Dependable Distributed Systems Master of Science in Engineering in Computer Science

AA 2023/2024

LECTURE 22B – MODELING THE WORKLOAD OF A SYSTEM

Part of the material inside is taken from slides of profs Menascé, Almeida, and Ciciani

Performance and Workload

The performance of a system depends heavily on the characteristics of its load

More in detail, the **performance** provided by a system are **mostly influenced by**:

- The **arrival pattern**
- The **service demands**

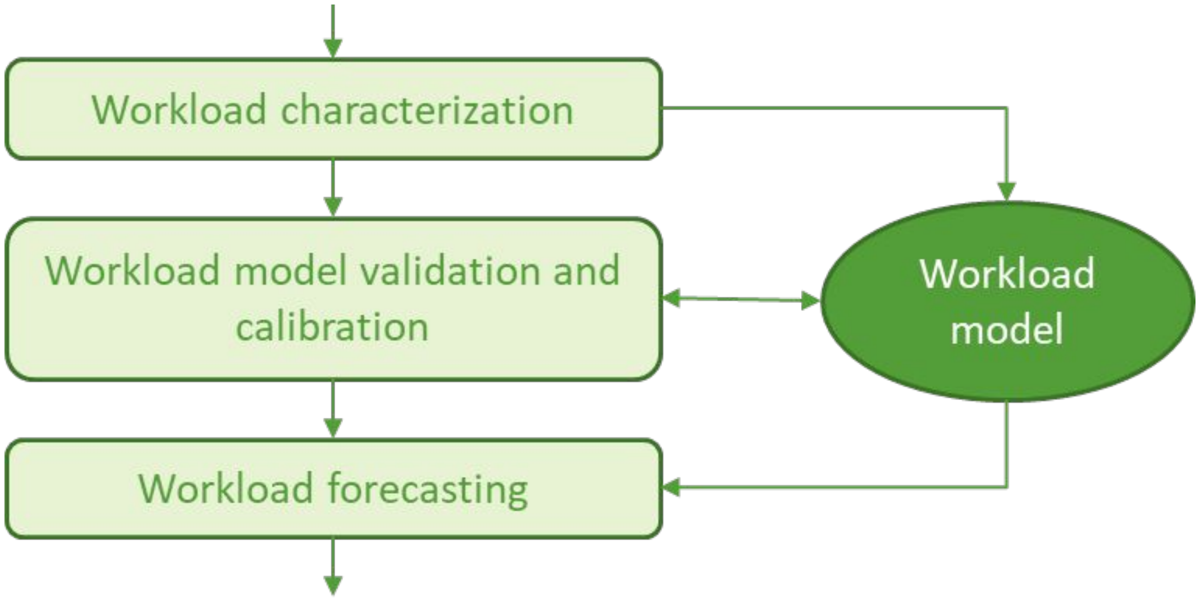
→ 1) Characterize the workload

Workload

The workload of a system is the **set of all inputs that the system receives from its environment during a given period of time**

Since a **real-user environment is generally not repeatable**, it is necessary to study the real-user environments, observe the key characteristics, and **develop a workload model** that can be used repeatedly

Building a Workload Model



A model is an abstraction of a generalized overview **of a real system**

The level of detail of the model **and the specific aspects** of the real system that are considered in the model **depends on the purpose of the model**

→ should not be more complex than is necessary to achieve its goal

Workload Model Construction: Common Steps

1. **Specification** of a **point of view** from which the workload will be analyzed (inside or outside the system?)

2. **Choice** of the **set of parameters** that capture the most relevant characteristics of the workload for the purpose of the study

3. **Monitoring** the system to obtain the raw performance data

4. **Analysis** and reduction of the performance **data**

5. **Construction** of a workload **model**

6. **Validation** that the characterization captures all the important performance information

Workload Model

Workload models are exploited as **input of analytic and simulation system models** and for the definition of benchmarking experiments

When building workload models, particular attention has to be paid to their **accuracy and representativeness (model validation and calibration)**, that is, their ability to **capture and reproduce the most relevant characteristics of the workloads** and how users behave

Workload Characterization

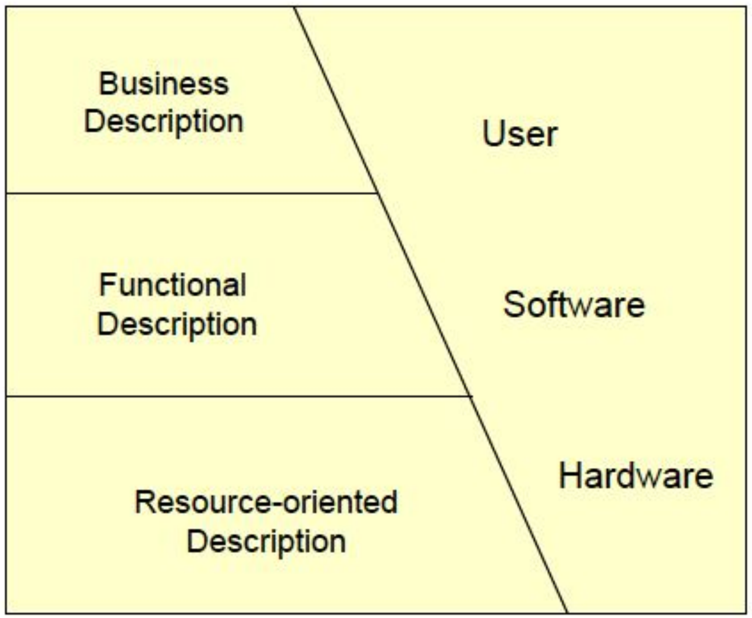
Workload characterization relies on experimental approaches based on the analysis of **measurements collected on the technological infrastructures while they are operating** (i.e., under their real workloads)

The ***workload characteristics*** are represented by a **set of information** (e.g. arrival, request parameters, etc.) collected for **each request**

The choice of the **characteristics** and parameters that will describe the workload **depends on the purpose** of the study

Workload Description

The workload of a computer system can be described at different levels:



- **Business characterization**: a user-oriented description that describes the load in terms such as number of employees, invoices per customer, etc.

- **Functional characterization**: describes programs, commands and requests that make up the workload

- **Resource-oriented characterization**: describes the consumption of system resources by the workload, such as processor time, disk operations, memory, etc.

