Computing Methods for Experimental Physics and Data Analysis

Data Analysis in Medical Physics

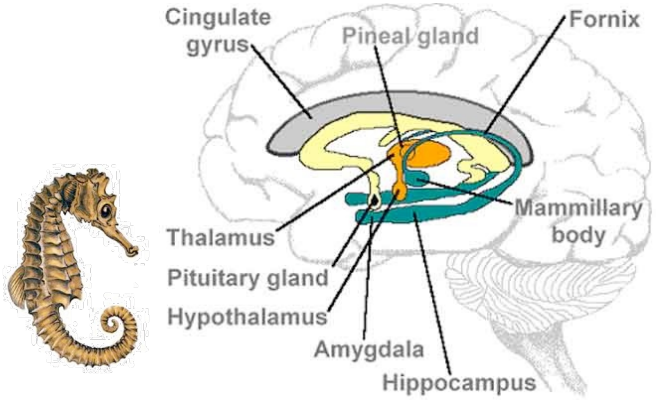
Lecture 5: Image segmentation and Radiomics

Alessandra Retico
alessandra.retico@pi.infn.it
INFN - Pisa

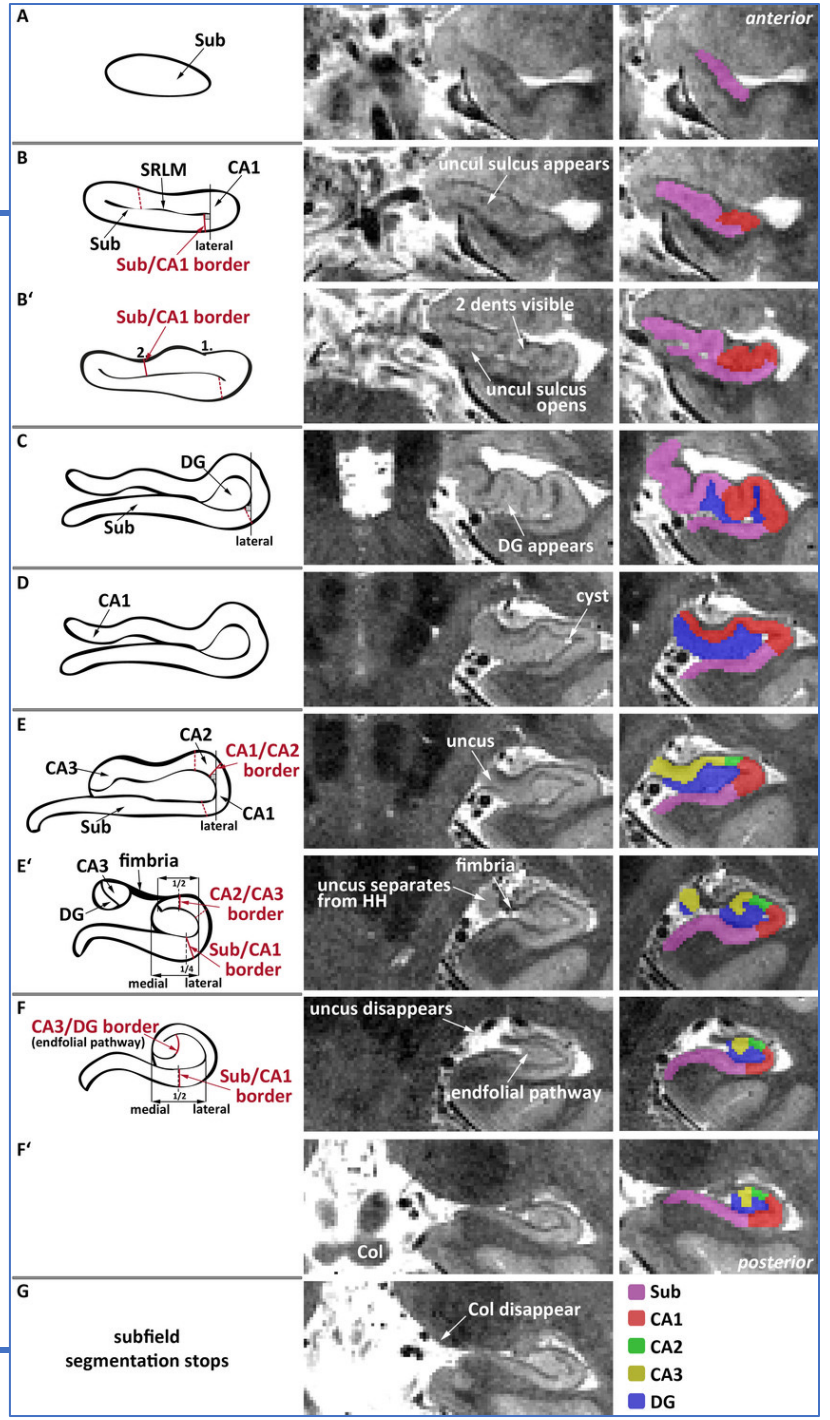
Image segmentation

- Image segmentation: the image is partitioned into its parts or regions (segments)
- Segmentation of tissues, organs, lesions or other **regions of interest (ROIs)** out of a medical image is often a useful step to extract meaningful information related to shape or texture of the object of interest

For example, in the study of Alzheimer's Disease (AD), the hippocampus is one of the first regions of the brain to suffer damage; memory loss and disorientation are included among the early AD symptoms.



D. Berron et al., A protocol for manual segmentation of medial temporal lobe subregions in 7 Tesla MRI, NeuroImage: Clinical 15(C) 2017



How to segment regions of interest (ROI) in medical images

- **Manual segmentation**: ROIs are manually drawn by experts
