INFORMATION TECHNOLOGY FOR STATISTICS

1) Introduction to IT & computers

a) Definition of information technology:

The definition of IT has evolved since the 1960's and 1970's as initially it was a technical term used within hospitals, banks and research institutions to describe early mechanical and electronic processes for storing and retrieving data.

In the 1980's and 1990's, the rise of personal computers, relational databases and networked environments transforms workspaces into "paperless" ecosystems. The expansion in computational capacity subsequently increased the demands on Information Technology (IT) to support more complex tasks.

In the business & industry perspective, IT is defined as: The study, design, development, application, implementation, support or management of computer-based information systems. Therefore, in finance, IT is the foundation of market data pipelines, algorithmic trading infrastructure, risk aggregation platforms, and portfolio environments.

In the academic and professional perspective IT is defined as the discipline responsible for:

1) Selecting appropriate hardware and software for an organization
2) Integrating systems into existing analytic and operational workflows
3) Maintaining, upgrading and securing the computational infrastructure

b) Role and components of IT in modern analytical professions

1) Data storage (clinical, genomic, epidemiological, market, transactional)
2) Data processing using appropriate computational models
3) Retrieve information rapidly for decision-making
4) Data transmission securely across networks
5) Protective sensitive or high value information

2) Fundamentals of computer operations

A computer is an integrated system whose components collaborate to execute complex analytical workflows.

The operational cycle of a computer is: Input Processing Output Storage

1) Input unit

The input unit acts as the system's sensory layer. Its role is to transform heterogenous real-world signals into coherent digital structures that the processor and memory hierarchy can manipulate

Human Interface Devices (HID)

- **Keyboard**

Converts keystrokes into ASCII/Unicode codes. In practice, this becomes part of command-driven bioinformatics environments or text-heavy trading terminals that require precise, low-latency input.

- **Mouse, Trackball, Joystick**

Generate spatial and kinematic data streams. Such devices are key in navigating three dimensional protein models or interacting with layered medical imaging datasets.

Automated Instruments

- **Scanners and Sensors**

  Sequencing machines, flow cytometers, and radiology devices produce high-volume continuous streams. Input controllers must buffer, encode, and checkpoint this data before passing it to memory.

- **Market Data Feeds**

  Tick-level pricing, news sentiment streams, and order-book snapshots arrive via high speed network cards. Failure to ingest a single tick can alter price -path simulations and arbitrage strategies.

Key Architectural Considerations

- **Latency sensitivity**

  Lossless ingestion of high-velocity is non-negotiable in high-frequency trading or real-time monitoring in ICUs.

- **Error handling**

Parity bits, CRC checks, and other integrity checks ensure that upstream noise does not propagate into statistical or financial inference engines.

Domain insight

In epidemiological modeling, input units must integrate real-time patient sensors with batch-upload records. Correct timestamping ensures that analytical models do not misinterpret asynchronous updates as causal associations.

2)Processing

Central Processing Unit (CPU)

The CPU is the "brain" of the computer. It executes instructions organized into programs (software) that determine the computer's actions.

Modern CPUs integrate two principal functional blocks:

- **Arithmetic and Logic Unit (ALU)**
- **Control Unit (CU)**

- **Arithmetic and Logic Unit (ALU)- Executes arithmetic and logical computations at high speed.**

  Functions of the ALU are:

  - Arithmetic operations: addition, subtraction, multiplication, division.
  - Logical operations: AND, OR, NOT, comparisons, bitwise manipulations.

Operational Flow:

1. Operand Fetching: Data retrieved from primary memory or high-speed CPU registers.
2. Computation: ALU executes the operations.
3. Result Storage: Outputs are written back to memory or held in registers i.e. subsequent operations.

Application to Financial Engineering

Rapid communication of floating-point numbers in option pricing or portfolio risk assessment.

- **Control Unit (CU)- Orchestrates execution, ensuring order, timing and synchronization**

  The CU coordinates all CPU operations, acting as the system's nervous systems

  Function of the CU

• Instruction decoding
• Signal generation
• Synchronization

Application to financial engineering

Ensures market feeds, analytical computations, and execution engines operate synchronously, preventing latency-induced arbitrage errors.

3 Internal organization of the CPU

Modern CPUs incorporate specialized registers and buffers to accelerate execution. Registers provide ultra-low-latency storage for intermediate computations • **Instruction Register (IR)** – Stores the current instruction; feeds control circuits for timing and sequencing

• **Program Counter (PC)-** Holds the address of the next instruction; increments sequentially or during conditional jumps.

• **Other internal Registers**

• **General Purpose Registers:** Hold operands and temporary results. • **Special Purpose Registers:** Include stack pointers, index registers, and flags. • **Cache Registers:** Store frequently accessed data and instructions to reduce memory latency.