Service Demands

Each request (arrival) is handled by one or more system resources (stations)

It has certain **demands**: access to a database, call a function, CPU time, memory usage etc.

Depending on the kind of system and on the purpose of the study different service demands can be considered

It must be understood **which** system components are interested by a request and **how**

e.g. a web service call may require an access to a database, a remote connection to another server etc.

Workload Components

The requests arriving to a system could be heterogeneous

Identify the workload components, namely the **different** (and relevant) **units of work** that arrive at the system from external sources (e.g. read requests, transactions, etc.)

Workload component requests share the same system demands

Each workload components must be characterized

Each workload component is **characterized by a set of parameters**

The kind of parameters strongly **depends on the kind of service**

Specific Workload Parameter Examples [2]

Web Workloads:

- page properties,
- traffic properties,
- access patterns,
- user behaviour.

Shopping Service Workloads:

- business level,
- session level,
- function level,
- protocol level.

Online Social Network Workloads:

- user behaviour,
- network structure and evolution,
- content propagation.

Video Service Workloads:

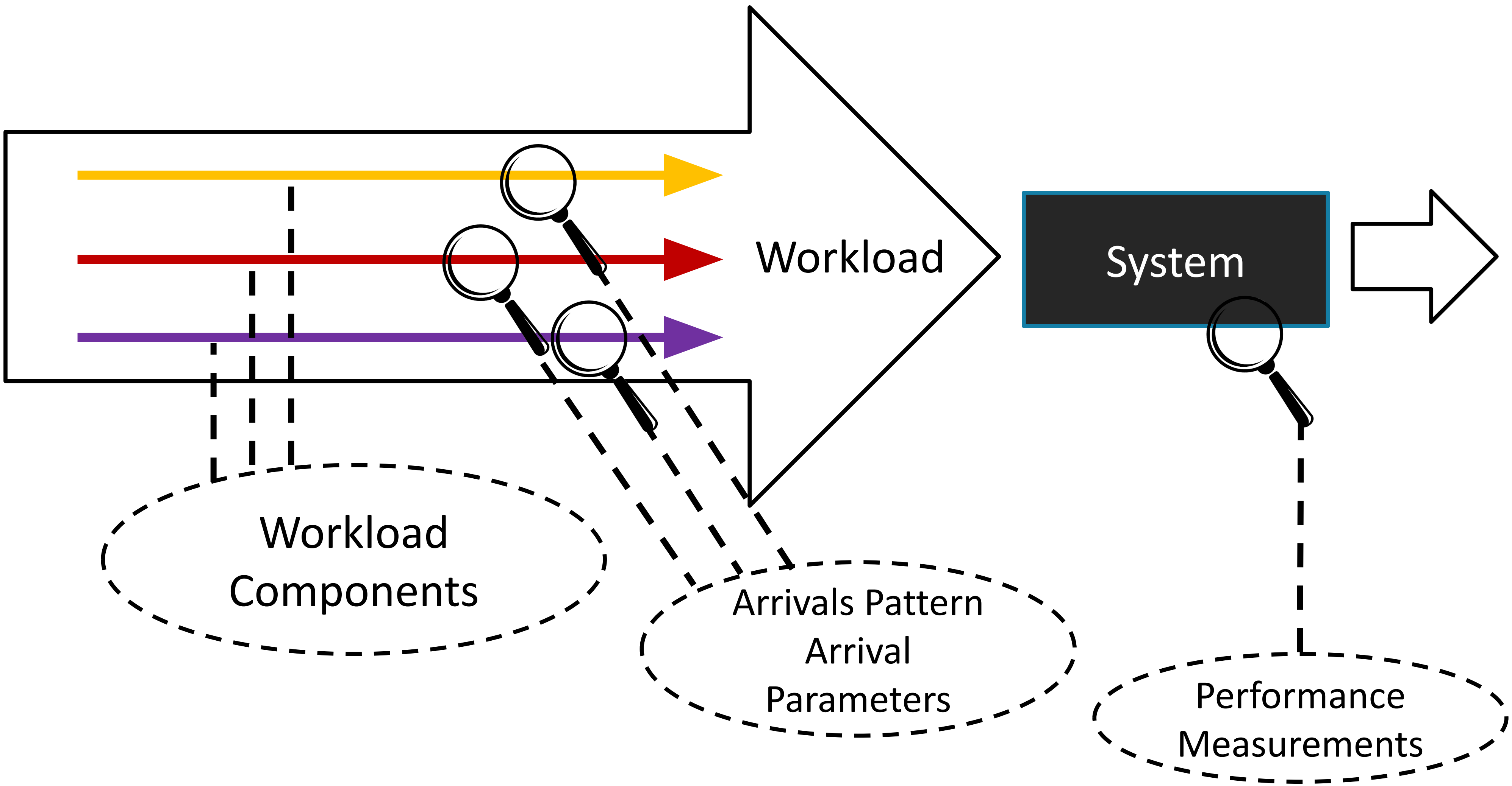
- media properties,
- traffic properties,
- user behaviour,
- social sharing properties.

Mobile Device Workloads

- traffic volume,
- access time,
- unique subscribers,
- Locations.

There exists dedicated literatures about the characterization of these and other specific parameters characterizing the workloads

Workload Characterization Recap Picture



Collect Measures

Workload characterization relies on **experimental approaches** based on the analysis of **measurements collected** on the technological infrastructures while they are operating (i.e., under their real workloads).

Active and passive monitoring techniques are often applied for obtaining workload measurements

Active techniques usually generate artificial (controlled) workloads on a real infrastructure with the only objective of monitoring

Passive techniques collect measurements on an infrastructure while it is operating, that is, under its actual workload

NOTE: take into account that monitors may introduce overhead!

Collect Measures: too many data?