  - Extremely time consuming, but necessary step to build the ground truth and/or templates for segmentation algorithm

- **Template matching:** Images can be warped to specific template images, which contain suitable ROI definitions
  - Applicable in processing almost-normal anatomy; not working in case of subject-specific alterations

- **Custom algorithms** can be set up to extract ROIs (tissues, organs, lesions), relying on image properties (e.g. object intensities, discontinuities in pixel values)

Some segmentation methods (and corresponding MATLAB functions)

Thresholding-based methods:

- Global/Multilevel image thresholding using Otsu's method (imbinarize, graythresh, multithresh)

Clustering-based methods:

- K-means clustering based image segmentation (imsegkmeans)

Use only histogram information

Edge-detection-based methods:

- finding edges in the input image (edge), then fill the region inside (imfill)
- Active contours (activecontour)

Region growing methods:

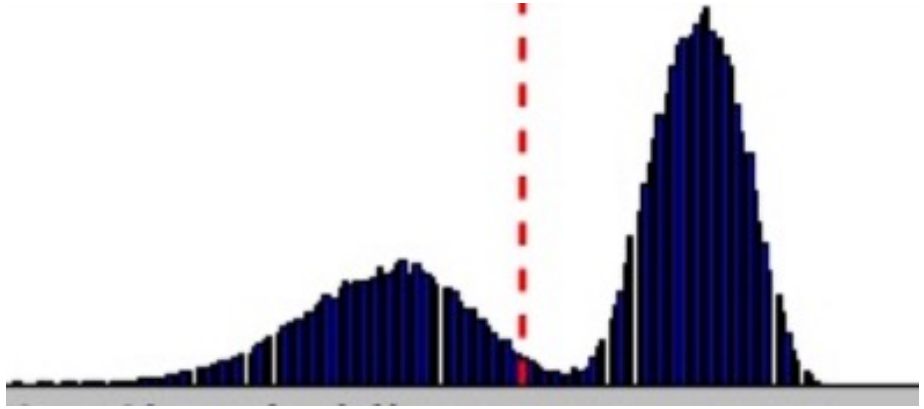
- Select contiguous image regions with similar gray values (grayconnected)

