Module 2

Biomedical Image Analysis

Biomedical Image Analysis

•An **image** is a **spatial map of one or more physical properties** of a **subject**.

•**Pixel intensity** represents the **value of a physical property of the subject** at that point.

•**Imaging the subject** is a way to **record spatial information**, **structure, and context information.**

•In this context, the **subject could be almost anything**: your **family sitting for a family photo taken with your smartphone**, the constellations of orion's belt viewed from a telescope, the **roads of your neighborhood imaged from a satellite**, **a child growing inside of its mother viewed using an ultrasound probe**.

•The list of possible subjects is endless, and the **list of possible imaging methods is long and ever expanding**.

•But the **idea of imaging is simple and straightforward**: **convert some scene of the world into some sort of array of pixels** that **represents that scene and that can be stored on a computer**.

- We are interested in **biomedical images**, which are a **subset of images that pertain to some form of biological specimen**, which is generally some part of human or animal anatomy.

- The **imaging modality** used to acquire an **image imaging (MRI), computed tomography (CT), positron emission tomographof that specimen generally falls into one of the categories of magnetic resonance y (PET), ultrasound (U/S) etc.**

- Wide range of **microscopy modalities** such as fluorescence, bright-field, and electron microscopy. Such modalities have **various purposes: to image inside of the body without harming the body or to image specimens that are too small to be viewed with the naked eye.**

- Biomedical image analysis is the **solution** to **this problem of too much data**

- Such analysis methods enable the extraction of quantitative measurements and inferences from images.

- Hence, it is possible to **detect and monitor certain biological processes and extract information about them.**

Biomedical Imaging Modalities

✓In this section, we provide a brief introduction to several biomedical imaging modalities with emphasis on unique considerations regarding image formation and interpretation.

✓Understanding the appearance of images resulting from the different modalities aids in designing effective image analysis algorithms targeted to their various features.

Computed Tomography

- Computed Tomography (CT) creates 2D axial cross-section images of the body.
- by collecting several 1D projections of conventional X-ray data using an X-ray source on one side and a detector on the other side.
- The 1D projection data are then reconstructed into a 2D image.
- Modern CT systems are capable of acquiring a large volume of data extremely fast by increasing the axial coverage.
- A CT image displays a quantitative CT number usually reported in Hounsfield units, which is a measure of the attenuation property of the underlying material at that image location.

Fig. anatomical imaging

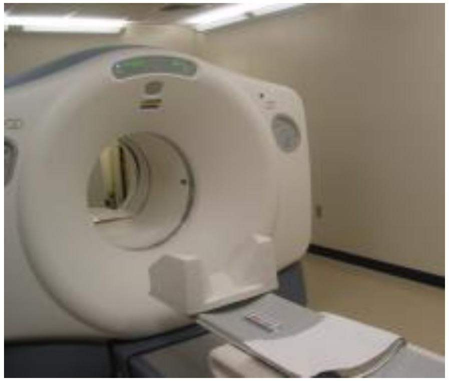
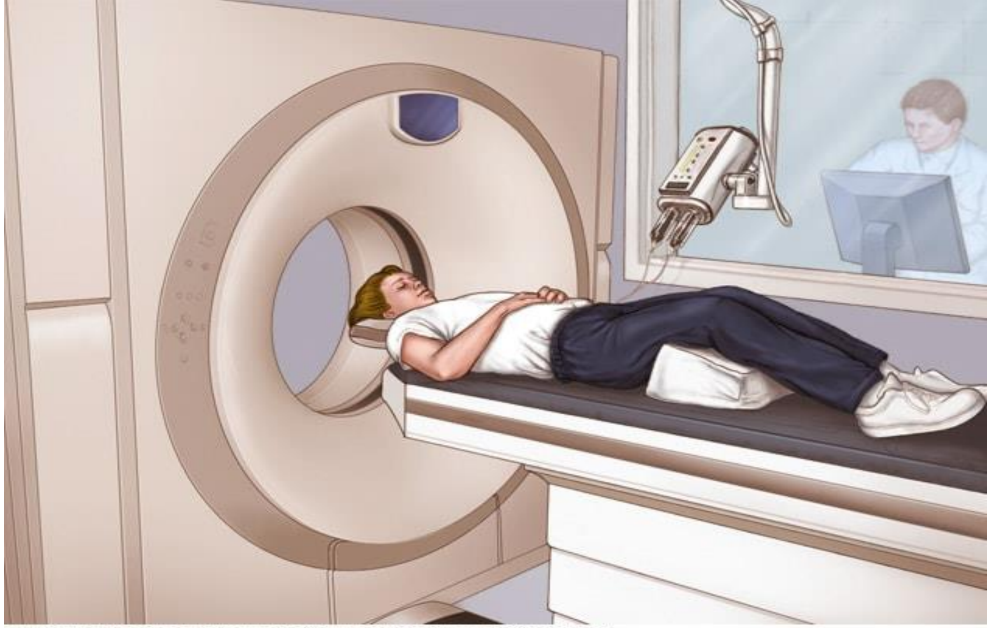


Fig. MRI



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Fig.   CT SCAN

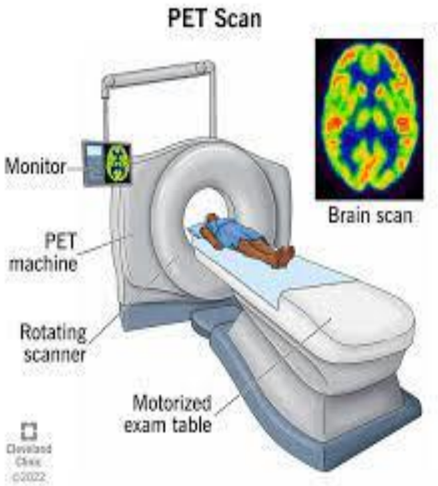


Fig. PET

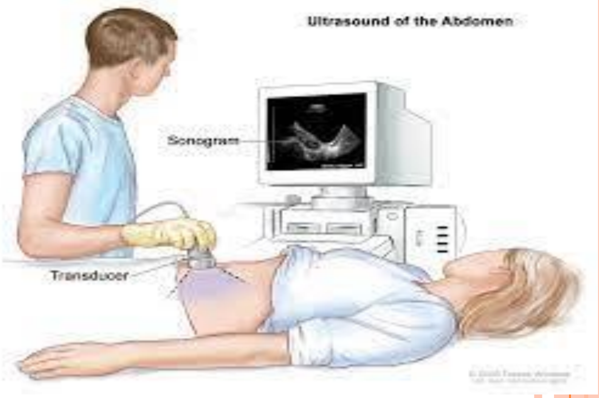


Fig. ultrasound (U/S)

- CT has become the mainstay of diagnostic imaging due to the very large number of conditions that are visible on CT images.
- Dual Energy CT systems, where CT images are acquired at two different energy levels. This makes it possible to do a very rich characterization of material composition using differential attenuation of materials at two different energy levels.
- modern iterative model-based methods are able to achieve excellent reconstruction while limiting doses to a patient.

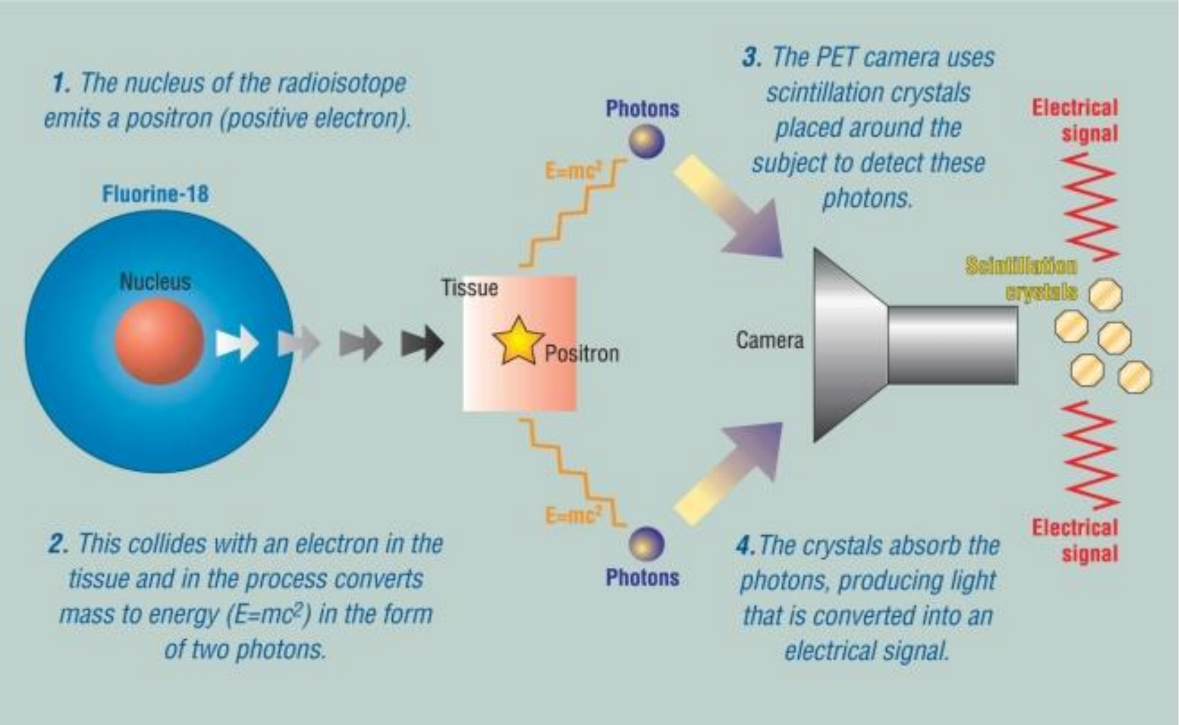
***Positron Emission Tomography (*PET):

•Positron emission tomography (PET) is a technique that measures physiological function by looking at blood flow, metabolism, neurotransmitters, and radiolabelled drugs.

•It is a nuclear imaging modality create activity maps inside the body based on uptake of a compound based on metabolic function.

•It allows assessment of important physiological and biochemical processes.

•Before meaningful and quantitatively accurate activity uptake images can be generated, corrections for scatter and attenuation must be applied to the data.


1. The nucleus of the radioisotope emits a positron (positive electron).
Fluorine-18
Nucleus
Tissue
Positron
Photons
E=mc²
E=mc²
Photons
3. The PET camera uses scintillation crystals placed around the subject to detect these photons.
Electrical signal
Camera
Scintillation crystals
Electrical signal
2. This collides with an electron in the tissue and in the process converts mass to energy (E=mc²) in the form of two photons.
4. The crystals absorb the photons, producing light that is converted into an electrical signal.

Magnetic Resonance Imaging (MRI)

•It is a high resolution, high contrast, non-invasive imaging modality with extremely rich and versatile contrast mechanisms that make it the modality of choice for looking at soft tissue contrast.

