the first empirical study that formally assessed the cost-benefit measure proposed by Chen and Golan [CG16].

Our project quantitatively evaluates the effects of different types of information on the reconstruction quality of a time series. Empirical study was conducted to make such an evaluation, in terms of accuracy and response time. The hypotheses on the difference of the performance mean among each type of information in each category were formulated. A reconstruction process of time series in this study involved using the given information about the time series to select a time series that satisfies all the given information. Three categories of information were given including statistical measure, time series type, and time series pattern. They represent machine-processed information, human soft knowledge information based on memory, and human soft knowledge information based on pattern recognition, respectively. For simplicity, we limited the number of time series that can be reconstructed to eight optional choices. The result from the experiment were analysed using descriptive and inferential analyses.

The stimuli were carefully designed and generated. Since the reconstruction quality is based on human-centric information such as known theories and past experiences, the time series data used in this study were required be taken from the real-world examples rather than artificially generated. This was also true for the information contained in the information screen. Besides the validity of the information, we had to design the information screen to be visually pleasing and the content needed to be interesting to avoid confounding effects. The software was developed and tested to ensure a precise user data collection.

The results suggest that human participants use minimum and maximum values, among other statistical indicators, to reconstruct time series most effectively. Regarding time series type, participants are more accurate when reconstructing temperature time series than the other two types, but they take the longest time to perform. For electrocardiogram time series, although the accuracy is slightly less than temperature time series, it takes the least amount of time to be reconstructed. When using information about time series pattern to reconstruct a time series, participants are more effective in reconstructing the simple pattern (global) than the complex pattern (local). More importantly, all the types of information given to the participant is significantly more effective for reconstruction process than a random guess.

Additionally, the results also show that machine-processed information alone is not good enough for reconstruction quality. By incorporating human soft knowledge through memory recall or pattern recognition, or both, human participants yield significantly higher performance.

The quantitative result is consistent with the subjective feedback on the difficulty of using each statistical indicator and time series type. Participants find it easiest to reconstruct a time series when minimum value and maximum value are given compared to average value and standard deviation. For time series type information, they find it least difficult to reconstruct time series of temperature type, compared to electrocardiogram and stock market.

7.2 Evaluation

To our knowledge, our research is the first empirical study that aims to understand the probabilistic nature of human knowledge, which are difficult to quantify, through the concept of information-theoretic metric called cost-benefit analysis proposed by Chen and Golan [CG16]. Our study successfully evaluate the effectiveness of each category of information, including those that require human intelligence to perceive, for reconstruction of a time series. Additionally, the study demonstrates that the effectiveness of all types of information is higher a random guess in reconstruction process. The validity of the result has shown that the metric proposed by Chen and Golan [CG16] is a valid evaluation of human ‘soft’ knowledge.

The study by Kothari [Kot15] shows the advantage of integrating human visual and cognitive abilities in machine-based decision making algorithm, which is a decision tree in this case. His research suggests that background knowledge possessed by human could be used in the classification process, which agrees with the background notes described in Section 2.1 on memory, pattern recognition, and reconstruction. Yet, his evaluation compares human-integrated framework with the entirely machine-based decision making algorithm. Our research focuses more on the comparison of the reconstruction quality between when performing a random guess, when semantic knowledge is given, and when statistical indicator is given.

The limitation in our study is the project schedule, which was a week late from what was originally planned. The cause of such a delay was the design of the stimuli, which required a reasonably large amount of time to complete. The design process that took the longest time was searching for a valid pattern along with its description. For example, when designing ECG stimuli, descriptions of several patterns were available. However, in order to find the actual data with those patterns, we had to go through each of the database in PhysioNet [GAG⁺00]. Although the search process took up a large amount of time, after all the data were retrieved, data manipulation

and visualisation image generation were done quickly. Thus, we were able to carefully conduct the experiment and analyse the result without a rush.

With the completed work, we have formally assessed Chen and Golan’s [CJ10] cost-benefit analysis measure. We also hope that the project would be useful for researchers to understand the probabilistic nature of knowledge derived from human intelligence and further examine each information’s effectiveness in the reconstruction process.

7.3 Future Work

There are a number of possible directions that can be extended from our research. First of all, more variations on the total number of information that is given to the participants may be used. Comparison among *not providing any information*, *provide only statistical measure*, and *provide statistical measure and time series type* can be investigated. Second, the effects of the time series type can be evaluate further by participants that have more in-dept knowledge on that time series type (professional users). Furthermore, time series types that are more closely related than the chosen ones may be further evaluate against the professional users.

In conclusion, further evaluations with different design can be carried out to deeper evaluate the reconstruction process. We hope that our study will be useful as it gives a significant evidence on the importance of human intelligence in data analysis.

Appendix A

Stimuli in the Study

Here are the examples of the stimuli used in the real experiment.

The image shows a web form titled "Demographic Questions". It contains the following sections:

- User ID:** A text input field with a "Enter User ID" placeholder.
- Age:** A group of radio buttons with options: "19 or less", "20 - 29", "30 - 39", "40 - 49", "50 - 59", and "60 or above".
- Gender:** A group of radio buttons with options: "Female" and "Male".
- Occupation:** A group of radio buttons with options: "Faculty member", "Staff", "Student", and "Others".
- Familiarity with time series:** A group of radio buttons with options: "Never heard of it before the study", "Heard of it, but do not understand it", "Moderately familiar", "Very familiar", and "Highly knowledgeable".

At the bottom of the form is a "Submit" button.

Figure A.1: Demographic questions

Normal ECG

In medicine, an **electrocardiogram (ECG)** is a record of electronic impulses generated from the heart. The doctor can use this ECG record to diagnose various heart conditions.

In a **normal ECG**, the waves are **regularly spaced and the baseline does not wander up and down**. One of the easiest components to spot is a spike, which is called a QRS complex.

Can you tell which of the time series represents **a normal ECG**?