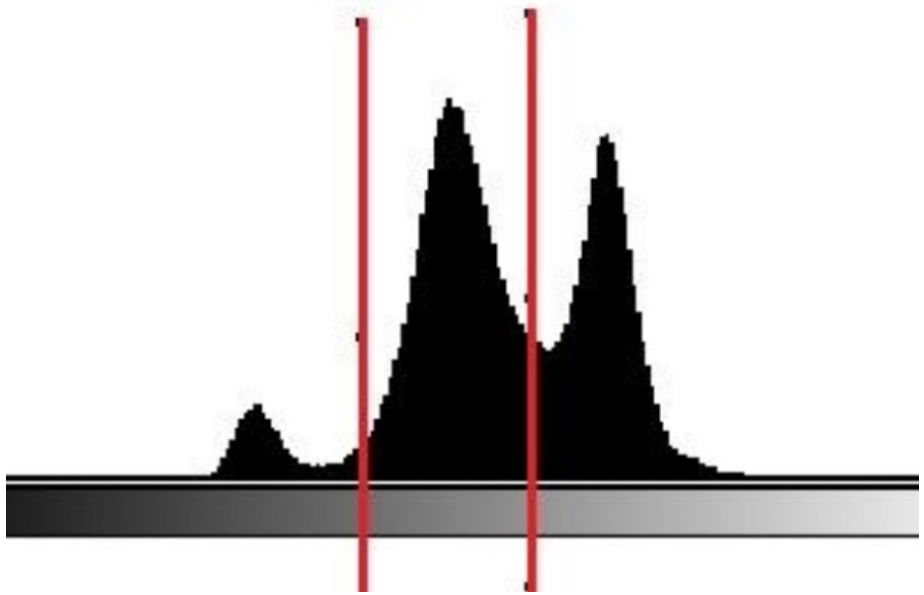
Exploit spatial relations between pixels

… → Machine-Learning and Deep-Learning based methods

Global and Multilevel image thresholds using Otsu's method

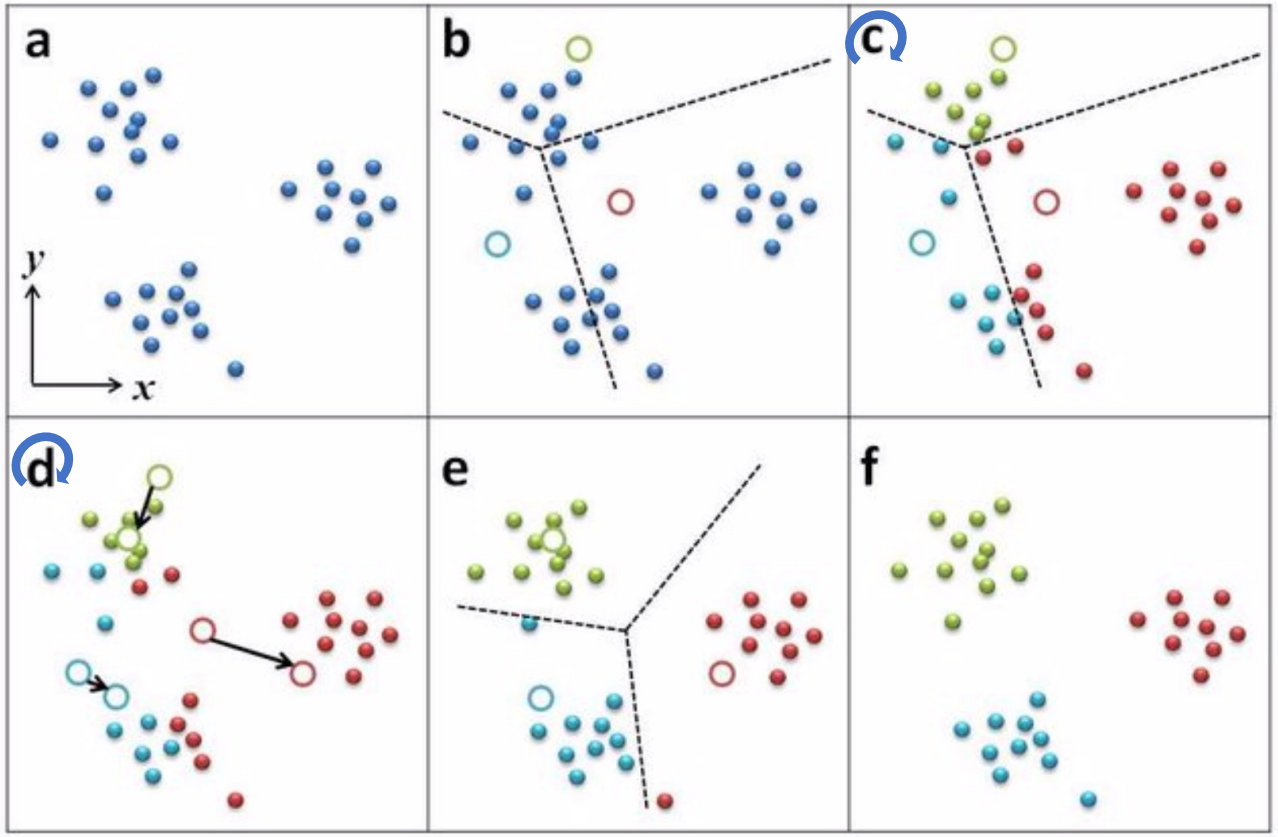
- The Otsu algorithm, in the simplest form, performs automatic definition of a single intensity threshold that separate pixels into two classes, foreground and background.
- This threshold is determined by minimizing intra-class intensity variance, or equivalently, by maximizing inter-class variance.
- In the multilevel form, the N-1 thresholds partitioning N different image levels are provided.