•In conventional MRI, signals are formed from nuclear magnetic response properties of water molecules that are manipulated using external static and varying magnetic fields and radio-frequency pulses.

•In addition to looking at anatomy and structure, image acquisition methods can be tailored to yield functional information such as blood flow.

•Images with very different contrasts can be created to selectively highlight and/or suppress specific tissue types.

•Spatially varying gradients of magnetic fields are used to form 2D or 3D images.

•Received data is typically reconstructed using Fourier methods.

Ultrasound

- It is one of the most ubiquitous imaging modalities due in large part to its low cost and completely non-invasive nature.

- IT transmits high frequency sound waves using specialized transducers, and then collects the reflected ultrasound waves from the body using specialized probes.

- The variable reflectance of the sound waves by different body tissues forms the basis of an image.

- Ultrasound can also depict velocities of moving structures such as blood using Doppler imaging.

- Due to very fast acquisition times, it is possible to get excellent real-time images using ultrasound to see functioning organs such as the beating heart.

- Modern ultrasound systems employ sophisticated electronics for beam forming and beam steering, and have algorithms for pre-processing the received signals to help mitigate noise and artifacts.

Microscopy

- In addition to in radiological imaging, clinical diagnosis as well as research frequently makes uses of in vitro imaging of biological samples such as tissues obtained from biopsy specimens.
- These samples are typically examined under a microscope for evidence of pathology.
- Traditional bright field microscopy imaging systems utilize staining with markers that highlight individual cells or cellular compartments or metabolic processes in live or fixed cells.
- Images from microscopy systems are traditionally read visually and scored manually.
- However, newer digital pathology plat-forms are emerging and new methods of automated analysis and analytics of microscopy data are enabling more high-content, high-throughput applications.
- Using image analysis algorithms, a multitude of features can be quantified and automatically extracted and can be used in data-analytic pipelines for clinical decision making and biomarker discovery.

❑ **Medical imaging data is commonly stored and managed using specialized systems known as Picture Archiving and Communications System (PACS).**

❑ **PACS systems house medical images from most imaging modalities and in addition can also contain electronic reports and radiologist annotations in encapsulated form.**

Object Detection

- Object detection is a computer vision task that involves identifying and locating objects in images or videos. It is an important part of many applications, such as self-driving cars, robotics, and video surveillance.

- Detection is the process through which regions of potential interest, such as anatomical structures.

- Often associated with detection is the localization of the targeted structures.

- Here, the word “detection” is used specifically to designate the joint detection and localization of a structure of interest.

1.Template Matching

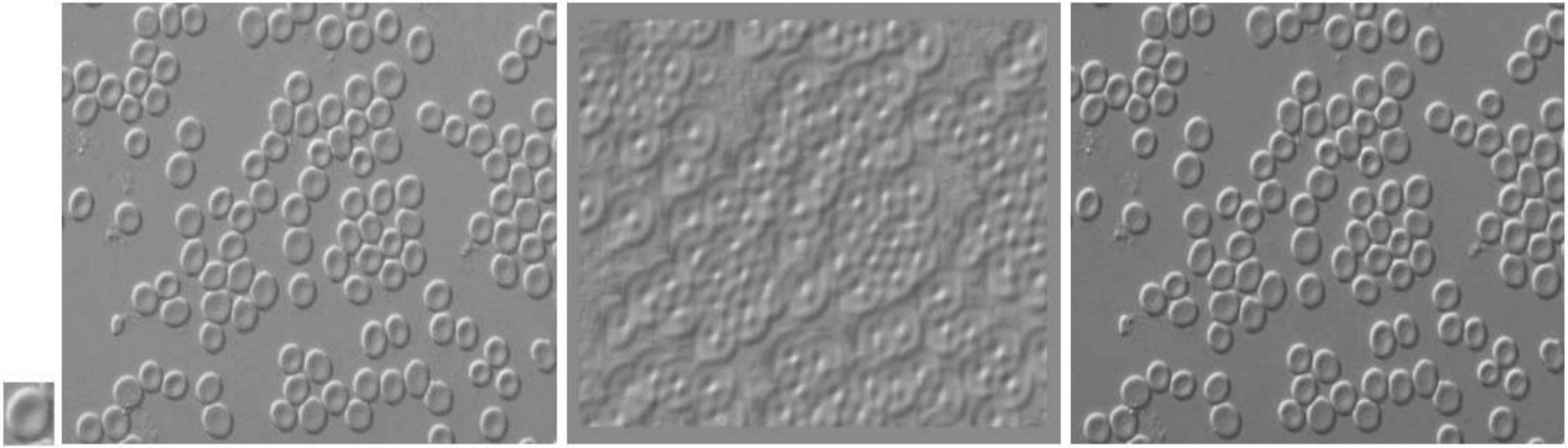
•It is an often-used method to detect objects of interest in an image is to choose a representative template and apply some variant of template matching to find similar regions in the image of interest.

•Using an approach such as normalized cross-correlation (NCC) measures the similarity between the two signals f1 and f2. This yields an output map showing the magnitude of the match, and this can be thresholded to find the best detections in the image.

•If we define f1 as the fixed image and f2 as the moving image or template image, the normalized cross-correlation between images f1 and f2 at a given (u,v) is defined as

$$\frac{\sum\left[\left(f_1(x,y)-\overline{f_{1,u,v}}\right)\left(f_2(x-u,y-v)-\overline{f_{2,u,v}}\right)\right]}{\sqrt{\sum\left(f_1(x,y)-\overline{f_{1,u,v}}\right)^2}\sqrt{\sum\left(f_2(x-u,y-v)-\overline{f_{2,u,v}}\right)^2}}$$

An example of the effectiveness of this approach can be seen in below figure , where a small template is matched with an entire image of cells imaged with differential interference contrast (DIC) microscopy, and the resulting NCC map is thresholding to yield strong detections in almost all of the



(a) Template (b) Input image (c) NCC map (d) Thresholded detections

Figure: Normalized cross-correlation (NCC) example for template matching. The thresholded NCC map serves as the detections for the cells in this DIC image.

2 Model-Based Detection

•Model-based detection methods are a generalization of template matching.

•Obtained by replacing the template and the NCC function with arbitrary models and figures of merit for the matching between the model and the data.

•In such methods, an arbitrary statistical model of features presumably found in the structure of interest.

•When presented with an image, such methods compute the selected features throughout the image and evaluate a figure of merit that indicates whether the computed features are consistent with the presence of the structure of interest at any given location.

•Early detection of colonic polyps has been associated with reduction in the incidence of colorectal cancer

•Optical colonoscopy has been shown to be an effective tool for polyp detection. However, optical colonoscopy is an invasive procedure, and discomfort to the patient.

•The use of cleansing materials and colonic fluids produces severe alterations in the appearance of the image.

•This poses a challenge to methods solely based on geometry therefore, a joint modeling of shape and appearance can applied.

3 Data-Driven Detection Methods

•Model-based methods, although powerful, are difficult to apply when expert knowledge is not available or is not in a format that can be easily encoded in algorithmic form.

•To address this problem, data-driven methods apply machine learning techniques to automatically extract features from labeled data.

•An additional difficulty of model-based methods is the need for explicit models for the structure or anatomical region of interest.

•Data-driven methods, on the other hand, can be used to construct models of normal regions, which are hopefully more common, and the detection problem is then translated into anomaly detection.

•here the objective is simply to locate structures or regions that do not conform to the norm, without explicit modeling of non-conforming structures.

•Unsupervised learning methods, such as PCA, can be used to discover and retain the more relevant modes of variation of the input, capturing the regularity of the input training data.

•When non-conforming data is presented to the algorithm, deviations from such regularity will become apparent, and abnormalities can therefore be detected.

•An **example** of an unsupervised data-driven method is found in the detection of carotid plaques.

•The availability of treatments that slow the progression of cardiovascular disease (CVD) increases the impact of early diagnosis in patient survival

•The presence of carotid plaque has been identified as a significant risk factor in the prognosis of CVD.

•A data-driven detection method was applied to the problem of detecting carotid plaques depicted in ultrasound images.

Image Segmentation

•The goal of image segmentation is to divide a digital image into separate parts or regions.

•The regions have a strong correlation with objects or areas of the real world contained in the image.

•This is used to locate objects and boundaries in images.

•Dividing the image into meaningful regions simplifies the representation of an image into something that is more meaningful and easier to analyze.

•Segmentation is one of the most important steps leading to the analysis of image data because it enables the further analysis of individual objects.

•in general image segmentation is not well defined and is very challenging because of the difficulty of defining and identifying the particular shapes of the segmented objects.

•A large number of different segmentation algorithms.

#Thresholding, #watershed, #region-growing, #classification, #wavelets, and #level-sets.

1 Thresholding

- The most simple and intuitive segmentation approach is thresholding.

- It separates an image into foreground/background using a cut-off value.

- It can be accomplished through one simple loop over the image with the following operation:

  if a pixel value xi is greater than t, set the new value to a foreground value (such as 255),
  and
  if it is less than t, set the new value to a background value (such as 0).

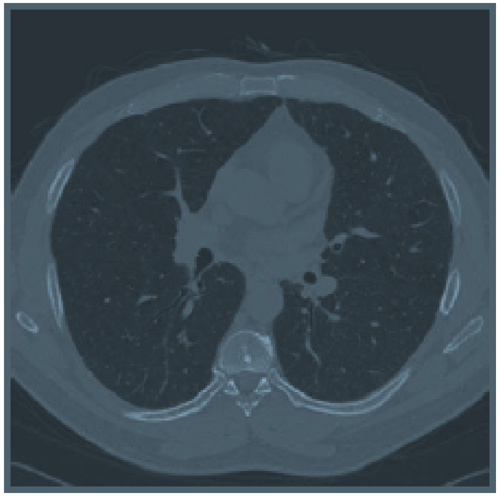
- The point is that the pixels are divided into two groups (creating a binary image) depending on their value relative to t.

- It is easy to see that increasing t increases the number of background pixels and vice versa.

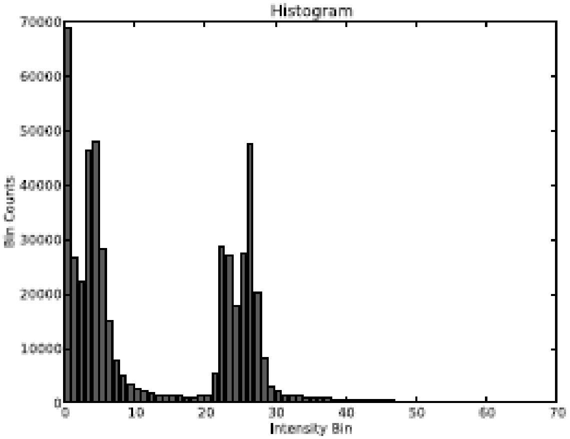
- The advantages of thresholding are speed, simplicity, and the ability to specify multiple thresholds.