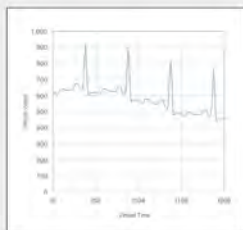


Source: <http://www.mayoclinic.org/tests-procedures/electrocardiogram/basics/definition/prc-20014152>

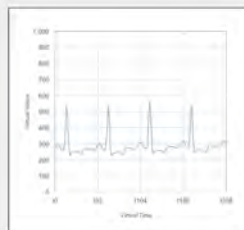
Next >

Figure A.2: Question 1

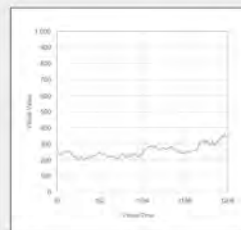
Which of the following plots represents a time series that has **a minimum value of 437** and likely reflects the information given in the previous page?



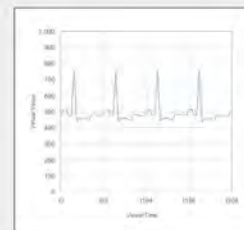
A



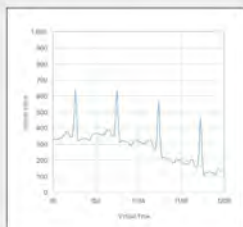
B



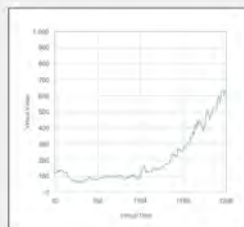
C



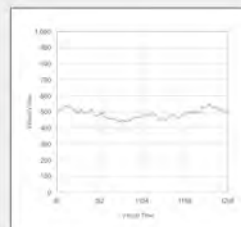
D



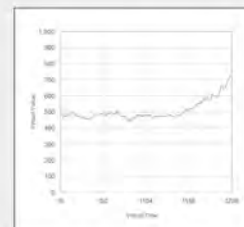
E



F



G



H

Next >

Figure A.3: Question 1

Morning News

VOL. 1, NO. 16
AUGUST 1, 2016




BEAR MARKET

In *financial trading*, trending market is a market that has a trend in one direction or another. A "bull" market is trending upward, while a "bear" market is *trending downward*.

The use of "bull" and "bear" to describe markets comes from the way the animals attack their opponents. A bull thrusts its horns up into the air while a bear swipes its paws down. These actions are metaphors for the movement of a market.

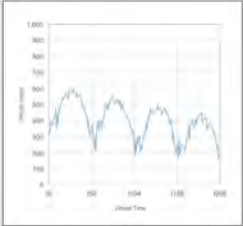
Which of the time series represents a *bear market*?

Source: <http://www.investopedia.com/terms/b/bullmarket.asp>

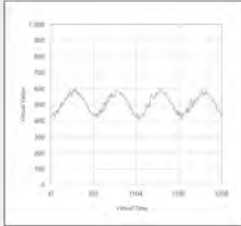
Next >

Figure A.4: Question 2

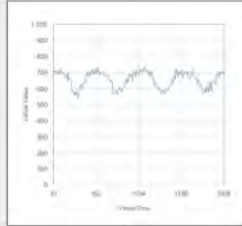
Which of the following plots represents a time series that has **a maximum value of 601** and likely reflects the information given in the previous page?



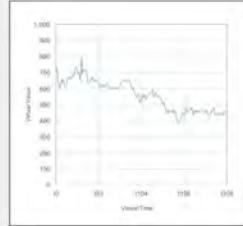
A



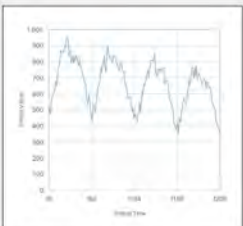
B



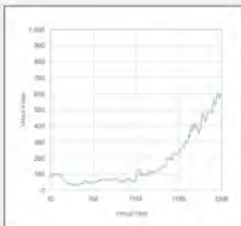
C



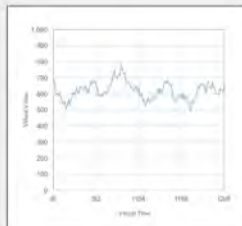
D



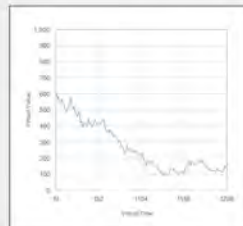
E



F



G



H

Next >

Figure A.5: Question 2

17 | MORNING NEWS

WEATHER

The Weather

TODAY
62|37
clearing sky,
partly cloudy

TOMORROW
58|41
clear, pleasant
cloudy

THE EQUATOR AND THE HEMISPHERES

The Earth is divided into two parts by an imaginary line called the equator. The two parts are named as Northern Hemisphere and Southern Hemisphere.

Countries in the Northern Hemisphere include, for example, the United States, the United Kingdom, and most of Asia.

A typical temperature of the Northern Hemisphere country is lowest in the beginning and at the end of the year. The temperature is highest around mid-year. A one-year temperature time series resembles a **bell curve**. The pattern then repeats in a cycle for each year.

Can you spot a **four-year** temperature time series of a **Northern Hemisphere country**?

Source: <http://www.kashif-ali.com/2010/03/23/northern-southern-hemispheres/>

Next >

Figure A.6: Question 3

Which of the following plots represents a time series that has **a minimum value of 297** and likely reflects the information given in the previous page?

A

B

C

D

E

F

G

H

Next >

Figure A.7: Question 3

Muscle Tremor

Muscle tremor is a kind of noise in the electrocardiogram (ECG). It is displayed in the ECG as **irregular spiky interference at the baseline**. Note that the R waves (peaks) of the ECG are still observable.

In order to reduce muscle tremors in the ECG, check the following:

- Patient's shoulders are relaxed
- Patient's arms are relaxed at their side, and their hands/fingers are still and not clenched
- Patient is not cold or shivering
- Patient is not talking

Which time series represents **an ECG containing muscle tremor**?