3)Output Unit: Rendering Data into Actionable Knowledge

The output subsystem converts internal digital structures into human-usable or machine usable forms.

Visual Output

High-resolution displays are essential for genomic visualizations, clinical dashboards, or real-time financial risk surfaces.

Hard Copy Output

Printers or report generators remain vital when regulatory compliance demands immutable physical documentation—such as clinical trial reports or financial audit logs.

Dual-Function Terminals

Touchscreen terminals in laboratories or trading floors enable direct manipulation of analytical outputs, creating feedback loops that blend machine inference with human judgment.

Advanced Considerations

- **PrecisionvsLatency**

  A genomics workstation may prioritize high-resolution plotting, while a trading dashboard sacrifices precision to maintain real-time updates.

- **Device-Specific Encoding** Output units may need to rasterize vector graphics or convert data streams into formats such as PostScript before rendering.

Domain Insight

In algorithmic trading, outputs are often consumed directly by automated execution engines rather than humans—illustrating that "output" may simply be another machine readable layer in a pipeline.

4)Storage /Memory unit

Computational performance and scalability depend on how the money hierarchy orchestrates speed, locality, and persistence.

Primary Memory (RAM)

- **RegistersvsRAM**

Registers offer sub-nanosecond access; RAM operates in the tens-of-nanoseconds range. Optimized algorithms exploit register locality to reduce instruction cycles. • **DataLocality**

Cache-aware and cache-friendly algorithms minimize memory stalls, crucial for real-time financial modeling and large-scale biostatistical simulations.

- **WordLengthAlignment**

The choice between 32-bit and 64-bit words affects precision and storage. For Monte Carlo option pricing or genomic imputation models, floating-point precision becomes a statistical determinant.

Domain Insight

Dynamic programming algorithms like Smith–Waterman require large in-memory matrices. Limited RAM forces disk-backed operations, drastically affecting runtime.

Secondary Memory

Long-term storage systems—HDDs, SSDs, tapes—support persistence but operate with significantly higher latency.

Historical Data Set storage

Terabyte-scale clinical archives or multi-decade financial price histories reside

here. **I/O Bottlenecks**
Algorithms handling large files must accommodate slow bandwidth and unpredictable seek times.

Advanced Techniques

- **Memory-Mapped Files** Provide a practical compromise when datasets exceed available RAM.

Tiered Storage Systems
Allocate SSDs for active models while reserving HDDs for deep archives.

Domain Insight

In high-frequency trading, primary and cache memory dominate active computation. Secondary storage serves mostly archival functions, whereas in biostatistics, it underpins reproducibility and multi-cohort meta-analysis.

Summary:

Component Purpose	**Financial Engineering**
CPU Executes instructions	Option pricing, risk simulations
RAM Temporary storage	Covariance matrices, scenario analysis
Relevance	**Component Purpose**
	Permanent data
	Financial Engineering Relevance

Historical market data, trade

Mass Storage
retention
logs

Market inputs, simulation

Input Devices Data entry

parameters

Low-latency financial

Output Devices

Data visualization Dashboards, risk metrics

computation

Bus Data transfer

3)Computer hardware basics

Hardware includes all the physical parts of the computer

system. **Essential hardware components**

- Central Processing Unit (CPU)
- Memory
  - RAM – Random Access Memory
  - ROM- Read Only Memory
- Storage devices i.e.
  - External storage devices- USB flash drives & SD cards

Internal storage devices – HDDs and SSDs

- Input devices-

  Keyboard

  Mouse

- Output devices

  Monitor

  Printer

4)Computer Software Basics

Computer Software refers to the programs of application and instructions that tell a computer what to do.

Categories of software

a. System software

It helps run control and manage computer hardware.

Forms the foundation on which application software operates.

Types of system software

1) Operating systems (OS) – Manages all hardware & software operations ie Windows, MacOS
2) Utility Programs- Help maintain, protect and optimize the computer. i.e. Antivirus, Backup tools.
3) Device Drivers- Small programs that allow hardware devices to communicate with the computer

b. Application Software

These are programs designed to perform specific tasks for users. Types of

Application Software

1) Productivity Software – Used for office work & daily tasks I,e Microsoft word 2) Web

browsers – Used to access the internet i.e. Chrome

3) Graphics and multimedia software - Used for creating and editing images videos and audios
4) Communication software - Enables messaging, video calls and collaboration i.e Whatsapp

5) Database software- Used to create, manage and store large amounts of data Microsoft

Importance of statistics

Controls and operates computer hardware

Helps uses perform tasks

Enables communication and internet access

Improves productivity and efficiency

Supports learning, business and entertainment

Difference between Hardware and Software

Hardware Software

Physical and tangible	Non- physical, intangible
Wears out overtime	Does not wear out but can become outdated
Hard to modify	Easy to modify or update
Executes Software	Direct hardware to perform tasks