- The disadvantages are that the objects must have similar appearance, it does not take into account any spatial information.

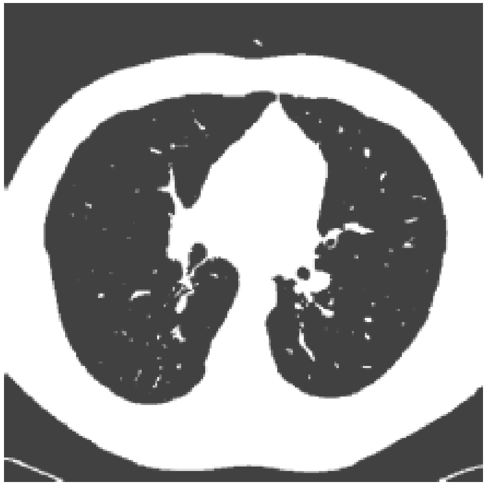
Biomedical Image Analysis



Original image



Histogram



Thresholded image

Many thresholding approaches are based on the image histogram, which is a simple transformation of the image whereby pixels with the same or similar intensities are grouped together into a one-dimensional array. In this array, the index represents an intensity value (or a small range of intensity values), and the value at each index represents the count of the number of pixels with that intensity (or range of intensities).We can use the image statistics to separate the image background from the foreground. For certain types of images this will provide good results.

2. Watershed Transform

•The watershed transform is an algorithm/solution framework that is very commonly used for image segmentation and for binary shape separation.

•The **watershed algorithm** is a computer vision technique used for image region segmentation.

•The segmentation process will take the similarity with adjacent pixels of the image as an important reference to connect pixels with similar spatial positions and gray values.

•Constitute a closed contour(outline), and this closure is an important feature of the watershed algorithm. Means it is an algorithm that correctly determines the "**outline of an object**".

•The watershed algorithm uses topographic information to divide an image into multiple segments or regions.

•The algorithm views an image as a topographic surface, each pixel representing a different height.

•The whole process of the watershed algorithm can be summarized in the following steps:-

1. **Marker placement:** The first step is to place markers on the local minima, or the lowest points, in the image. These markers serve as the starting points for the flooding process.

2. **Flooding**: The algorithm then floods the image with different colors, starting from the markers. As the color spreads, it fills up the catchment basins until it reaches the boundaries of the objects or regions in the image.

3. **Catchment basin formation**: As the color spreads, the catchment basins are gradually filled, creating a segmentation of the image. The resulting segments or regions are assigned unique colors, which can then be used to identify different objects or features in the image.

4. **Boundary identification**: The watershed algorithm uses the boundaries between the different colored regions to identify the objects or regions in the image. The resulting segmentation can be used for object recognition, image analysis, and feature extraction tasks.

3 Region Growing

•Region-growing methods rely mainly on the assumption that the neighboring pixels within one region have similar values.

•The common procedure is to compare one pixel with its neighbors.

•If a similarity criterion is satisfied, the pixel can be set to belong to the cluster as one or more of its neighbors.

•The selection of the similarity criterion is significant and the results are influenced by noise in all instances.

•This method takes a set of seeds as input along with the image.

•The seeds mark each of the objects to be segmented.

•The regions are iteratively grown by comparison of all unallocated neighboring pixels to the regions.

•The difference between a pixel's intensity value and the region's mean, is used as a measure of similarity.

3. Region Growing

- The pixel with the smallest difference measured in this way is assigned to the respective region.

- This process continues until all pixels are assigned to a region.

- **Pros:**

  Since it performs simple threshold calculation, it is faster to perform.
  Region-based segmentation works better when the object and background have high contrast.

- **Limitations**:

  It did not produce many accurate segmentation results when there are no significant differences b/w pixel values of the object and the background.

4. Clustering Segmentation

- Clustering is the process of grouping similar data points together and marking them as a same cluster or group.

- It is used in many fields including machine learning, data analysis and data mining.

- We can consider segmentation as a clustering problem.

- We need to cluster image into different object, each object’s pixels has common features for example same color or same intensity.

- Talking about similarity criteria takes us to what so-called feature.

- A feature is a value that measures or identify characteristic of a subject, it must discriminate between different subjects.

- **K means Clustering:** K means clustering Initially assumes random cluster centers in feature space. Data are clustered to these centers according to the distance between them and centers. Now we can update the value of the center for each cluster, it is the mean of its points. Process is repeated and data are re-clustered for each iteration, new mean is calculated till convergence. Finally we have our centers and its related data points.

Image registration

•Image registration is an image processing technique used to align multiple scenes into a single integrated image.

•For Image registration problems are encountered in the following types of applications:

*For motion correction or motion estimation, where two images contain the same anatomy but with some motion or deformation due to time difference between the two images.

*For multimodality registration, where two or more images represent different acquisition modalities for the same subject such as registering a CT image of a subject with an MRI image of the same subject. This is sometimes referred to as "Fusion."

*For intersubject comparisons, where images from two different subjects are registered to establish a spatial correspondence between the two images.

•Image registration is often used in medical & satellite imagery to align images from different camera sources.

Registration Transforms

The transform T is a function that maps physical points between the fixed and the moving image.

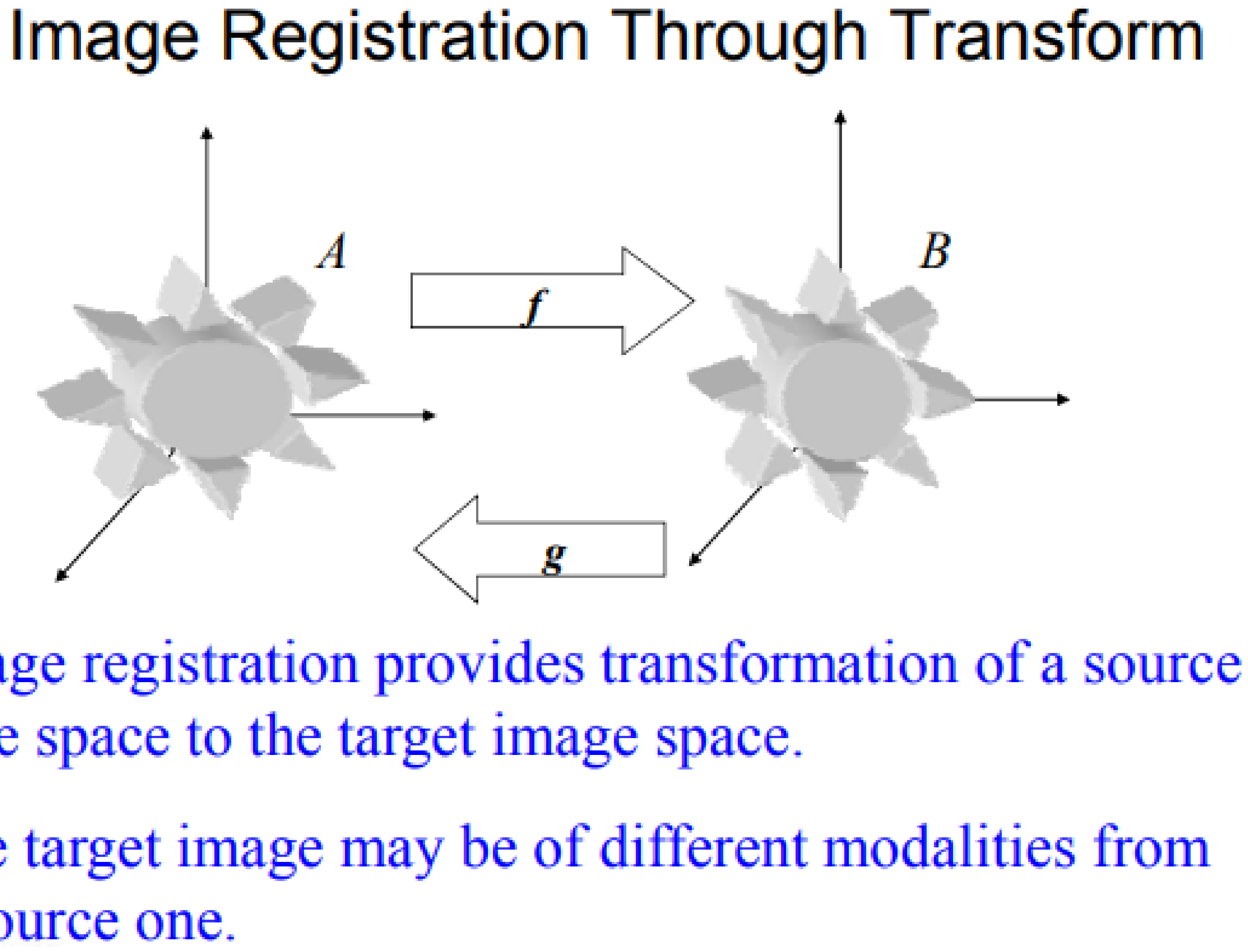
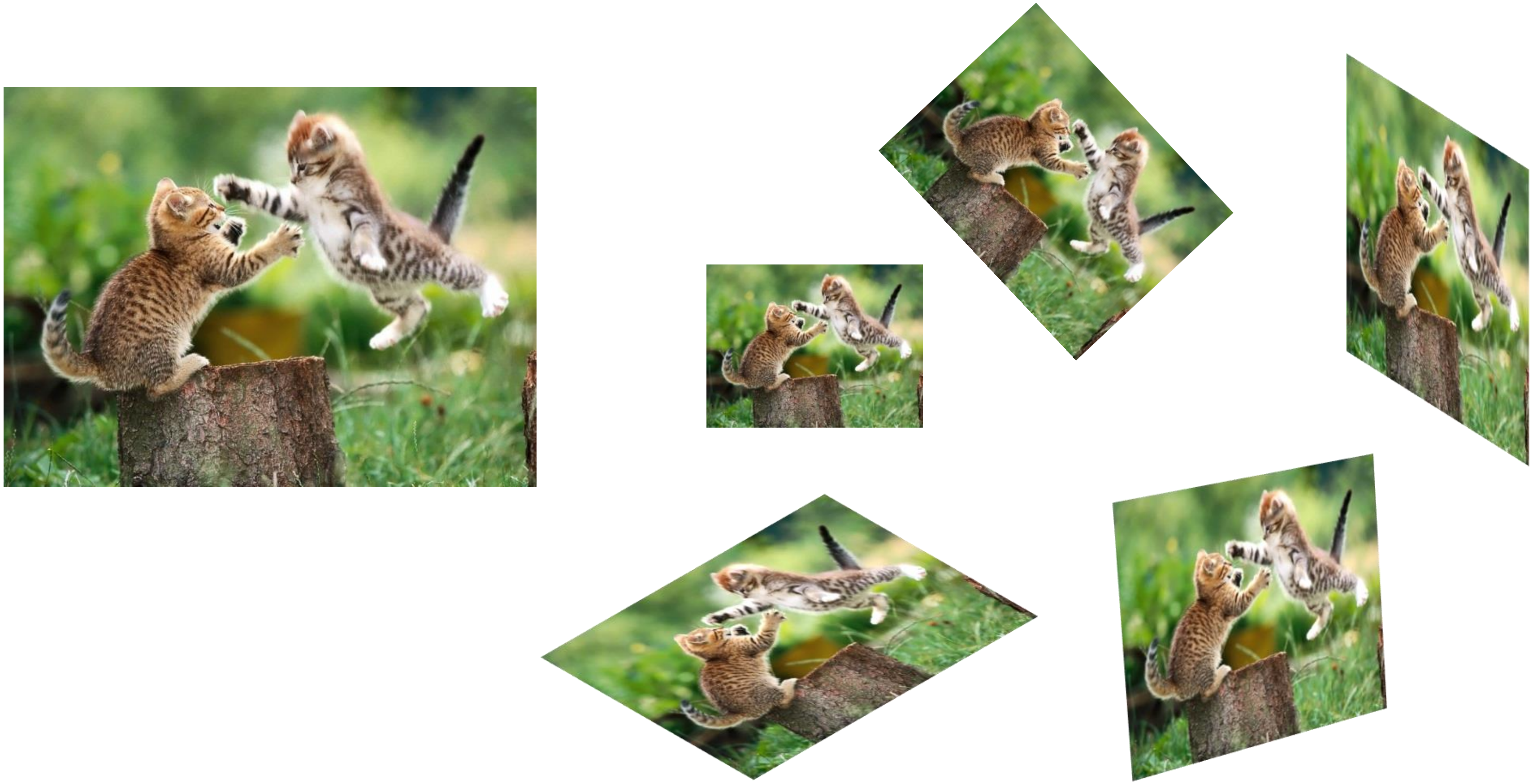