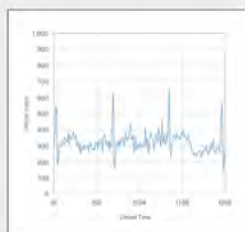


Source: <https://heartlearning.org/labyrinths?id=41221&parent=41275&sessID=1>

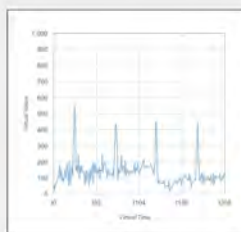
Next >

Figure A.8: Question 4

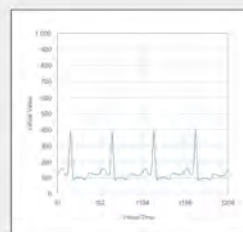
Which of the following plots represents a time series that has **an average value of 134** and likely reflects the information given in the previous page?



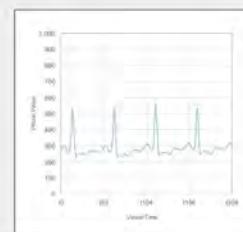
A



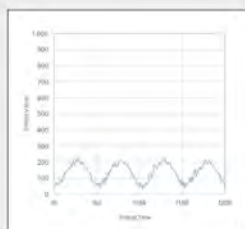
B



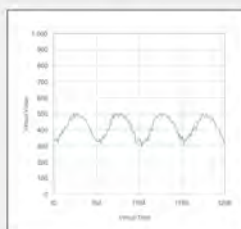
C



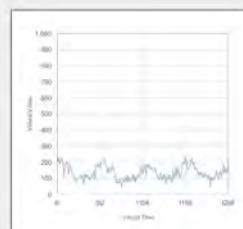
D



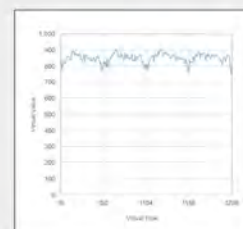
E



F



G



H

Next >

Figure A.9: Question 4

Morning News

VOL. 1, NO. 28
AUGUST 1, 2016

BLACK FRIDAY IN FINANCIAL MARKETS



Black Friday marks the beginning of the Christmas shopping season. The idea behind this term is that this is the day in which retail stores have enough sales to put them "in the black" - an accounting expression that alludes to the practice of recording losses in red and profits in black.



However, the term "black" has also been used to describe disastrous days in financial markets. Thus, Black Friday can also refer to **a one-day sharp drop in the stock market**.

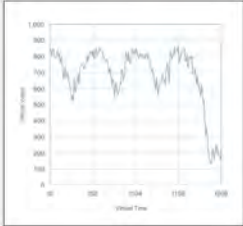
Can you spot a *sharp drop in a stock price*?

Source: <http://www.investopedia.com/terms/b/blackfriday.asp>

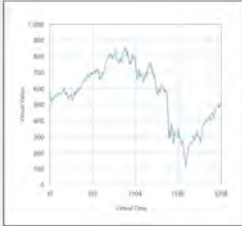
Next >

Figure A.10: Question 5

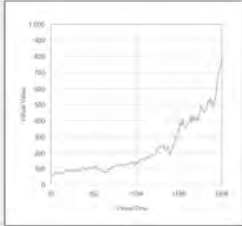
Which of the following plots represents a time series that has **a standard deviation of 103** and likely reflects the information given in the previous page?



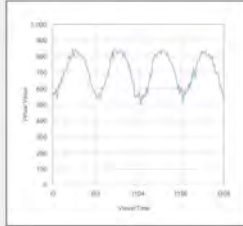
A



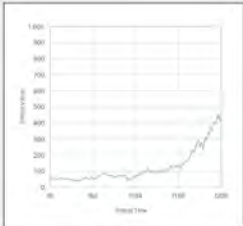
B



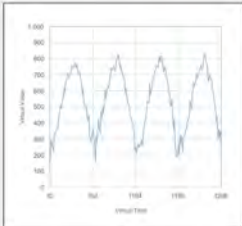
C



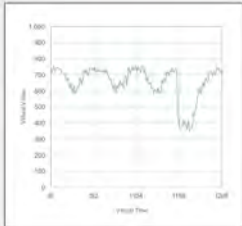
D



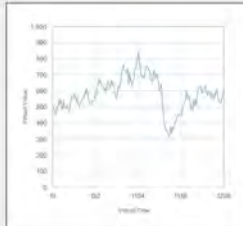
E



F



G



H

Next >

Figure A.11: Question 5

The Weather



WINTER IN ALASKA

You might be wondering what winter is like in Alaska. In the winter, the temperature in Alaska is going to be between 40° to -40° F. *The temperature tends to fluctuate between the two extremes during winter.* Yet, it is not difficult to see a clear bright night sky filled with stars, a large moon on most nights. And if one is lucky you get to see the aurora borealis.

Let us find out how *the temperature time series of Alaska* look like.

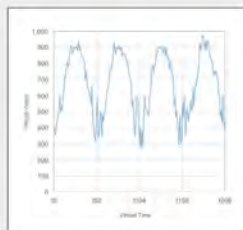


Source: <https://www.quora.com/What-is-it-like-to-visit-Alaska-during-the-winter>

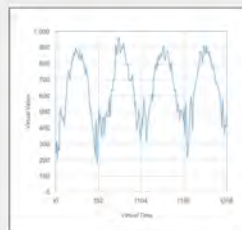
Next >

Figure A.12: Question 6

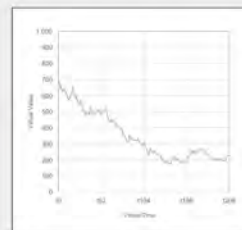
Which of the following plots represents a time series that has **a minimum value of 175** and likely reflects the information given in the previous page?



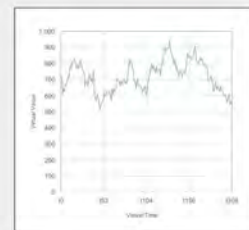
A



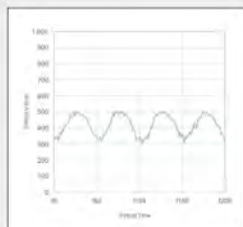
B



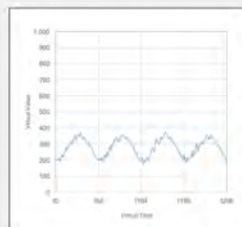
C



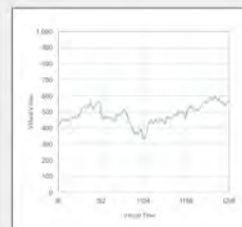
D



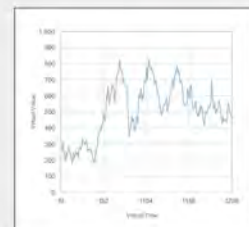
E



F



G



H

Next >

Figure A.13: Question 6



NOT Asystole

More commonly known as '**flat line**', **asystole** refers to a cardiac arrest rhythm or simply as an absence of the heartbeat.