In general, the amount of **data being collected can become quite large** and sometimes even intractable

→ **Identify the time window**

Appropriate **sampling techniques** may need to be applied. Since there might be the danger of ignoring events referring to rare significant workload components, it is very important to ensure the representativeness of the data sample being considered

Analysis

Statistical Analysis Techniques: application of statistical and visualization techniques

Descriptive statistics and measures of dispersions (e.g. mean, range, variance, coefficient of variation, skewness, median, percentiles) are useful to summarize the properties of each attribute

Analysis

Real workload can be viewed as a **collection of heterogeneous components**

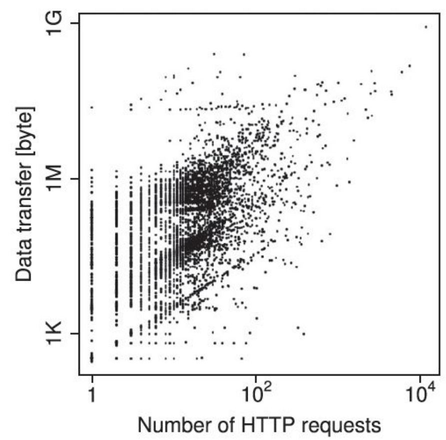
→**Partition the workload**, i.e. divide it into a series of classes such that their population are formed by quite homogeneous components

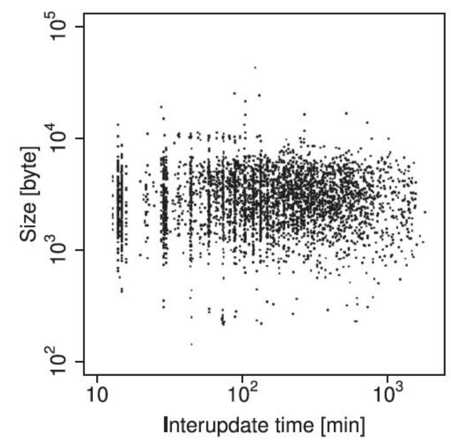
For analysis purposes, it is useful to classify these components into a small number of classes or clusters such that **the components within a cluster are very similar to each other**

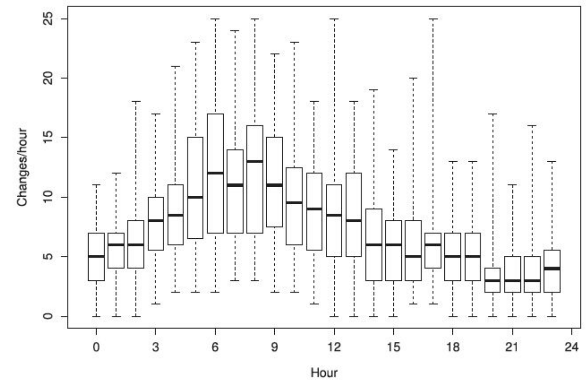
Analysis

Diagrams, such as histogram, scatter plot or box plots, may provide initial hints to interpret collected data

Scatter plots highlight the **correlation** between attributed, whereas box plots summarize their **distribution**







Analysis

atypical behaviour

The term **outlier** denotes the workload components characterized by an atypical behaviour of one or more attributes.

It is critical to take the right approach toward outliers because of their potential effects on the workload models.

Outliers **could indicate phenomena or properties previously unknown**, thus worth exploring.

On the contrary, they **could correspond to anomalous operating conditions** of the infrastructures or even errors in the measurements, thus worth discarding.

Analysis: Identify Components, Clustering

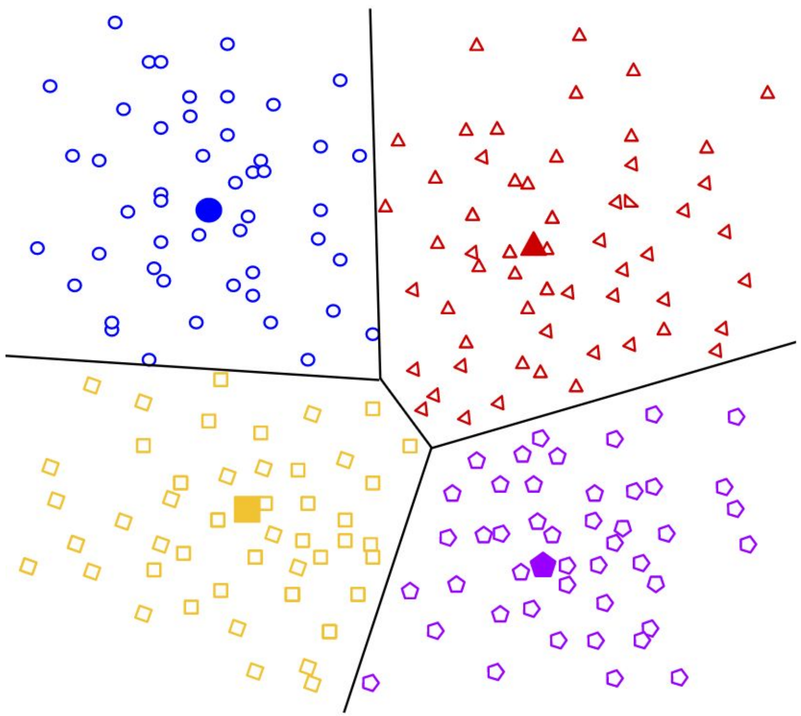
Clustering is an **unsupervised** process that **subdivides a set of observations** (i.e., workload components) **into homogeneous groups** (i.e., clusters).

The components of each group are very similar, whereas the components across groups are quite distinct.

The centroids (i.e., the geometric center of the clusters) are often used as representatives of the groups.

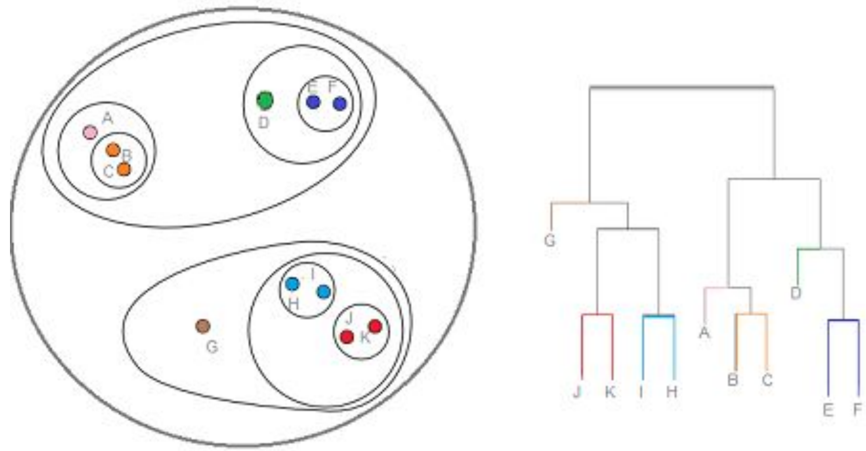
Distance-based clustering techniques **differ** for the **algorithms applied** (e.g., hierarchical, iterative) and their **similarity measures** (e.g., Euclidean distance, Manhattan distance).