Image Registration Through Transform

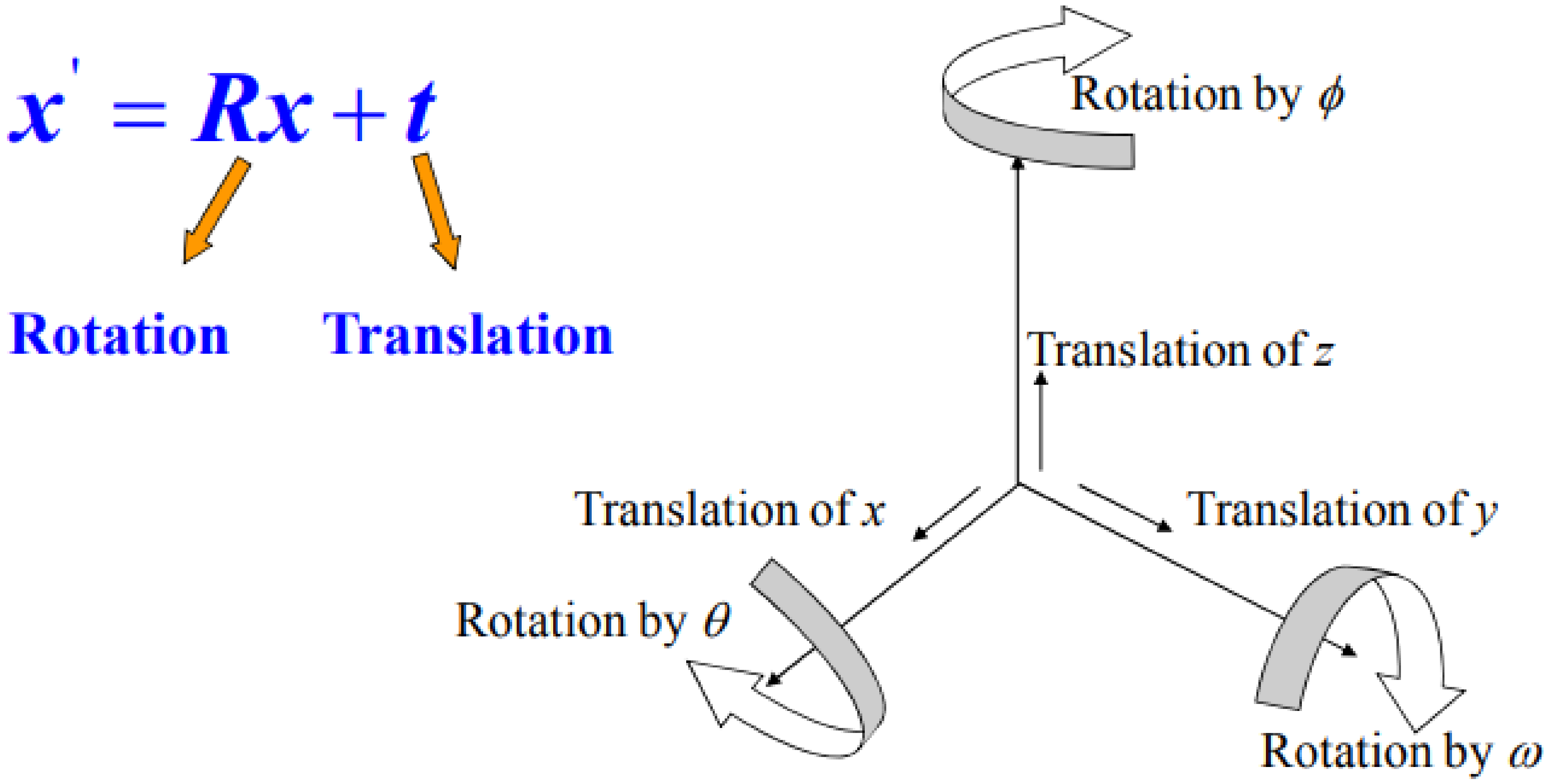
- Image registration provides transformation of a source image space to the target image space.
- The target image may be of different modalities from the source one.

What are geometric transformations?



1) **Rigid Body Transform:** A rigid body transform is comprised of image translation and rotation, and is represented by T(x) = R ∗ x+t, where R and t are the rotation matrix and translation vector respectively. In a rigid body transformation, distances, and angles between points and lines are preserved.

Rigid-Body Transformation

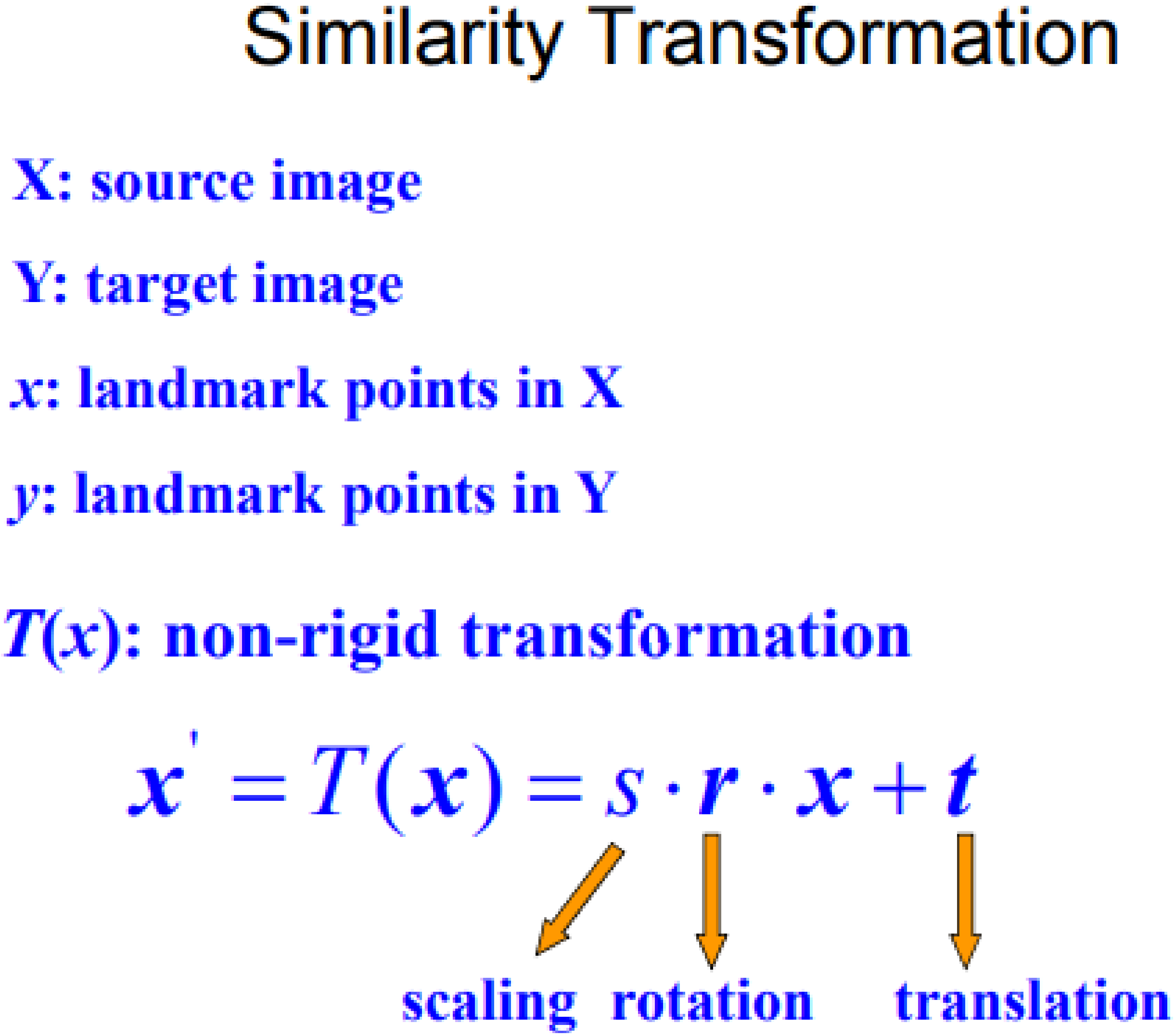


$$x' = Rx + t$$

Rotation

Translation

2) **Similarity Transform**: A similarity transform consists of an isotropic scaling factor in addition to the rigid body transformation of a rotation and translation. In a similarity tranformation, angles between lines are preserved, and objects change size proportionately in all dimensions.