The following are common causes of a flat line that is NOT asystole:

1. Loose or disconnected leads
2. Loss of power to the ECG monitor
3. Low signal gains on the ECG monitor



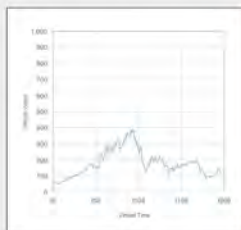
Which time series is an electrocardiogram (ECG) containing a **flat line** due to disconnected leads **at the beginning of the record**?

Source: <https://acls-algorithms.com/asystole/>

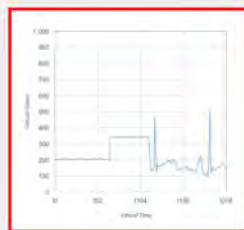
Next >

Figure A.14: Question 7

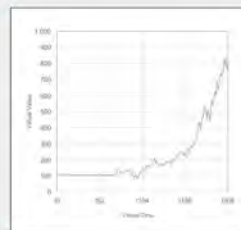
Which of the following plots represents a time series that has **a standard deviation of 80** and likely reflects the information given in the previous page?



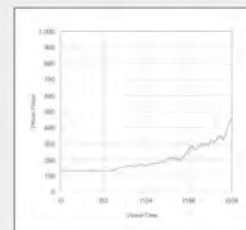
A



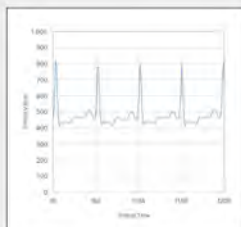
B



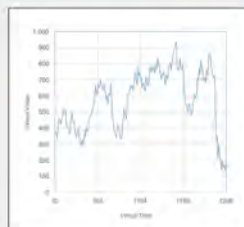
C



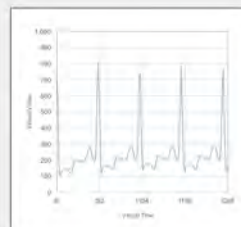
D



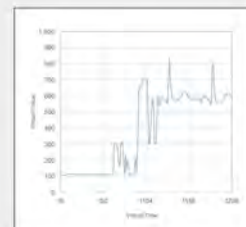
E



F



G



H

Next >

Figure A.15: Question 7

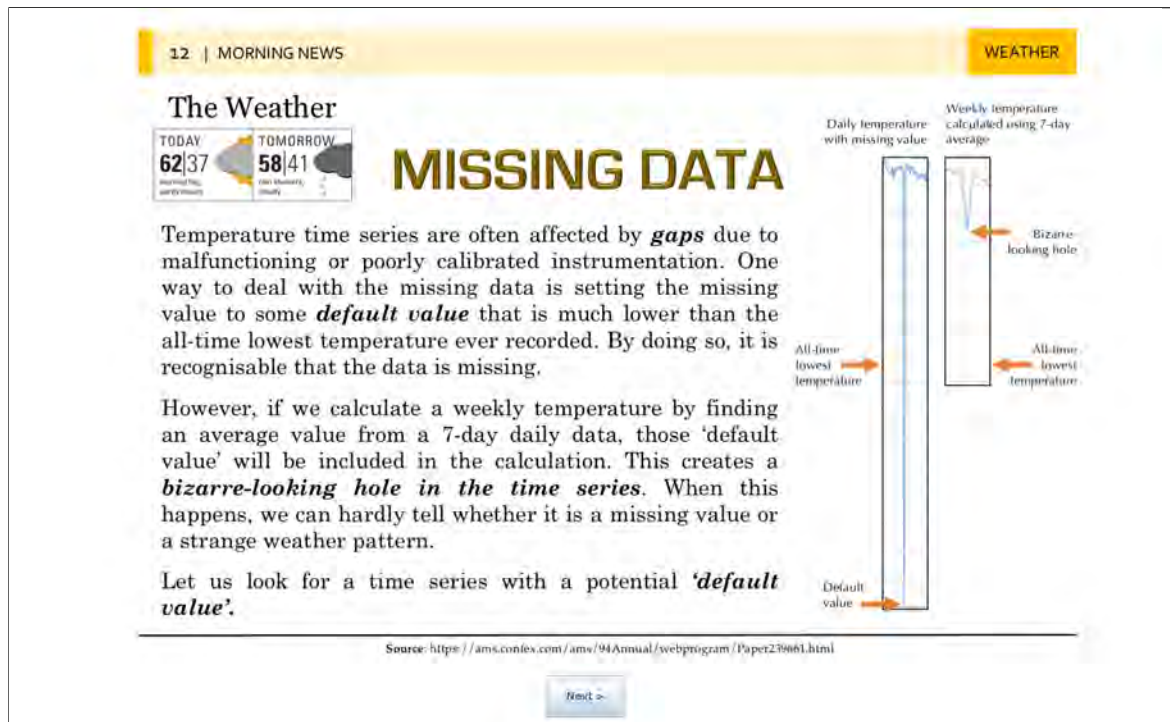


Figure A.16: Question 8

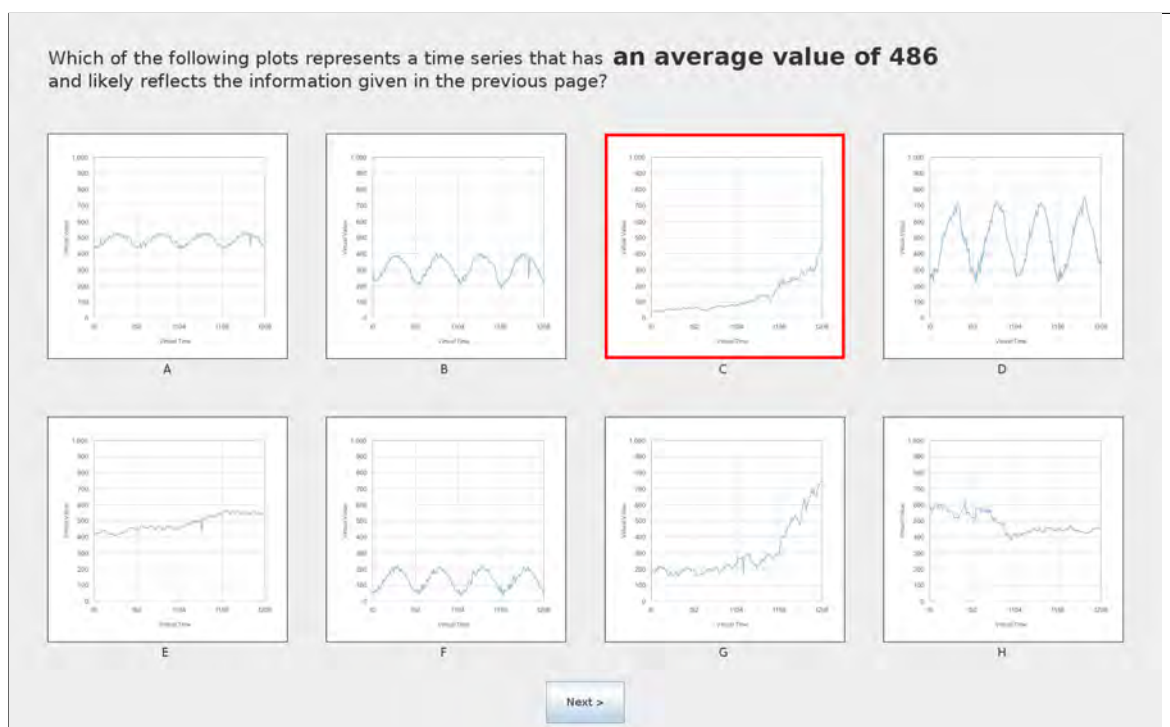


Figure A.17: Question 8

Morning News

VOL. 1, NO. 33
AUGUST 1, 2016

INVESTING IN CATS AND DOGS



Cats and dogs is a slang term referring to speculative stocks that have short or suspicious histories for sales, earnings, dividends, etc. The origin of this term may have stemmed from the use of "dog" to refer to an underperforming stock.

Investing in cats and dogs may create a large financial gain or loss in a short amount of time, which sometimes results in a **high volatility** in the stock price.