Some Clustering Algorithms



Centroid-based Clustering

(e.g. K-means)



Hierarchical Clustering

Clustering

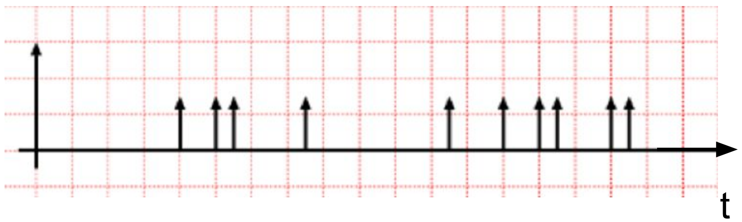
Example https://bit.ly/32XSJrf

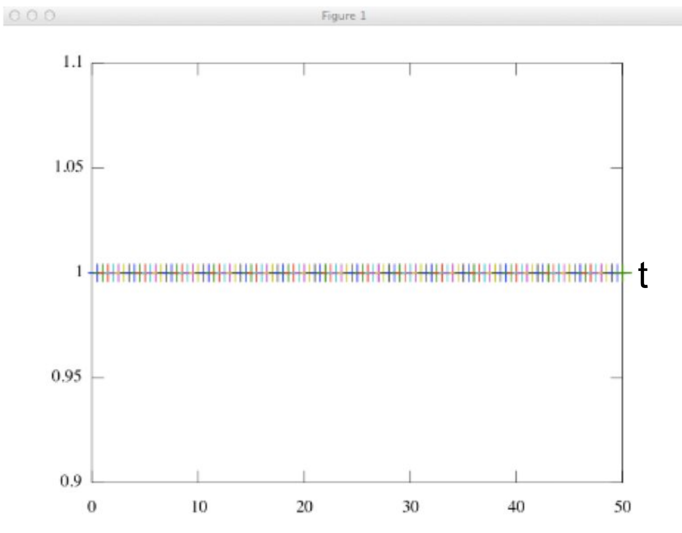
An overview about clustering is available at

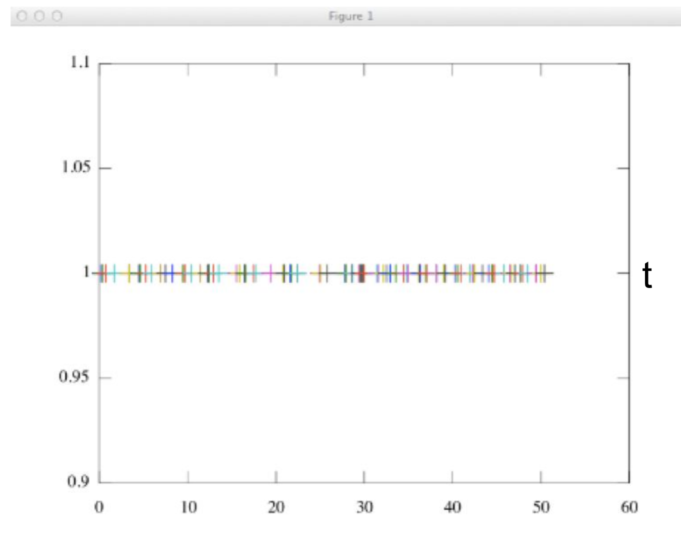
https://developers.google.com/machine-learning/clustering/overview

Characterizing the Arrival Pattern of a Single Component

Arrivals → System/Component → Completions

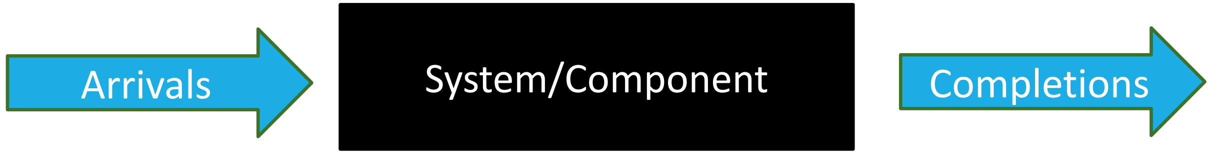


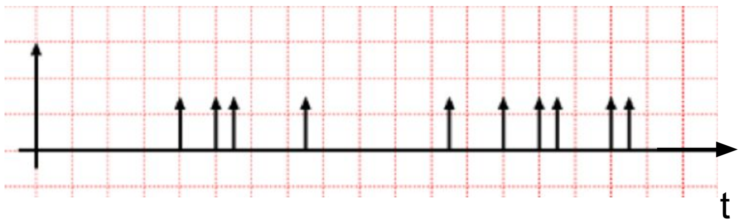




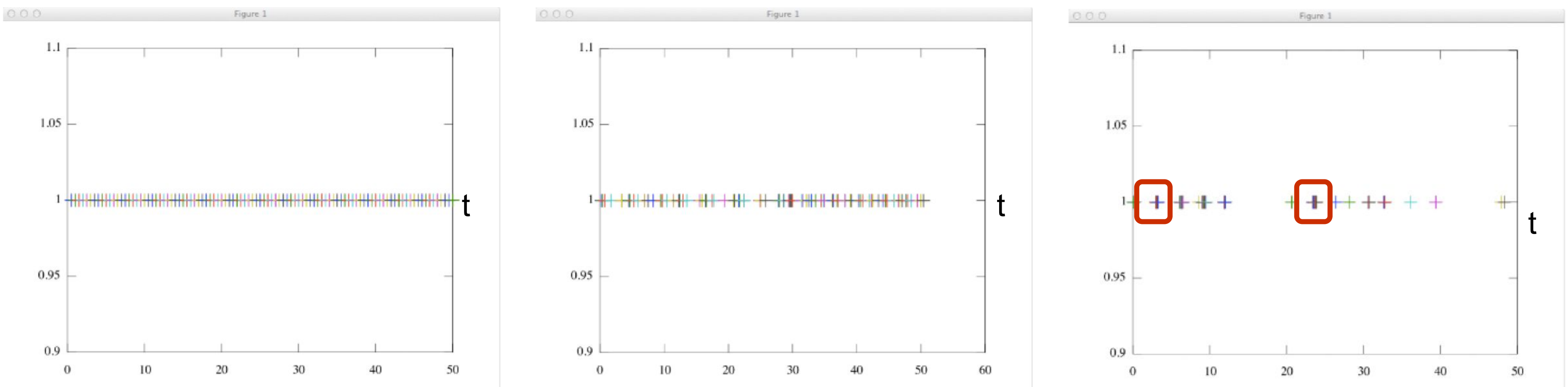
The mean arrival rate λ of the two workloads is the same