Similarity Transformation

X: source image

Y: target image

***x*: landmark points in X**

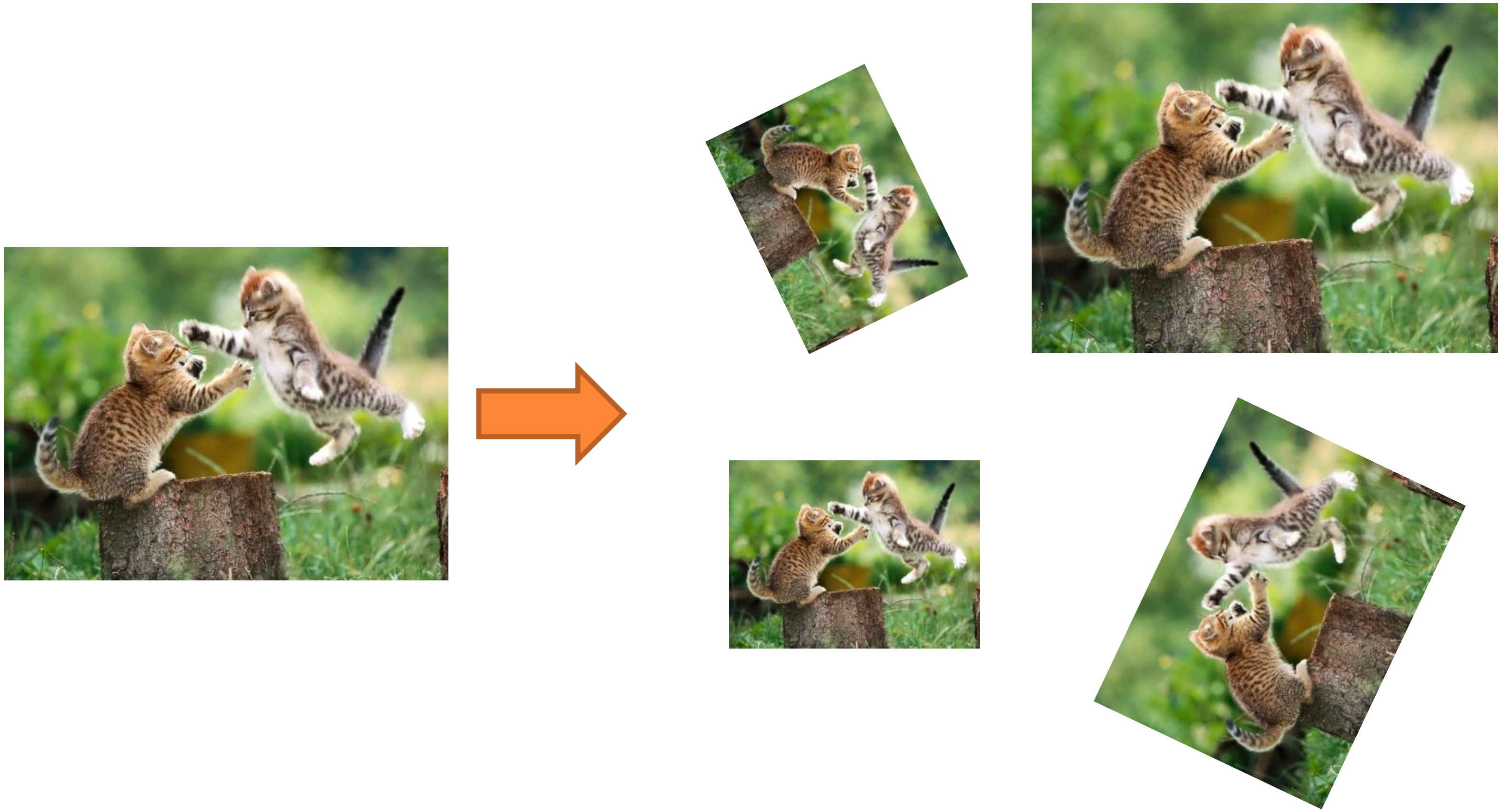
***y*: landmark points in Y**

***T*(*x*): non-rigid transformation**

$$x' = T(x) = s \cdot r \cdot x + t$$

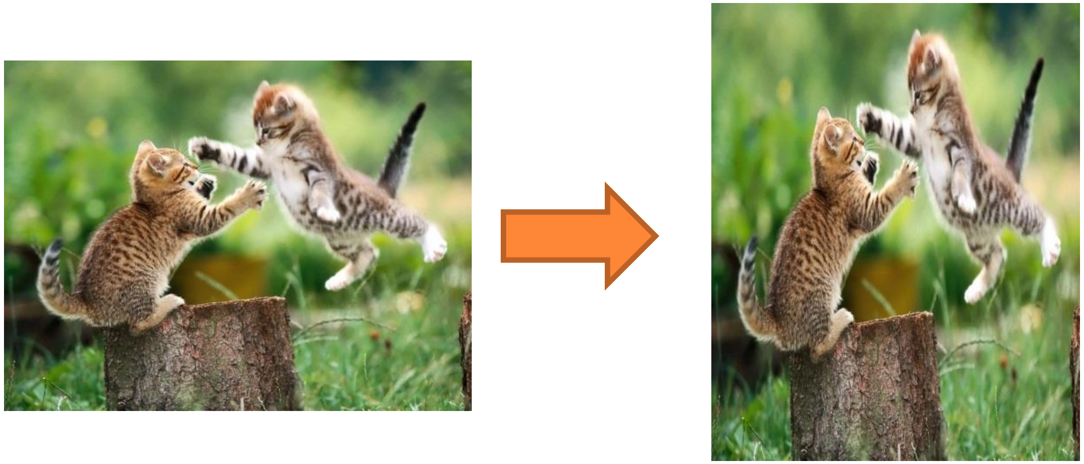
scaling rotation translation

Similarity transformations

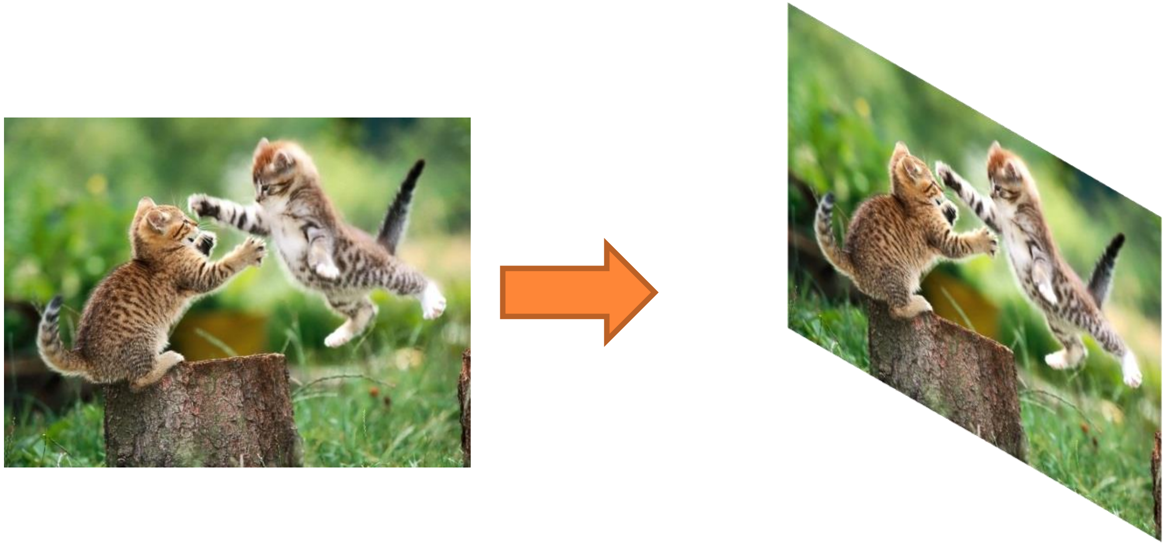


Similarity transform = translation + rotation + scale

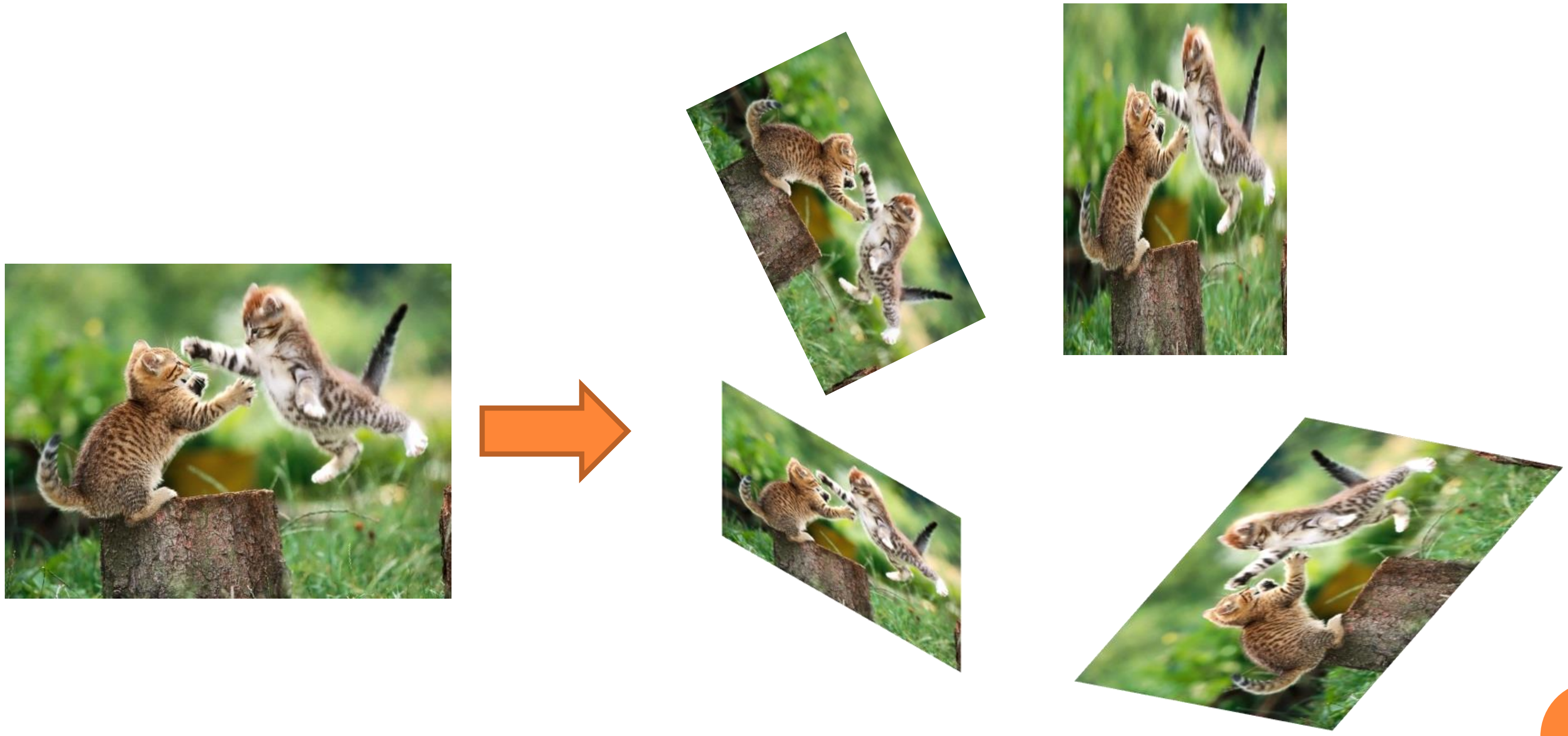
Aspect ratio



Shear



3) Affine Transform: An affine transform is a general linear transform in which straight lines remain straight lines after transformation but distances and angles may not be preserved, although ratios of distances between points are preserved.