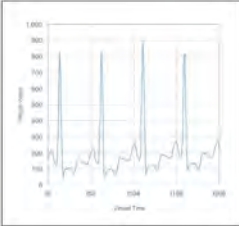
Which time series represent **a stock price with high volatility**?

Source: <http://www.investopedia.com/terms/c/catsanddogs.asp>

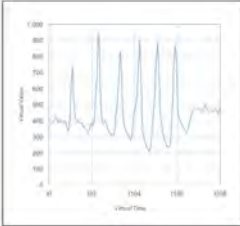
Next >

Figure A.18: Question 9

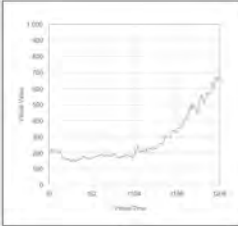
Which of the following plots represents a time series that has **a standard deviation of 148** and likely reflects the information given in the previous page?



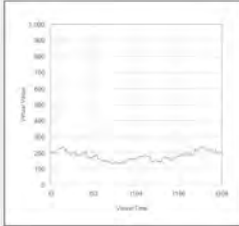
A



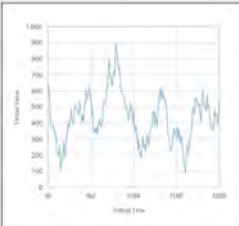
B



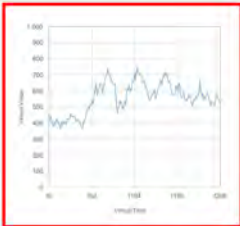
C



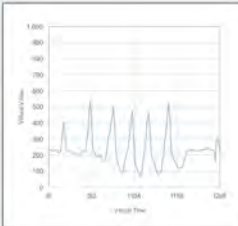
D



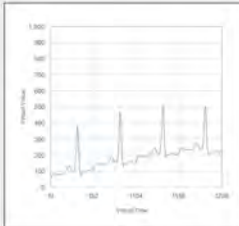
E



F



G



H

Next >

Figure A.19: Question 9

Wandering Baseline

An electrocardiogram (ECG) is a record of electronic impulses generated from the heart.

Sometimes, an ECG record shows a **wandering baseline**, which can be caused by several factors such as patient's motion, deep breathing, and loosely connected electrodes.

Can you spot a patient's **ECG record that contains a wandering baseline that is trending upwards?**

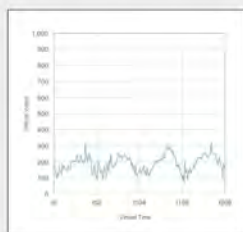


Source: http://www.nursingcenter.com/upload/static/592775/take5_monitor_problems.pdf

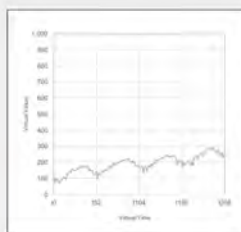
Next >

Figure A.20: Question 10

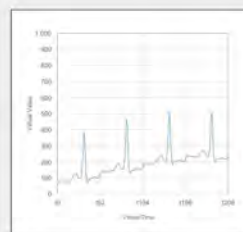
Which of the following plots represents a time series that has **an average value of 189** and likely reflects the information given in the previous page?