Characterizing the Arrival Pattern of a Single Component

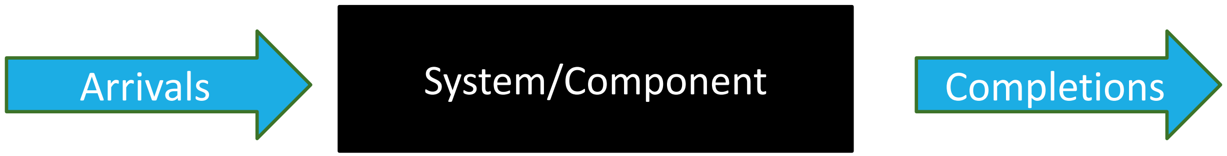


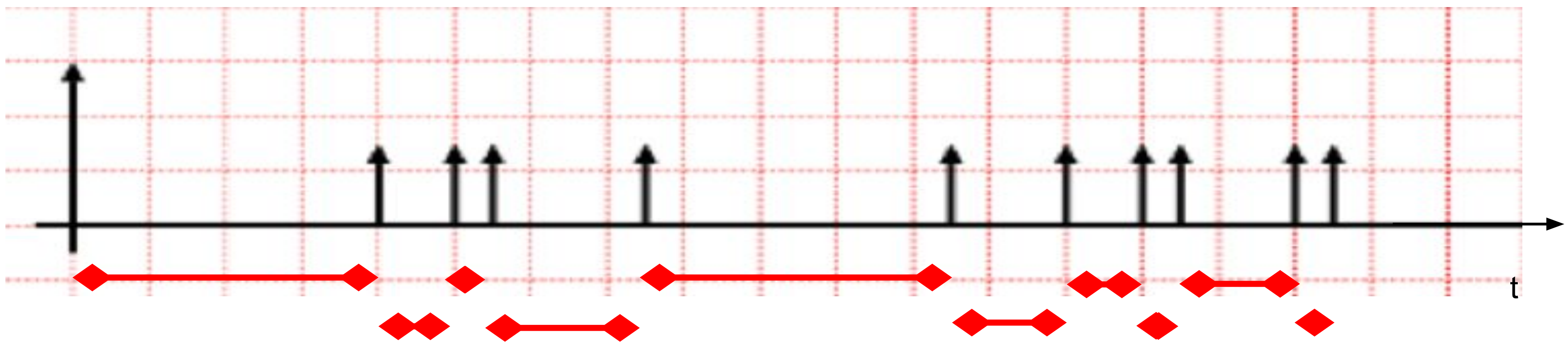


- ***rate*** (i.e. how fast they arrive)
- ***regularity*** (i.e. the time that passes between two occurrences)
- ***correlation*** (informally, inter-arrival are independent or there is a correlation?



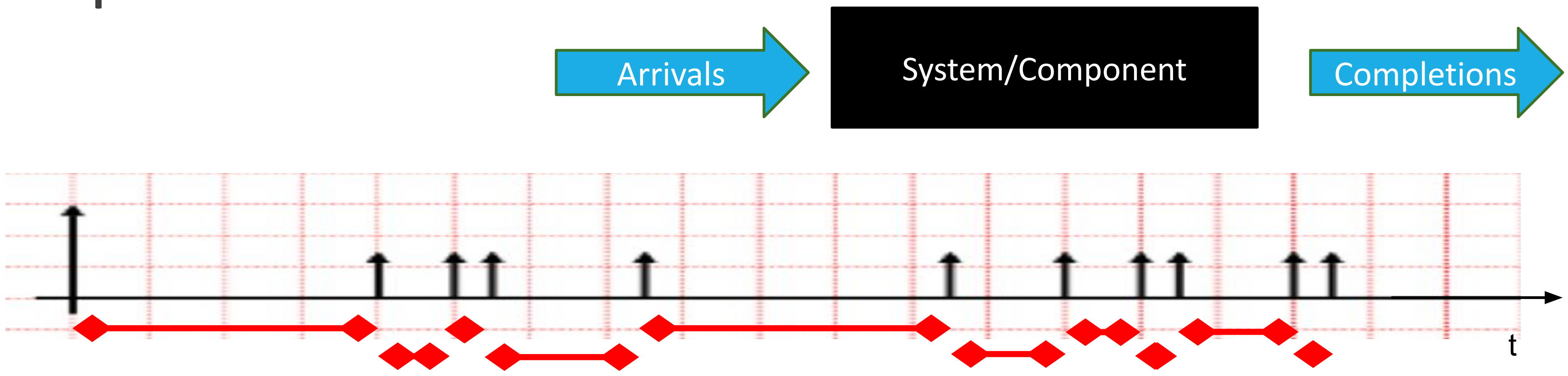
Characterizing the Arrival Pattern of a Single Component





Inter-arrival time: time that passes between an arrival and the following one

Characterizing the Arrival Pattern of a Single Component



Most of the times, the arrivals to a system are not deterministic, namely they do not occur exactly with the same pattern (same arrivals with the same inter-arrival times). **They are random**.

e.g., let us assume we deployed an e-commerce; we cannot know at what times requests will arrive to the service (search an item, add to cart, etc.) but, looking at the log of previous request (or asking to some expert) we may predict the possible target of such request (what is the mean inter-arrival time of all the request, if there are bursts at specific hours, etc.)