Affine transform = translation + rotation + scale + aspect ratio + shear

Feature Extraction: Object Features

Feature extraction is the process of summarizing or converting raw image data into expressive presentations that are more informative or show better association with an underlying biological phenomenon.

•In order to compute object features, an object must first be determined. Given a segmentation step, the output is a set of foreground regions that are separated from the background.

•An algorithm like as connected components can be used to separate the foreground regions into individually labeled regions where each unique label indicates a unique object in the image.

•Given individually segmented objects, a feature is a number that describes some aspect of an object.

•There are a large number of features that are generally fall into the basic categories of shape, size, brightness, and texture.

•In constructing features, it is convenient to first compute the image moments, which are particular averages of either binary objects (unweighted) or their pixel intensities (weighted).

•They are useful to describe objects and form the building blocks of many useful features of the objects.

•For example, they can be used to compute a variety of shape features such as volume and centroid.

•They can also be used to compute the eigen-values and eigen-vectors of shapes, which can then be used to compute additional features such as eccentricity, elongation, and orientation.

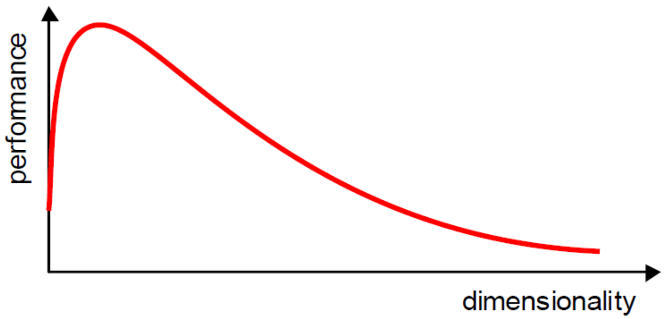
A good feature is

(1) discriminatory: significantly different for each class,
(2) information rich: has a small variance/spread within each class,
(3) easy to compute, and
(4) statistically independent: not correlated with another feature in use.

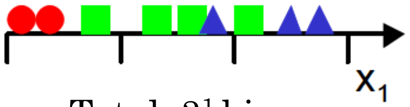
•The goal of feature selection is to reduce the complexity and dimensionality of the feature space in order to decrease the computational burden of the algorithms.

Curse of Dimensionality

- Increasing the number of features will not always improve classification accuracy.



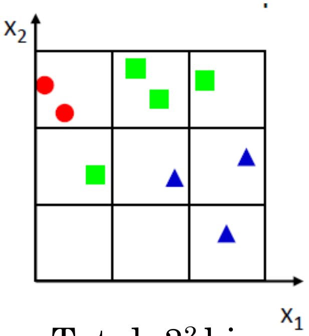
- In practice, the inclusion of more features might actually lead to worse performance.



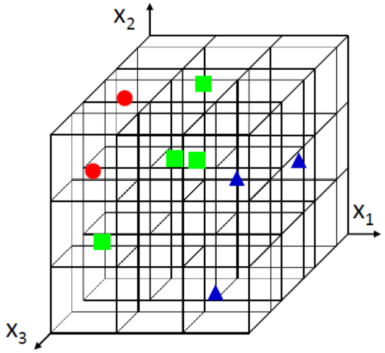
Total: 3^1 bins

k=3 bins per feature

- The number of training examples required increases exponentially with dimensionality **D** (i.e., k^D).



Total: 3^2 bins

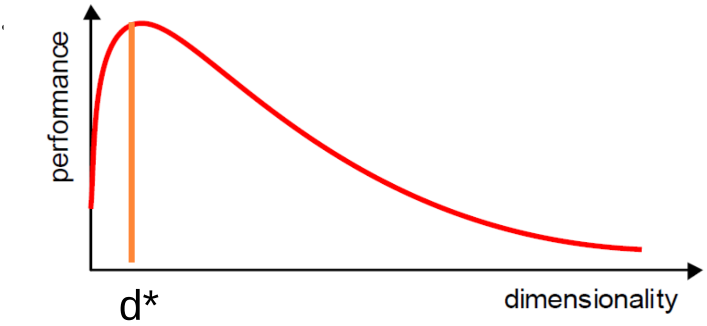


Total: 3^3 bins

k: number of bins per feature

Dimensionality Reduction

- What is the objective?
    - Choose an optimum set of features d* of lower dimensionalit uracy.



- Different methods can be used to reduce dimensionality:
    - Feature extraction
    - Feature selection

DIMENSIONALITY REDUCTION (CONT'D)

Feature extraction: computes a new set of features from the original features through some transformation f() .

f() could be linear or non-linear

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_D \end{bmatrix} \xrightarrow{f(\mathbf{x})} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_K \end{bmatrix}$$

K<<D

Feature selection: chooses a subset of the original features.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_D \end{bmatrix} \rightarrow \mathbf{y} = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ . \\ . \\ x_{i_K} \end{bmatrix}$$

K<<D

Feature Extraction

- Linear transformations are particularly attractive because they are simpler to compute and analytically tractable.

- Given x ϵ R^D, find an K x D matrix T such that:

  y = Tx ϵ R^K where K<<D

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_D \end{bmatrix} \overset{T}{\xrightarrow{f(\mathbf{x})}} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_K \end{bmatrix}$$

This is a projection transformation from D dimensions to K dimensions.

Each new feature y_i is a linear combination of the original features x_i

Feature Extraction (cont'd)

- From a mathematical point of view, finding an optimum mapping $\mathbf{y}=f(\mathbf{x})$ can be formulated as an optimization problem (i.e., minimize or maximize an **objective** criterion).

- Commonly used objective criteria:

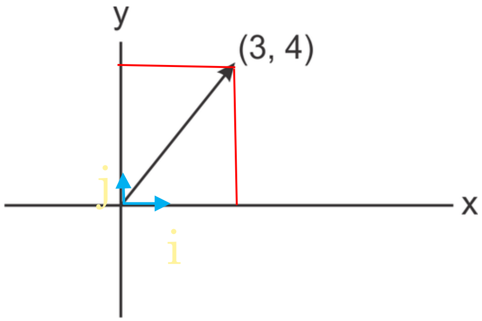
  - Minimize Information Loss: projection in the lower-dimensional space preserves as much **information** in the data as possible.

  - Maximize Discriminatory Information: projection in the lower-dimensional space increases **class separability**.

VECTOR REPRESENTATION (CONT'D)

- **Example** assuming D=2:

$$\mathbf{x}: \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$



- Assuming the standard base <v_1=i, v_2=j>, x_i can be obtained by projecting x along the direction of v_i:

$$x_1 = \mathbf{x}^T i = \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 3$$

$$x_2 = \mathbf{x}^T j = \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 4$$

- **x** can be "reconstructed" from its projection coefficients as follows:

$$\mathbf{x} = 3i + 4j$$

PCA - STEPS

- Suppose we are given $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ (D x 1) vectors

D: # of features

M: # data

Step 1: compute sample mean

Step 2: subtract sample mean (i.e., center the data at zero)

Step 3: compute the sample covariance matrix Σ_x

Step 4: compute the eigenvalues/eigenvectors of Σ_x

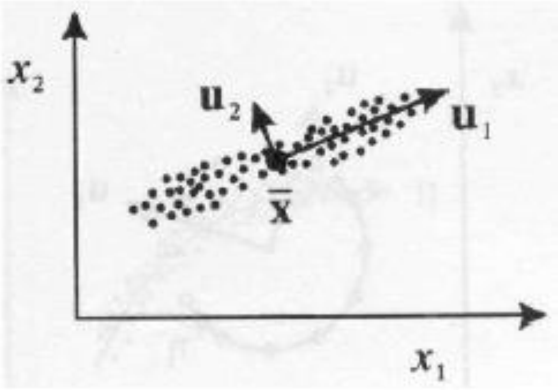
Step 5: dimensionality reduction step – approximate using only the first K eigenvectors (K<<D) (i.e., corresponding to the K largest eigenvalues where K is a parameter):

Note : most software packages return the eigenvalues (and corresponding eigenvectors) is decreasing order – if not, you should explicitly put them in this order)

Note : most software packages normalize u_i to unit length to simplify calculations; if not, you should explicitly normalize them)

INTERPRETATION OF PCA

- PCA chooses the eigenvectors corresponding to the largest eigenvalues.
- The eigenvalues correspond to the variance of the data along the eigenvector directions.
- Therefore, PCA projects the data along the directions where the data varies most.
- PCA preserves as much information in the data by preserving as much variance in the data.



u_1: direction of max variance
u_2: orthogonal to u_1

Mining of Sensor Data in Healthcare

- With progress in sensor technologies, the instrumentation of the world is offering unique opportunities to obtain fine grain data on patients and their environment.

- It not only facilitates design of sophisticated clinical decision support systems capable of better observing patients’ physiological signals and helps provide situational awareness to the bedside.

- This necessitates better monitoring and understanding of patients, their physiological signals, and their context.

- Recently, with advances in sensor and wearable technologies, several new data sources are available to provide insights on patients. For instance, Bluetooth enabled scales, blood pressure cuffs, heart rate monitors, and even portable electrocardiogram monitors are now available for early diagnosis.

- Several remote health monitoring solutions for chronic disease management, and wellness management have been proposed

- While rapid growth in healthcare sensor data offers significant promise to impact care delivery, it also introduces a data overload problem, for both systems and stakeholders

Mining Sensor Data in Medical Informatics: Scope

•Sensors measure physical attributes of the world and produce signals, i.e., time series consisting
•of ordered measurements of the form (timestamps, data elements).

•For example, in intensive care, respiration rates are estimated from measurements of the chest impedance of the patient. The resulting time series signals are consumed either by a human or by other sensors and computing systems.

•For instance, the output of the chest impedance sensor can be consumed by an apnea detection system to produce a signal measuring apnea episodes.

•The data elements produced by sensors range from simple scalar (numerical or categorical) values, to complex data structures.

•Examples of simple data elements include measures such as hourly average of temperature in a given geographical location, output by a temperature sensor.

•More complex data elements include summaries of vital signs and alerts measured by a patient monitor sensor in a medical institution.