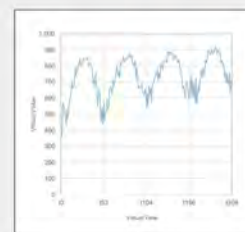
A



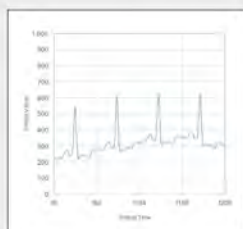
B



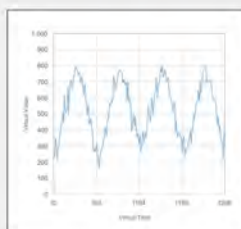
C



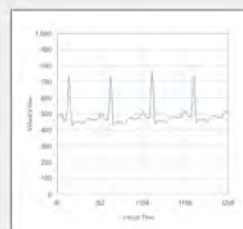
D



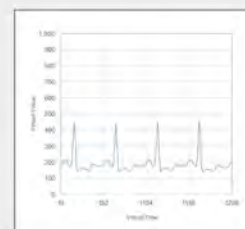
E



F



G



H

Next >

Figure A.21: Question 10

35 | MORNING NEWS

WEATHER

The Weather

TODAY

62|37

current temp, per by viewing

TOMORROW





58|41

low humidity, cloudy

MISSING A PERIOD OF DATA

Apart from machine errors and occasional missing temperature values, a temperature time series in some countries can have *long missing gaps*, which look like *a true straight line* at a certain period of the year. This may be the result of leaving a broken data-capturing machine unrepaired for a period of time.

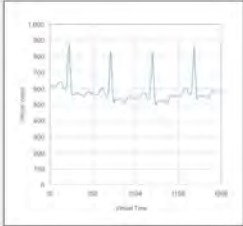
The countries with numerous gaps are mostly those in South Africa. These include Mozambique, Kenya, Congo, and Uganda. Can you tell which time series is a *temperature time series, which contains long missing gaps* ?

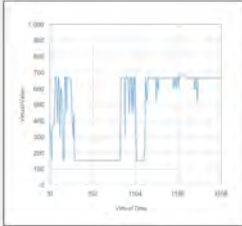
Next >

Figure A.22: Question 11

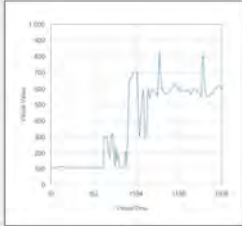
Which of the following plots represents a time series that has **a standard deviation of 131** and likely reflects the information given in the previous page?



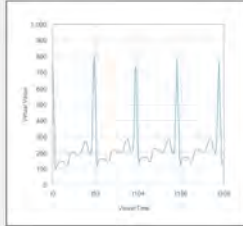
A



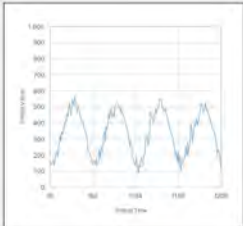
B



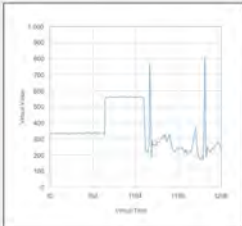
C



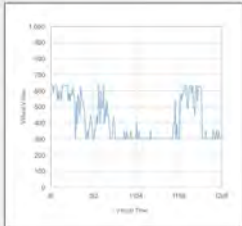
D



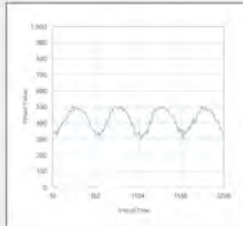
E



F



G



H

Next >

Figure A.23: Question 11

96

Morning News

VOL. 1, NO. 9
AUGUST 1, 2016

JANUARY EFFECT



The January effect is a seasonal **increase in stock prices during the month of January**. Analysts generally attribute this rally to an increase in buying, which **follows the drop in price that typically happens in December** when investors, engaging in tax-loss harvesting to offset realized capital gains, prompt a sell-off. Another possible explanation is that investors use year-end cash bonuses to purchase investments in the following month.

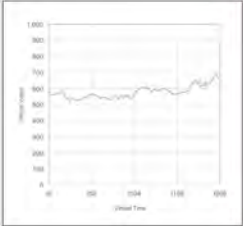
Which of the time series is a stock price whose **January effect is observable in the 4th year**?

Source: <http://www.investopedia.com/terms/j/januaryeffect.asp>

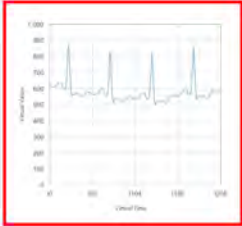
Next >

Figure A.24: Question 12

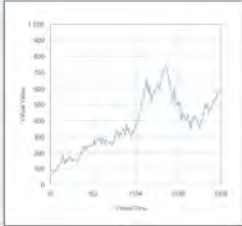
Which of the following plots represents a time series that has **an average value of 578** and likely reflects the information given in the previous page?



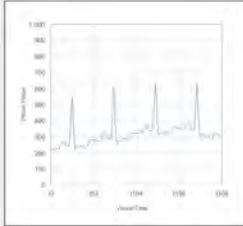
A



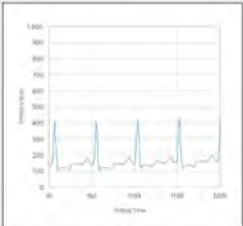
B



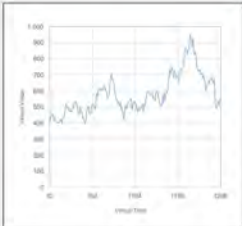
C



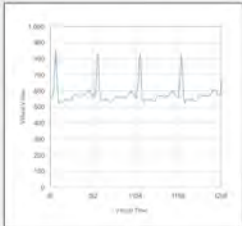
D



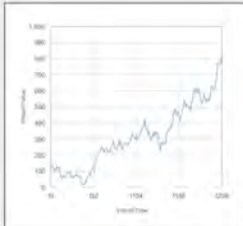
E



F



G



H

Next >

Figure A.25: Question 12



THE TRUE CAUSE OF PVCs IS NOT ALWAYS CLEAR. HOWEVER, PVCs MAY BE ASSOCIATED WITH ALCOHOL, TOBACCO, CAFFEINE, AND EXERCISE.