We model the arrivals of our system through a **random variable**

Random Variable in a Nutshell

Informally: a variable is called a random variable if **it takes one of a specified set of values with a specified probability**

It can either be continuous or discrete

The description of **how likely** a random variable takes one of its **possible values** can be given by a **probability distribution**

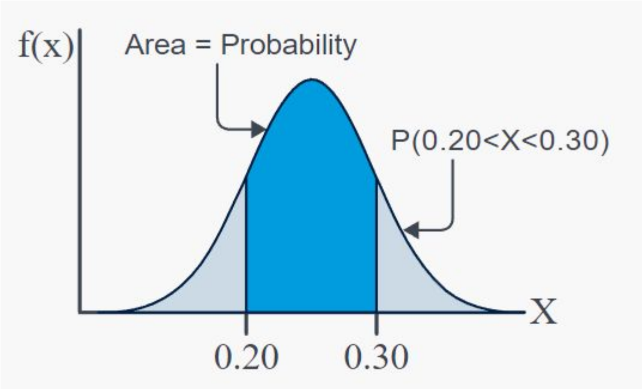
The probability distribution is a mathematical function that gives the probabilities of different outcomes for an experiment

Random Variable in a Nutshell

Continuous Random Variable, Probability Distribution

_> **Probability Density Function (PDF)**

PDF is used to **specify the probability of the random variable falling within a particular range of values** (a continuous random variable takes on an uncountably infinite number of possible values, the probability that the variable takes on any particular value is 0)

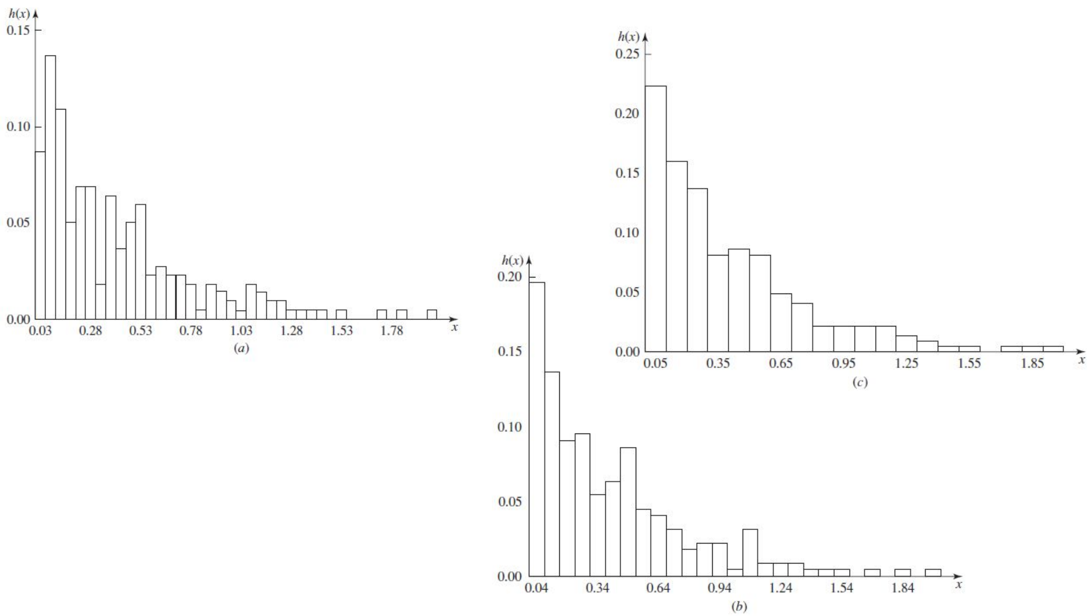


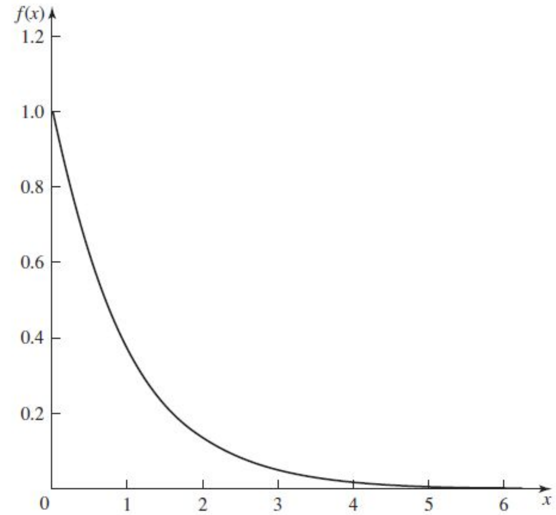
https://online.stat.psu.edu/stat414/lesson/14/14.1

https://towardsdatascience.com/understanding-random-variables-and-probability-distributions-1ed1daf2e66

From Data to a Random Variable (simplified)

1. **Take the measures** (e.g., inter-arrival times between arrivals) **and order them** in increasing order.
2. **Partition data** in k adjacent intervals (bins) of the same size, then count the number of occurrences in every bin and plot _> make a histogram
3. **Get a graphical estimate** of the PDF





PDF exponential distribution

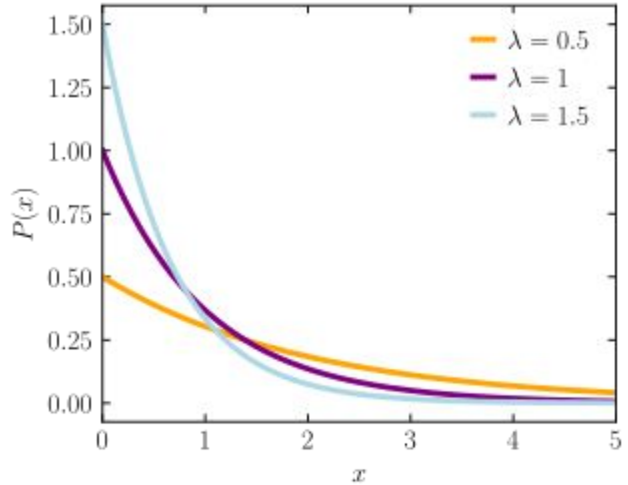
Exponential Distribution

We say that a random variable X is distributed Exponentially with rate λ,

$$X \sim \text{Exp}(\lambda)$$

if X has the probability density function:

