MAXWELL RUPANTE SANE
SCM222-0558/2025

SUMMARY NOTES - IT FOR STATISTICS

INTRODUCTION

Information Technology – refers to the use f computers, software networks and other technologies to manage, process, store and communicate information.

TOPIC 1: ROLE OF INFORMATION TECHNOLOGY AND COMPUTER IN TODAY'S WORLD

IT is about using computers and software to manage, process, and communicate information.

1. Medical research – IT tools to facilitate medical research, analysis and discovery
2. Education: E-learning platforms and digital resources enhance information making it accessible and flexible.
3. Agriculture: IT tools are used to deduce weather patterns to predict when to plant and to harvest.
4. Environmental sustainability: IT can help reduce carbon footprint through establishment of remote work which reduces carbon emissions by reducing travel and fuel consumption
5. Communication & Collaboration: Cloud platforms (like Google Workspace or Microsoft Teams) and APIs (Application Programming Interfaces) allow for seamless sharing of models and data between quants, traders, and risk managers across the globe.

- Data Processing & Analysis: We use computers to clean, sort, and analyze vast datasets like historical stock prices, economic indicators, or real-time trade feeds to identify patterns and build predictive models.

E.g.: Running a Monte Carlo simulation to forecast the potential future price of a complex derivative. Doing this by hand is impossible; with a computer, it takes seconds.

- Storage & Retrieval: IT systems provide the "memory" for our financial world, storing everything from client portfolios to decades of market data in databases that can be queried instantly.

- Automation: Computers execute repetitive tasks flawlessly and at high speed.

Instance: High-Frequency Trading (HFT) algorithms that make thousands of trades per second based on predefined criteria; a process entirely driven by IT systems.

TOPIC 2. FUNDAMENTALS OF COMPUTER OPERATIONS

The core operation of the computer is the Input-Process-Output (IPO) Cycle, guided by stored instructions.

Input: It is the keying in of commands and instructions to a computer by using some of the computer software such as the mouse, and keyboard

Processing: The Central Processing Unit (CPU) fetches instructions from the software (e.g., Python, R) and performs calculations on the data.

Output: The computer presents the results of the processing.

The result, the computer outputs what the commands were made and keyed in via the keyboard.

Storage: Crucially, results or intermediate data can be stored for later use (e.g., saving the result to a database or your hard disk).

The Fetch-Decode-Execute Cycle: This is how the CPU does its job. It continuously:

Fetches the next instruction from memory,

Decodes it to understand what to do (e.g., "add these two numbers"),

Executes the command. This happens billions of times per second.

TOPIC 3. COMPUTER HARDWARE AND SOFTWARE

a) Hardware – This is the tangible part of the computer

I. CPU (Central Processing Unit): The brain. For statistical work, parallel processing running multiple calculations simultaneously, which speeds up tasks like bootstrapping or large-scale simulations.

II. RAM (Random Access Memory): The computer's short-term memory. It's super-fast but volatile i.e., clears when powered off. When you load a dataset into R or Python, it lives in RAM. The more RAM you have, the larger the datasets you can work with without slowing down. 16GB is a modern minimum; 32GB+ is better for large financial models.

III. Storage (Hard Disk Drive (HDD) and Solid-State Drive (SSD)): The long-term memory. This is where your Operating System, software, and data files permanently reside. SSDs are much faster than HDDs, leading to quicker boot times, faster software loading, and speedier file access.

IV. Motherboard: The nervous system that connects all the components.

V. Input and Output Devices: These key in commands to the computer. Examples are: Keyboard, mouse, monitor.

b) Software These are computer parts that are intangible thus can't be seen, or touched

System Software:

Operating System (OS): Windows, macOS, Linux. It's the intermediary between you and the hardware. It manages resources, runs applications, and provides a user interface. Linux is heavily used on servers for its stability and performance.

Application Software:

General-Purpose: Excel (ubiquitous in finance for quick analysis and prototyping).

Specialized Statistical Software: R, Python (with libraries like Pandas, NumPy, SciPy), MATLAB, SAS. These are our primary tools for statistical modelling and data analysis. Database Software: SQL-based systems (MySQL, PostgreSQL) for managing and querying large, structured datasets.

UNIT 4: COMPUTER SOFTWARE BASICS

4.1 Operating System Functions: A Deeper Look

The OS acts as a resource manager, abstracting the complexity of hardware away from the application programs. Kernel: The central component of the OS, managing the interaction between hardware and software. Virtual Memory: A technique where the OS uses hard disk space (called a swap file or paging file) to simulate additional RAM when physical RAM is full. This allows the computer to run more programs than its physical RAM allows, albeit slowly. Multitasking: The OS allows multiple processes to run concurrently by rapidly switching the CPU between tasks (time-slicing), giving the appearance of simultaneous execution.

4.2 Software Licensing and Acquisition

Software is typically governed by licensing models. Proprietary Software: Owned by a company or individual; users must purchase a license to use it (e.g., Microsoft Office). Open-Source Software (OSS): Source code is publicly available and modifiable. Encourages community development and collaboration (e.g., Linux, Python). Software as a Service (SaaS): Software is hosted by a vendor and accessed over the Internet (Cloud Computing model), eliminating the need for local installation (e.g., Google Workspace, Salesforce).

4.3 Translating Software: Compilers vs. Interpreters

High-level programming code must be translated into machine code (binary). Compiler: Translates the entire source code into an executable machine code file before execution. The resulting program runs very fast (e.g., C, C++). Interpreter: Translates and executes the source code line by line during execution. Slower execution speed, but development and debugging are often faster (e.g., Python, JavaScript).

TOPIC 5. CONSTRUCTING DATA FILES

A data file is a structured collection of information. How we structure it is critical for both humans and software to read it correctly.

Common Data File Formats

CSV (Comma-Separated Values): The most common format for raw data. It's a plain text file where each line is a row, and values are separated by commas.

Examples:

Date, Open, High, Low, Close, Volume

2023-10-01,150.25,152.80,149.50,151.75,1567800

2023-10-02,151.80,154.20,151.10,153.90,1985200

Excel (.xlsx): Good for manual work and formatting, but can be more complex for programs to read directly.

Plain Text (.txt): For unstructured

Methods of File Organization

Serial (sequential) files: They store records one after another just as they are created

Pro- simple and fast to create

Con- slow to access

Random

It is a storage method where records are placed at random locations on the storage medium, usually determined by a hash function based on a key field.

TOPIC 6. DISK STORAGE

How data is stored physically affects speed, cost, and capacity.

Magnetic Storage (HDD - Hard Disk Drive):

How it works: Uses spinning magnetic platters and a read/write head.

Pros: Cheap, high capacity (good for archiving old trade data).

Cons: Slower, mechanical parts can fail.

Solid-State Storage (SSD - Solid State Drive):

How it works Uses flash memory chips (like a large USB drive). No moving parts.

Pros: Very fast (leads to much quicker system responsiveness and data loading), more durable, less power consumption.

Cons: More expensive per GB.

Verdict: Essential for your primary work machine. The speed boost for loading software and datasets is invaluable.

Cloud Storage:

How it works: Data is stored on remote servers accessed via the internet (e.g., AWS S3, Google Cloud Storage, Microsoft Azure).

Pros: Scalable (pay for what you use), accessible from anywhere, built-in redundancy and backup.

Cons: Ongoing cost, dependent on internet connection.

Instance: A hedge fund might store petabytes of historical tick data on AWS S3 and use cloud computing (EC2) to run analysis on it.