Premature Ventricular Contraction (PVC)

Premature ventricular contraction (PVC) is a type of ventricular ectopic beat, a beat that occurs at an abnormal position. PVC commonly occurs and is usually harmless in normal hearts.

Normally each beat in a normal electrocardiogram (ECG) has similar height and width. However, in an ECG containing a PVC beat, we can observe a **beat that is taller than other beats in the same ECG**.

Although PVC can occur in healthy, normal hearts, they are also sometimes a sign of underlying cardiac disease. Let us find an ECG time series whose **LAST beat is a PVC (where the last beat is taller)**.

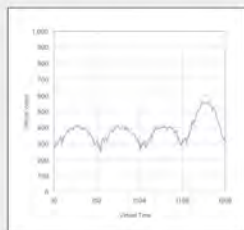
Source: <http://www.cardionetics.com/ventricular-ectopic-beats>

<http://www.healthcentral.com/encyclopedia/hc/premature-ventricular-contractions-3168414/#causes>

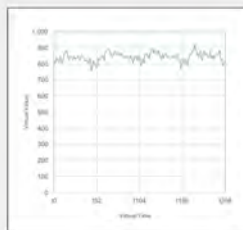
Next >

Figure A.26: Question 13

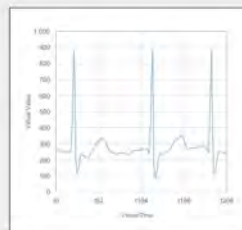
Which of the following plots represents a time series that has **a standard deviation of 112** and likely reflects the information given in the previous page?



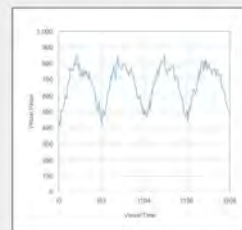
A



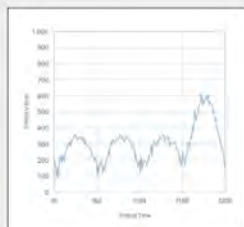
B



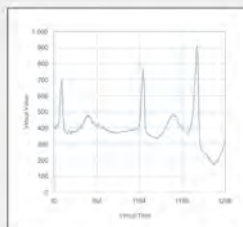
C



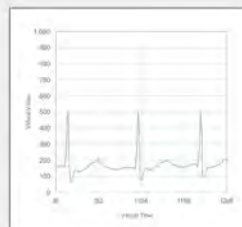
D



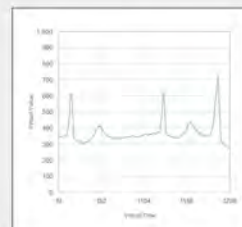
E



F



G



H

Next >

Figure A.27: Question 13

The Weather



MACHINE ERROR

Getting an accurate measurement of air temperature across the entire planet is not simple. Ideally, scientists would like to have thousands of standardized weather stations spaced evenly all around Earth's surface. The trouble is that there are some pretty big gaps over the oceans, the polar regions, and even parts of Africa and South America.

In the past, Africa was underdeveloped so their weather station tended to be more problematic than the other regions.

Can you distinguish which time series is Madagascar temperature containing *erratic temperature data between the end of the 3rd year and the beginning of the 4th year*?

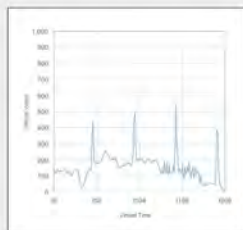


Source: <http://earthobservatory.nasa.gov/blogs/earthmatters/2015/01/21/why-so-many-global-temperature-records/> | <http://www.xocities.org/bhupinder Singh2/ddk/ssp/problems.htm>

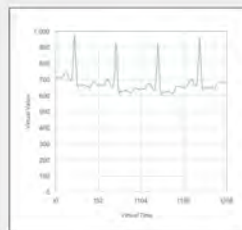
Next >

Figure A.28: Question 14

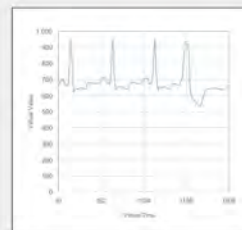
Which of the following plots represents a time series that has **an average value of 677** and likely reflects the information given in the previous page?



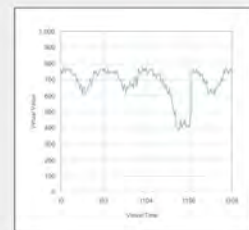
A



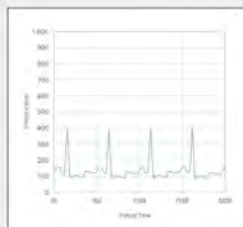
B



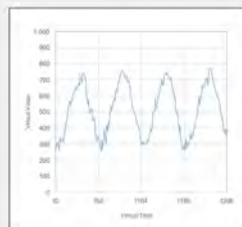
C



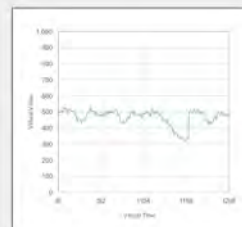
D



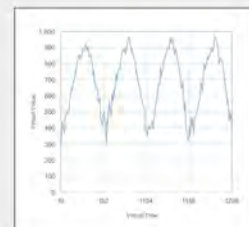
E



F



G



H

Next >

Figure A.29: Question 14

Morning News

VOL. 1, NO. 15
AUGUST 1, 2016

DINOSAUR ENERGY: A DEFENSIVE STOCK



A **defensive stock** is a stock that provides a constant dividend and stable earnings regardless of the state of the overall stock market. Dinosaur Energy stock price exhibits this characteristic.

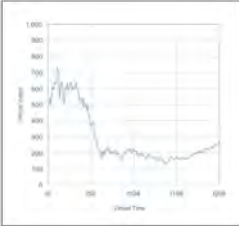
Dinosaur Energy stock price is relatively ***much less influenced by this year's economic crisis***. This is opposed to a typical situation when the price drops dramatically. In general, utility stocks are mostly defensive stocks because, during all phases of the business cycle, people still need gas and electricity.

Can you tell which time series is the **stock price of Dinosaur Energy (not influenced by economic crisis)**?