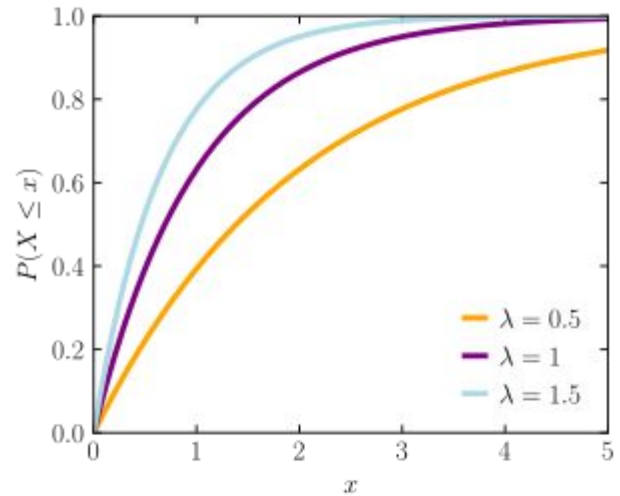
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0. \\ 0 & x < 0. \end{cases}$$



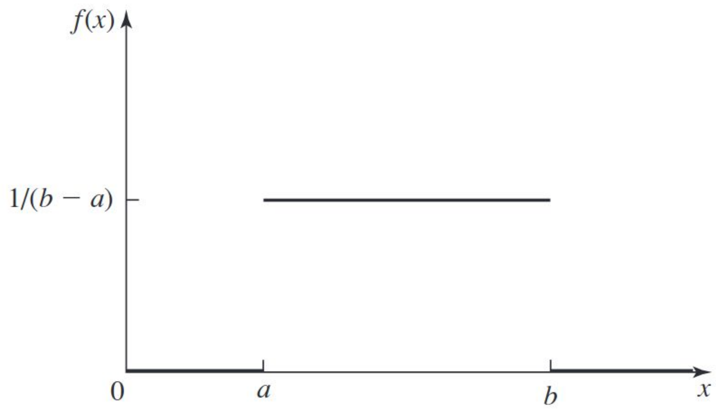
The cumulative distribution function, $F(x) = \mathbf{P}\{X \leq x\}$, is given by

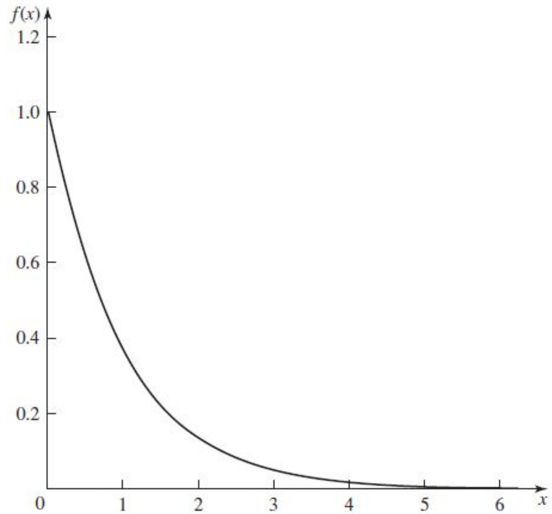
$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x} & x \geq 0. \\ 0 & x < 0. \end{cases}$$

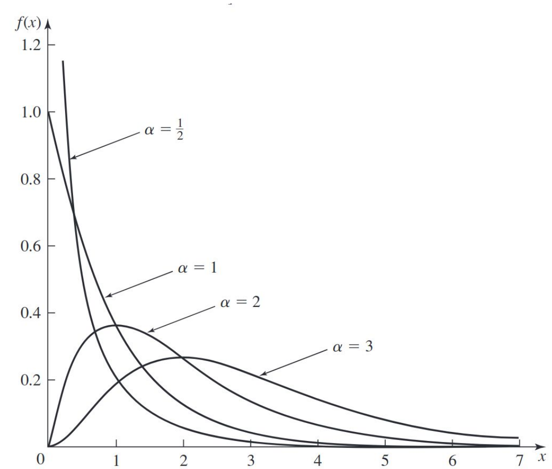
$$\overline{F}(x) = e^{-\lambda x}, \ x \geq 0.$$



Some Common Distributions (PDF)





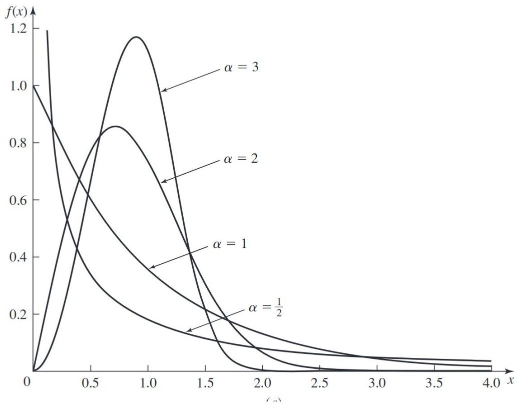


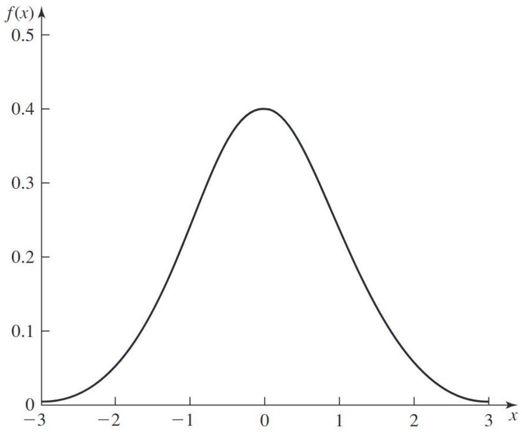
Uniform
Used as a "first" model for a quantity that is felt to be randomly varying between a and b but about which little else is known

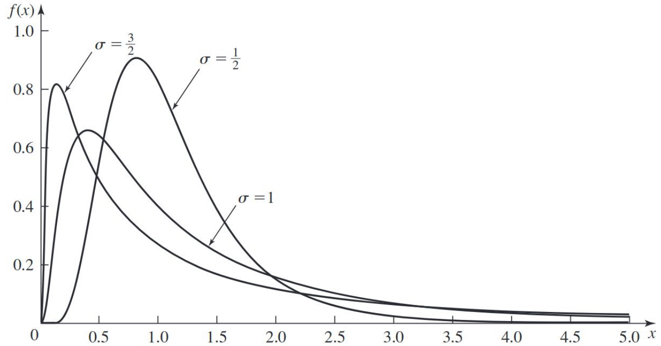
Exponential
Interarrival times of "customers" to a system that occur at a constant rate, time to failure of a piece of equipment