Taxonomy of Sensors Used in Medical Informatics

we categorize sensors in medical informatics as follows:

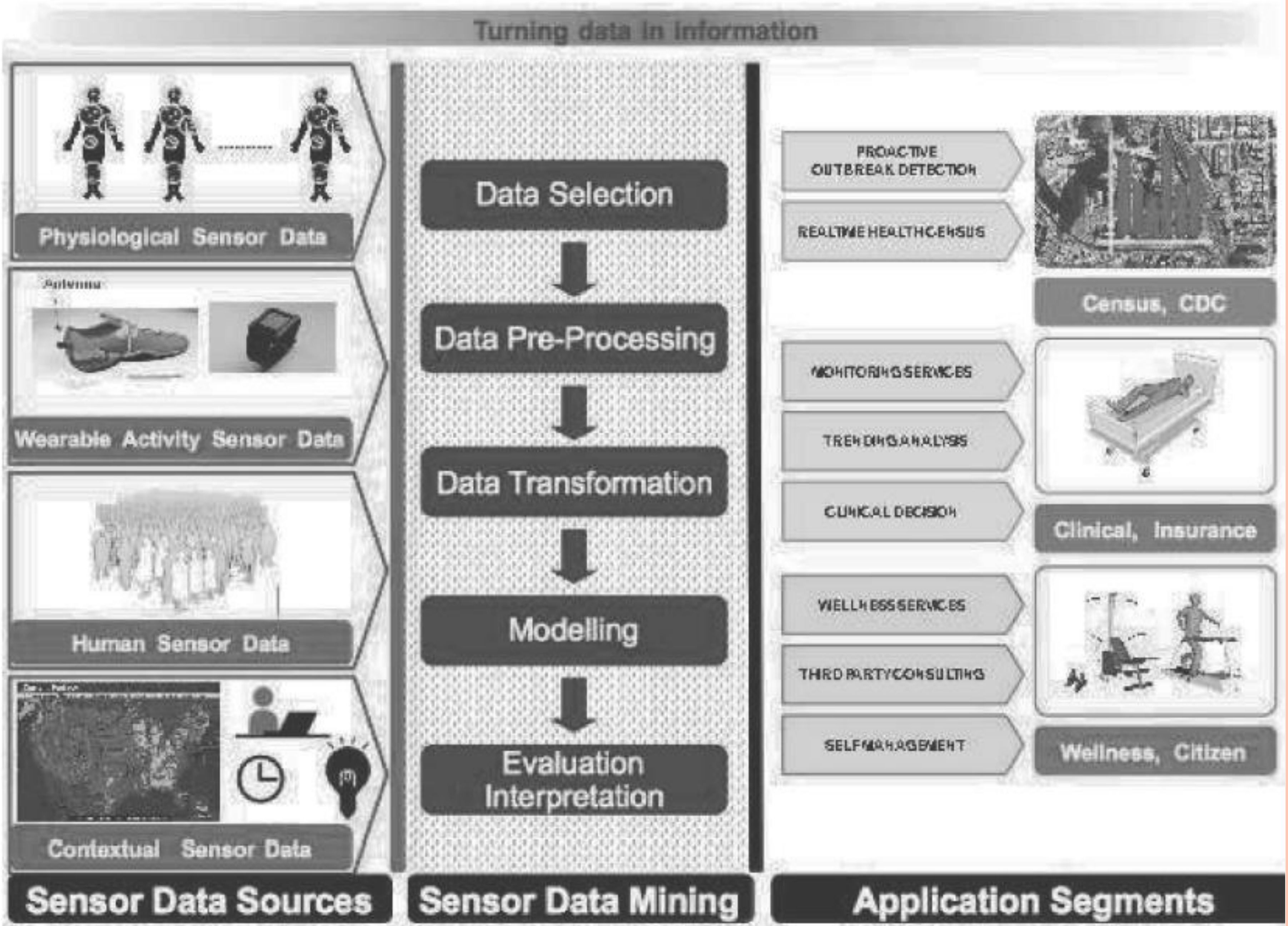
1) Physiological sensors: These sensors measure patient vital signs or physiological statistics. They were first used to measure vitals on astronauts before appearing in medical institutions.

2) Wearable activity sensors: These sensors measure attributes of gross user activity, different from narrowly focused vital sign sensors. Good example is Shoe manufacturers like Nike have enabled many of their running shoes with sensors
capable of tracking walking or jogging activities.

3) Human sensors: Humans play an integral role in the sensing process. For instance, physicians introduce important events that relate to the patient health status during examinations. Lab technicians follow rigorous processes to provide blood content information.

4) Contextual sensors: These sensors are embedded in the environment around the user to measure different contextual properties. Examples include motion detection sensors, audio and video sensors, temperature sensors, weather sensors, etc.

Taxonomy of Sensors Used in Medical Informatics



Challenges in Healthcare Data Analysis

•Despite several standardization efforts, medical sensor manufacturers tend to design proprietary data models and protocols to externalize sensed signals.

•In healthcare, standard bodies like HL7 and the Continua Health Alliance address data modeling issues while several IEEE standard protocols address device interoperability issues. However, there is a lack of incentives for sensor data manufacturers to adhere to these standards.

•With this lack of adherence to standards, mining medical sensor data across multiple data sources involves several nontrivial engineering challenges, and the design of custom solutions specific to each sensor data mining application.

•Another key challenge in the acquisition process is related to the protection of user privacy.

1. Acquisition Challenges

•There are 4 different classes of sensors that generate and collect healthcare relevant information. physiological sensors, contextual sensors and human sensors, activity sensors.

•Acquiring and integrating this data is nontrivial because of the inherent heterogeneity and lack of standards and protocols.

•Physiological sensor manufacturers have mostly designed proprietary data models and protocols to externalize sensed signals, despite the efforts of standard bodies like HL7 , IEEE standard etc.

•Additionally, there is little standardization or interoperability studies of contextual and activity sensors.

•Data from healthcare providers is captured poorly, often requiring manual entry and transcription.

•data aggregators operate only on a narrow set of sources, and often do not interoperate with each other.

•Hence, mining medical sensor data across multiple data sources has involved several nontrivial engineering challenges.

•These acquisition challenges are compounded by the need to provide privacy protection for this often very sensitive personal information

2 Pre-processing Challenges

•Data in the real world is inherently noisy. The preprocessing stage needs to address this problem with sophisticated data filtering, sampling, interpolation, and summarization techniques.

•The preprocessing also needs to account for the heterogeneity of data, and the lack of standards adoption by medical sensor manufacturers.

•Indeed, data generated in different formats needs to be syntactically aligned and synchronized before any analysis can take place.

•Clocks across sensors are often not synchronized, aligning the data across sensors can be quite challenging.

•Different types of structured and unstructured data.

•A semantic normalization is often required to cope with differences in the sensing process.

3 Transformation Challenges

•Data transformation involves taking the normalized and cleaned input data and converting it to a representation such that attributes or features relevant to the mining process can be extracted.

•This may include applying different types of linear (e.g., Fourier Transform, Wavelet Transform) and nonlinear transformations to numeric data.

•Converting unstructured data such as text and images into numeric representations (e.g., using a bag of words representations, or extracting color, shape, and texture properties), and applying dimensionality reduction and de-correlation techniques

•Summarizing the result with a set of representative features that can then be used for analysis and modeling.

•The choice of the appropriate transformations and representations for the features is heavily dependent on the task that needs to be performed.

•For instance a different set of features may be required for an anomaly detection task, as opposed to a clustering or classification task.

•.

- Additionally, the choice of appropriate features requires understanding of the healthcare problem at hand (e.g., the underlying physiology of the patient) and often requires inputs from domain experts.

- In addition, Human sensing adds different types of unstructured data that need to be effectively integrated.

This includes textual reports from examinations (by physicians or nurses) that need to be transformed into relevant features, and aligned with the rest of the physiological measurements

4 Modeling Challenges

•The time series nature of the data often requires the application of sequential mining algorithms that are often more complex than conventional machine learning techniques.

•Non-stationarities in time series data necessitate the use of modeling techniques that can capture the
•dynamic nature of the state of the underlying processes that generate the data.

•Known techniques for such problems, including discrete state estimation approaches (e.g., dynamic Bayesian networks
•and hidden Markov models) and continuous state estimation approaches (e.g.recurrent neural networks) have been used only in limited settings.

•Another challenge arises due to the inherent distributed nature of these applications.

•In many cases, communication and computational costs, as well as sharing restrictions for patient privacy prevent the aggregation of the data in a central repository. As a result, the modeling stage needs to use complex distributed mining algorithms.

•In remote settings, there is limited control on the data acquisition at the sensor. Sensors may be disconnected for privacy reasons or for resource management reasons (e.g., power constraints), thereby affecting the data available for analysis.

•Optimizing the modeling process becomes a challenging distributed data mining problem that has received only limited attention in the data mining community.

•Modeling in healthcare mining is also hindered by the ability to obtain ground truth on the data.

•Labels are often imprecise and noisy in the medical setting. For instance, a supervised learning approach for the early detection of a chronic disease requires well-labeled training data.

•However, domain experts do not always know exactly when a disease has started to manifest itself in a body,and can only approximate this time.

•Additionally, there are instances of misdiagnosis that can lead to incorrect or noisy labels that can degrade the quality of any predictive models.

•In clinical settings, physicians do not have the luxury of being able to try different treatment options on their patients for exploration purposes.

•As a result, historical data sets used in the mining process tend to be quite sparse and include natural biases driven by the way care was delivered to the patient.

5 Evaluation and Interpretation Challenges

•Data mining results consist of models and predictions that need to be interpreted by domain experts.

•Many modeling techniques produce models that are not easily interpretable.

•For example, the weights of a neural network may be difficult to grasp for a domain expert. But for such a model
•to be adopted for clinical use, it needs to be validated with existing medical knowledge.

•It becomes imperative to track metadata describing the process used to derive any results from data mining to help domain expert interpret these results.

•Furthermore, the provenance of the data sets, and analysis decisions used during the modeling are also required by the experts to evaluate the validity of the results.

•This imposes several additional requirements on the selected models and analysis.

6 Generic Systems Challenges

•Beyond analytical challenges, sensor data mining also comes with a set of systems challenges that apply to medical informatics applications.

•The mining of sensor data typically requires more than conventional data management

reasons:

- The temporal aspect of the data produced by sensors sometimes generate large amounts of data that can overwhelm a relational database system. For example, a large population monitoring solution requiring the real-time analysis of physiological readings.
- Sensor mining applications often have real-time requirements.
- The unstructured nature of some of the data produced by sensors coupled with the real-time requirements imposes requirements on the programming and analysis models used by developers of sensor data mining applications.

•Hence, sensor mining in healthcare requires the use of emerging stream processing system technology in conjunctionwith database and data warehousing technologies.

•They are also time sensitive and analyze data within small latency bounds.