4) Data and data files

Data – Data refers to raw facts, figures, instructions that have not much meaning to the user. **Types of data**

a) Numerical data – Whole numbers (integers) and decimals

b) Text (Alphanumerical) data – Letters numbers and symbols

c) Image data – Pictures Graphics and formats

d) Audio data – Sound recordings Formats

e) Video data – Moving pictures

f) Boolean data – True or false values

Data files

A data file is a collection of related data stored under a single name on a storage device

It allows data to be; stored, organized, retrieved, updated and shared

Types of data files

a) Text files – Contain plain readable text Used in documents
b) Binary Files - Contain data in machine readable text. Used in Programs, system files c) Database files – Stores structured data in tables. Used in Record management systems d) Spreadsheet files- Used for numerical analysis and calculation. Used in Data analysis; Budgets
e) Multimedia files – Contains images, audio and videos

Files Organization Methods

This explains how records are arranged inside a file . There are two types:

a) Sequential Datafiles

A sequential file stores records one after another in a fixed logical order I.e Alphabetical order
Characteristics

1) Records are arranged sequentially
2) Access is from the beginning to the end
3) Efficient for batch processing
4) Slow for random searches

Advantages

1) Easy to create
2) Efficient for large volume batch processing
3) Simple to update when order is preserved

Disadvantages

1) Slow searching (must read from start)
2) Updating requires rewriting the entire file
3) Not suitable for systems needing instant access

Constructing a sequential Data file

a) Define the records (fields like; name, ID, marks)
b) Choose a key field (Admission number)
c) Arrange data in order (001)
d) Store records consecutively without gaps

Example

10	Name	Score
001	Alex	74
002	Brian	68
003	Chris	97

b) Random (Direct) Data Files

A random file stores data anywhere in the storage location using a formula (hash function) to compute where each recording should go.

Characteristics

1) Data stored at arbitrary locations
2)Access uses a key not physical order
3)Very fast random access
4)Used in real time systems

Advantages

1) Fast searching and updating
2) Suitable for real time transactions

3) No need to read the entire file

Disadvantages

1) Complex to design
2) Collision (when two keys map to same location)
3) Requires more memory

Constructing a random file

1) Choose a key field
2) Apply a hash function to determine the storage location
3) Store the record at the computed address
4) If collision occurs use ; -open addressing
   -overflow area
   - chaining

Differences between sequential and random files

Feature	Sequential file	Random file
Access method	Start to end	Direct using key
Speed	slow	Very fast
Use case	Batch processing	Real time suggestions
Storage order	Fixed	Unpredictable
Updating	Hard	Easy

Bits and Bytes

1) Bits

A bit (binary digit) is the smallest unit of data in a computer
Can only take two values: 0 – off; 1- on
Computers store and process all data in binary form.

2) Bytes

It is the standard unit used to represent a single
character A byte = 8 bits

Memory units

1 Byte(B) = 3 bits

1 Kilobyte (KB)= 1024 bytes

1 Megabyte (MB)= 1024 KB

1 Gigabytes (GB)= 1024 MB

1 Terabyte (TB) = 1024 GB

6)Disk Storage Fundamentals

Disk storage is a type of permanent storage that uses magnetic, optical or solid-state technologies to store data

Examples are; Hard Disk Drive (HDD), Solid State Drive (SSD) Optical Disk (OD) **Types of disk storage**

a) Magnetic storage

Uses magnetism to store data

Features

1) Large capacity

2) Slower than SSD

3) Uses moving parts (spinning platters read / write heads)

Examples

1) Hard Disk Drives (HDD)

2) Magnetic topes

3) Floppy disks

b) Solid state storage

Uses flash memory (no moving parts)

Features

Fast read or Write speeds

More durable

Silent operations

Higher cost per GB compared to HDD

Examples

SSDs

Flash drives

c) Optical storage

Uses lower technology to read and write data.

Examples

CDs

DVDs

Features

Portable

Good for media storage

Slower and lower capacity compared to HDD

Structures of disk storage

a) Platters (HDD)

Disk inside the HDD coated with magnetic material

Spins at high speed (5400-7200+RPM)

Data is stored on surfaces of platters

b) Tracks- Concentric circles on disks where data is written

c) Sectors- Small physical unit storage unit on a track (typical size is 512 bytes) d) Clusters- Groups of sectors used as the smallest unit for file storage e) Cylinders- Formed by stacking tracks located in the same position on multiple platters

f) Read or Write heads- Tiny electromagnetic heads that float just above the platter surface

_ Read data from magnetic surfaces

- Write data onto platters

DEBRA WAMBUI MWANGI

SCM222-1285/2025