Gamma
Time to complete some task, e.g., customer service or machine repair

Some Common Distributions (PDF)







Weibul
Time to complete some task, time to failure of a piece of equipment; used as a applications rough model in the absence of data

Normal
Errors of various types

Lognormal
Time to perform some task

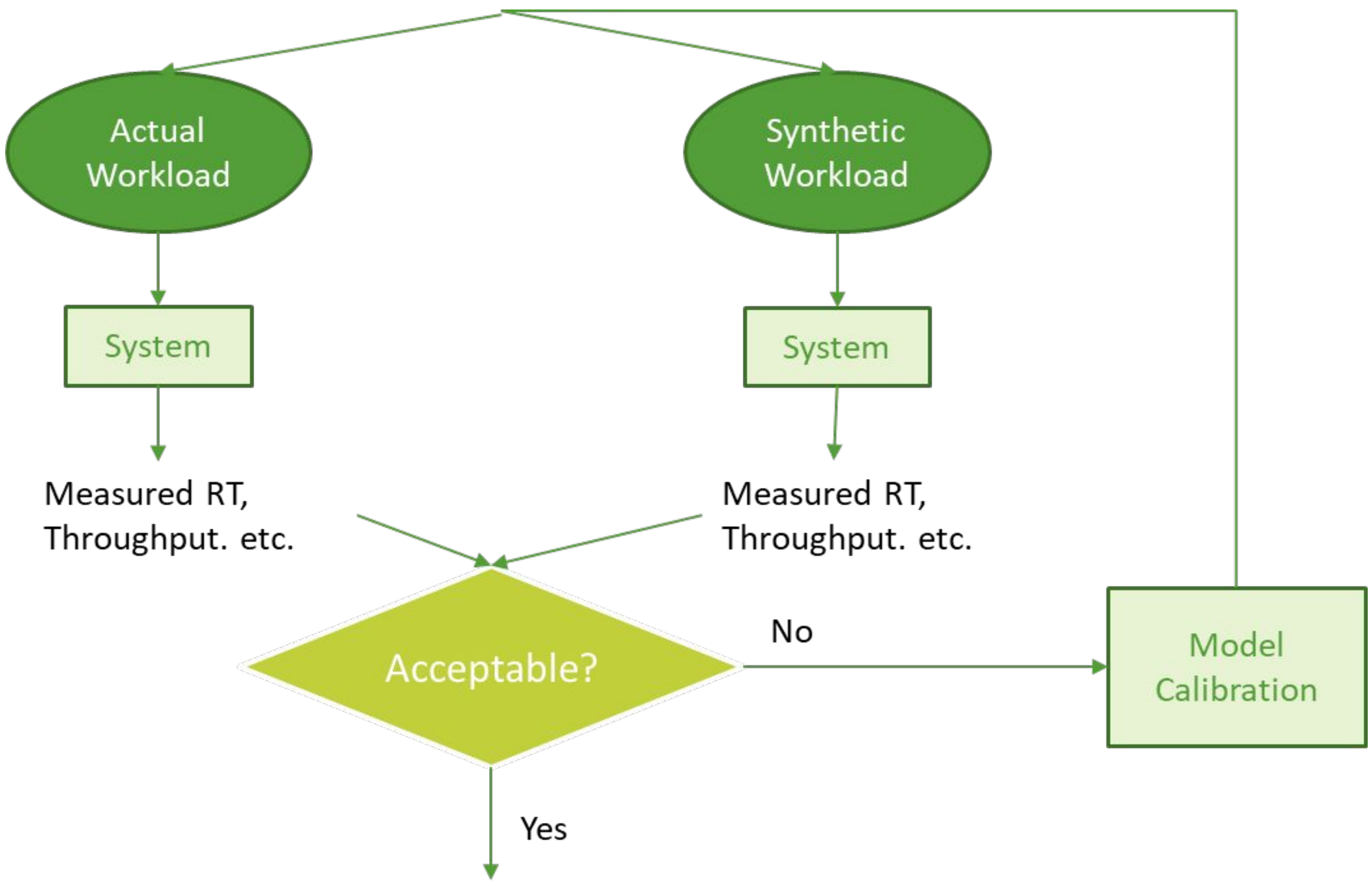
Distribution Fitting

Fitting distributions to data consists in **choosing a probability distribution modeling the random variable**, as well as finding parameter estimates for that distribution

https://medium.com/the-researchers-guide/finding-the-best-distribution-that-fits-your-data-using-pythons-fitter-library-319a5a0972e9

http://www.cs.unitn.it/~taufer/Readings/2014-JSS-fitdistrplus-An%20R%20Package%20for%20Fitting%20Distributions.pdf

Workload model validation and calibration



Workload forecasting

Forecasting is the art and science of predicting future events

Predicting how system workloads will vary in the future

It is a set of scenarios and assumptions

- Evaluating the organization’s workload trends;
- analyzing historical usage data;
- analyzing business or strategic plans;
- mapping plans into business processes

Benchmark

A benchmark is the act of running a computer program, a set of programs, or other operations, in order to **assess the relative performance of an object**, normally by running a number of standard tests and trials against it

Benchmarks are designed to mimic a particular type of workload on a component or system

References

- Chapter 3, 10 - D. A. Menascé, V. A. F. Almeida: *Capacity Planning for Web Services: metrics, models and methods*. Prentice Hall, PTR
(Available in the library inside Dipartimento di Ingegneria informatica, automatica e gestionale Antonio Ruberti)

- Chapter 6 - A. M. Law - Simulation modeling and analysis
https://industri.fatek.unpatti.ac.id/wp-content/uploads/2019/03/108-Simulation-Modeling-and-Analysis-Averill-M.-Law-Edisi-5-2014.pdf

- Calzarossa, Massari, Tessera. "Workload characterization: A survey revisited." ACM Computing Surveys (CSUR) 48.3 (2016): 1-43 https://doi.org/10.1145/2856127

- R. Jain, "The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling,“
https://www.cin.ufpe.br/~rmfl/ADS_MaterialDidatico/PDFs/performanceAnalysis/Art%20of%20Computer%20Systems%20Performance%20Analysis%20Techniques.pdf