Biomedical Signal Analysis

- Biomedical Signal Analysis consists of measuring signals from biological sources, the origin of which lies in various physiological processes.

- These signals based on their origin are classified into different types e.g., physiological signals originating because of electrical activity in the heart are called electrocardiogram (ECG), while those originating because of electrical activity in the brain are called electroencephalogram (EEG).

- Biological signals manifest themselves into different forms such as electrical, acoustic, chemical, and many others.

- The analysis of these signals is vital in diagnosing the pathological conditions and in deciding an appropriate care pathway insights on patients.

- Many times the underlying pathological processes result in different signatures and a good understanding Biomedical Signal Analysis of the physiological system is necessary to understand the status of the system.

- For instance, a rise in the temperature of the human body can convey infections in the body. Sometimes it can be a consequence of a blood clot, which is good if it helps in stopping the bleeding but carries a risk of heart attack or stroke.

Types of Biomedical Signals

- Types of biomedical signals, their origins and importance for diagnosis purpose

- The most basic form of measurement is body temperature, which although quite simplistic to measure can convey the well-being of the human system.

- This section looks into the signals originating from the cellular level, such as the action potential, to the macro level, for instance the heart sound, which is produced as a consequence of contractile activity of the cardiohemic system.

1) Action Potential:

2) Electroneurogram (ENG):

3) Electromyogram (EMG) :

.

4) Electrocardiogram (ECG) :

5) Electroencephalogram (EEG):

6) Electrogastrogram (EGG):

7) Phonocardiogram (PCG):

8) Other Biomedical Signals :

ECG Signal Analysis

- The recorded PQRST complex of the ECG waveform contains substantial information and certain features such as cardiac rate, rhythm, PR interval, QRS duration, ST segment, QT interval, and T waves indicate the underlying pathological condition of the patient.

- However, the ECG waveform is corrupted with several sources of noise and before any feature could be extracted, a proper signal conditioning is necessary.

- Various kinds of noise that affect the ECG signals are:

(a) Power line interference
(b) Electrode contact noise
(c) Motion artifacts
(d) Muscle contraction (electromyographic, EMG)
(e) Baseline drift and ECG amplitude modulation with respiration
(f) Instrumentation noise generated during signal acquisition
(g) Electrosurgical noises, and many other less significant noises

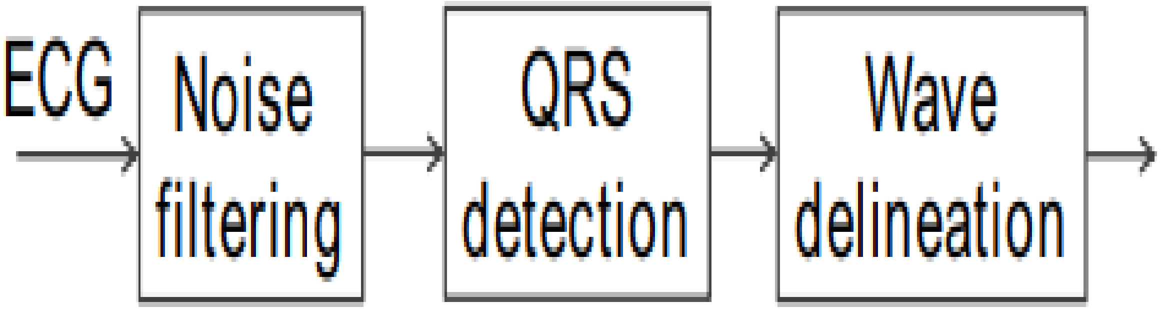


FIGURE : Algorithm for processing ECG signal

- The heart rate is estimated from successive QRS beats.
- Other features, for instance T-wave alternans, are estimated once the waves are delineated.

- Common to all kinds of means by which an ECG signal is recorded, whether in an ambulatory or resting state or during a stress test, is the processing of an ECG signal.

- Figure shows a frequently used signal processing routine deployed on ECG machines to minimize the interference due to the above-mentioned sources of noise.

- Signal processing has contributed immensely in deciphering information from an ECG signal and has substantially improved our understanding of the ECG signal and its dynamic properties as expressed by changes in rhythm and beat morphology (PQRST complex).

- For instance, detection of alternating changes in a T wave from one PQRST complex to another in the form of oscillations, an indicator of life-threatening arrhythmias cannot be perceived by the naked eye or from a standard ECG printout, but needs careful signal processing to unmask the information buried in noise.

- While designing signal processing algorithms for reducing noises in the measurement, it is important to note that an electrocardiograph should meet or exceed the requirement of IEC 60601-2-51 (2003).

- The ECG measuring devices should be programmed in accordance with American Heart Association (AHA) specifications.

- For instance, according to the guidelines, the low frequency filter should be set no higher than 0.05 Hz to avoid distortion of the ST segment and the high frequency filter should be set no lower than 100 Hz to prevent loss of high frequency information

- The are various signal processing approaches applied to remove the noises affecting the ECG measurement and also the approaches commonly used for extracting certain morphological features from ECG, such as QRS detection, QT interval.

Genomic Data Analysis for Personalized Medicine.

•Empowered by newly emerging biotechnologies and hence the fast generation of biological and medical information, advanced genomic research promises the whole field unprecedented opportunities and hopes for genome scale study of challenging problems in life science.

• For example, advances in genomic technology made it possible to study the complete genomic landscapes of healthy individuals or of any complex diseases,

•the genome-wide responses to certain genetic and chemical perturbations or drug treatment, and the large-scale molecular changes that are associated to various disease phenotypes.

•Many of such research efforts have proven to be highly promising to generate new insights into the biology of human disease and to predict the individual response to treatment, which therefore could enhances our understanding of the underlying mechanisms, promote the knowledge exchange between doctors and patients, and facilitate clinical decision making

1. Genomic Data Generation

- There are 3 different classes of data generation:

Microarray Data Era
Next-Generation Sequencing Era
Public Repositories for Genomic Data

Methods and Standards for Genomic Data Analysis

- Empowered by newly emerging biotechnologies and hence the fast generation of biological and medical information, advanced genomic research promises the whole field unprecedented opportunities and hopes for genome scale study of challenging problems in life science.

- For example, advances in genomic technology made it possible to study-

- the complete genomic landscapes of healthy individuals or of any complex diseases,

- the genome-wide responses to certain genetic and chemical perturbations or drug treatment,
- and the large-scale molecular changes that are associated to various disease phenotypes.

- Many of such research efforts have proven to be highly promising to generate new insights into the biology of human disease and to predict the individual response to treatment, which therefore could enhances our understanding of the underlying mechanisms, promote the knowledge exchange between doctors and patients, and facilitate clinical decision making

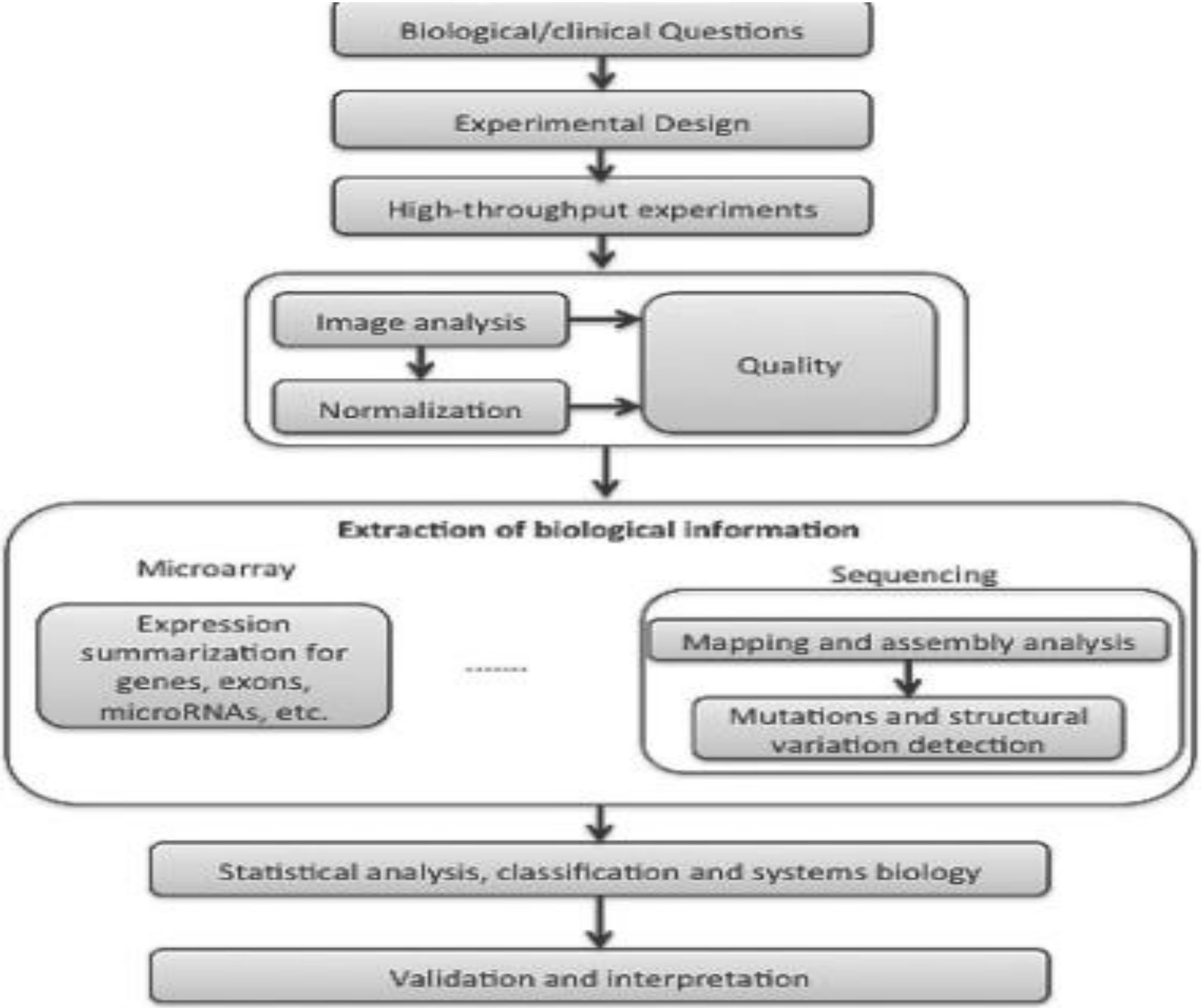

Biological/clinical Questions
Experimental Design
High-throughput experiments
Image analysis
Normalization
Quality
Extraction of biological information
Microarray
Expression summarization for genes, exons, microRNAs, etc.
....
Sequencing
Mapping and assembly analysis
Mutations and structural variation detection
Statistical analysis, classification and systems biology
Validation and interpretation

FIGURE 6.3: The standard bioinformatics workflow to analyze the genomic data.