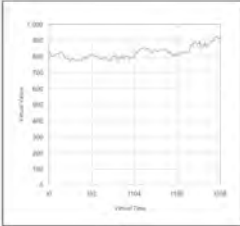
Source: <http://www.investopedia.com/terms/d/defensivestock.asp>

Figure A.30: Question 15

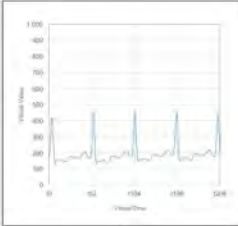
Which of the following plots represents a time series that has **a minimum value of 129** and likely reflects the information given in the previous page?



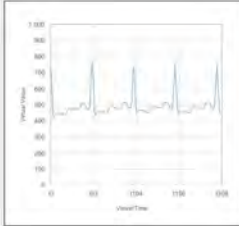
A



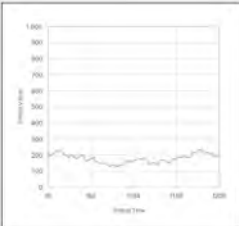
B



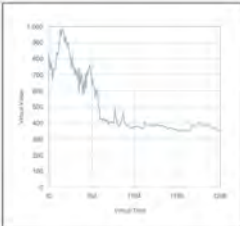
C



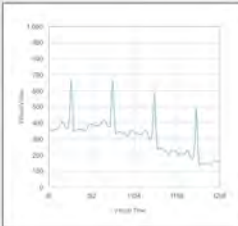
D



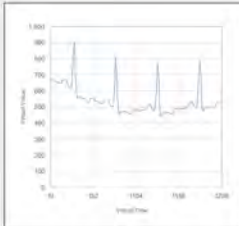
E



F



G



H

Figure A.31: Question 15

Appendix B

The Experiments



Figure B.1: Pre-study presentation.



Figure B.2: Experiment.

Bibliography

- [AB14] John R Anderson and Gordon H Bower. *Human associative memory*. Psychology press, 2014.
- [AJB16] Muhammad Adnan, Mike Just, and Lynne Baillie. Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5444–5455. ACM, 2016.
- [AMM⁺07] Wolfgang Aigner, Silvia Miksch, Wolfgang Müller, Heidrun Schumann, and Christian Tominski. Visualizing time-oriented dataa systematic view. *Computers & Graphics*, 31(3):401–409, 2007.
- [AMST11] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.
- [BB33] Frederic Charles Bartlett and Cyril Burt. Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, 3(2):187–192, 1933.
- [Bri01] David R Brillinger. *Time series: data analysis and theory*, volume 36. Siam, 2001.
- [CFB14] Min Chen, Luciano Floridi, and Rita Borgo. What is visualization really for? In *The Philosophy of Information Quality*, pages 75–93. Springer, 2014.
- [CFV⁺16] Min Chen, Miquel Feixas, Ivan Viola, Anton Bardera, Han-Wei Shen, and Mateu Sbert. *Information Theory Tools for Visualization*. A K Peters Ltd, 2016.

- [CG16] Min Chen and Amos Golan. What may visualization processes optimize? 2016.
- [CJ10] Min Chen and Heike Jaenicke. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, 2010.
- [CJK09] Charles Chang, Jing Jiang, and Kenneth A Kim. A test of the representativeness bias effect on stock prices: A study of super bowl commercial likeability. *Economics Letters*, 103(1):49–51, 2009.
- [CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [Few06] Stephen Few. Visual pattern recognition: Meaningful patterns in quantitative business information. 2006.
- [GAG⁺00] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. <https://www.physionet.org/>.
- [Gir92] Ellen R Girden. *ANOVA: Repeated measures*. Number 84. Sage, 1992.
- [Goo09] C James Goodwin. *Research in psychology: Methods and design*. John Wiley & Sons, 2009.
- [HA14] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.
- [HB89] Glyn W Humphreys and Vicki Bruce. Visual cognition, 1989.
- [JWC⁺11] Heike Jänicke, Thomas Weidner, David Chung, Robert S Laramée, Peter Townsend, and Min Chen. Visual reconstructability as a quality metric for flow visualization. In *Computer Graphics Forum*, volume 30, pages 781–790. Wiley Online Library, 2011.
- [KC⁺14] Evzen Kocenda, Alexandr Cerný, et al. Elements of time series econometrics. *University of Chicago Press Economics Books*, 2014.

- [Knu] Donald Knuth. University of Dayton - Environmental Protection Agency Average Daily Temperature Archive. <http://academic.udayton.edu/kissock/http/Weather/default.htm>.
- [Kot15] Vivek Kothari. Visual Analytics Methods for Image Classification. Master's thesis, University of Oxford, 2015.
- [Ler11] Gondy Leroy. *Designing User Studies in Informatics*. Springer Science & Business Media, 2011.
- [Lev14] Daniel J Levitin. *The organized mind: Thinking straight in the age of information overload*. Penguin, 2014.
- [LN13] Peter H Lindsay and Donald A Norman. *Human information processing: An introduction to psychology*. Academic Press, 2013.
- [LS96] Nikos K Logothetis and David L Sheinberg. Visual object recognition. *Annual review of neuroscience*, 19(1):577–621, 1996.
- [MGSP08] Michael L Mack, Isabel Gauthier, Javid Sadr, and Thomas J Palmeri. Object detection and basic-level categorization: Sometimes you know it is there before you know what it is. *Psychonomic Bulletin & Review*, 15(1):28–35, 2008.
- [MJK15] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.
- [Nei67] Ulric Neisser. *Cognitive psychology*. 1967.
- [Ric06] John Rice. *Mathematical statistics and data analysis*. Nelson Education, 2006.
- [Rut01] Andrew Rutherford. *Introducing ANOVA and ANCOVA: a GLM approach*. Sage, 2001.
- [Sel15] Howard J Seltman. *Experimental design and analysis*. 2015.
- [Sha01] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

- [TC05] James J Thomas and Kristin A Cook. National v, analytics c. illuminating the path. los alamos. *CA: IEEE Computer Society*, 2005.
- [Yah] Yahoo! Finance. <https://finance.yahoo.com>.