Clustering methods: the *K*-means algorithm

A schematic illustration of the K-means algorithm for two-dimensional data clustering



Chen and Lai, https://arxiv.org/abs/1611.01849

(a) The data points (solid blue circles) to be clustered in a 2D feature space. There are three clusters, so K = 3. The K-means algorithm is capable of assigning each data point into one of the clusters.

(b) For random locations of the cluster centers (empty circles), each data point can be associated with the closest center.

(c) The 2D space is divided into three regions through three decision boundaries, each containing the corresponding data points whose closest center is within. The data points currently in the regions with the aqua, green, and red centers are assigned to the corresponding clusters, respectively.

(d) Each center moves to the centroid of the data points currently assigned to it (movements shown by the black arrows).

(e) The updated cluster assignments of the data points are obtained according to the new center locations.

The steps in (c) and (d) are repeated until convergence is achieved.

(f) The final cluster assignments

Clustering methods: the *K*-means algorithm

- The image is iteratively partitioned in *K* clusters (distinct, non-overlapping subgroups of pixels).
- The objective is to make the inter-cluster elements as similar as possible while also keeping the clusters as different (far) as possible

Algorithm steps:

- Definition of the number of clusters *K*
- Initialize centroids randomly selecting *K* data points for the centroids
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the data points that belong to each cluster.

Expectation-Maximization.
The **E**-step is assigning the data points to the closest cluster.
The **M**-step is computing the centroid of each cluster.

Objective function

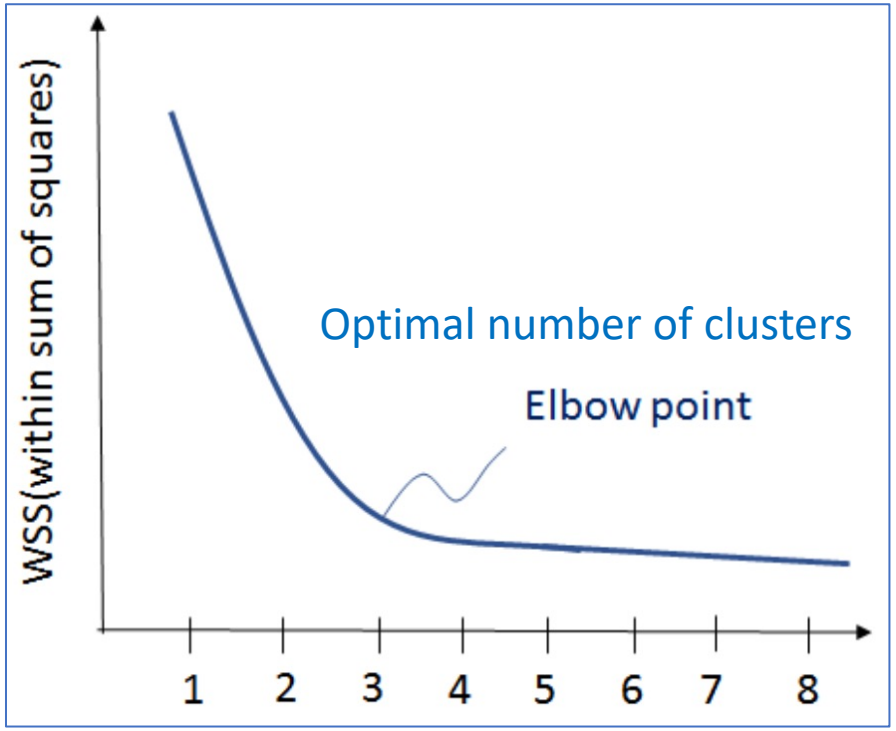
$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

K is the number of clusters;
w_{ik} is 1 for points in the k^{th} cluster, 0 outside;
m is the total amount of data points.

The optimization consists in two steps, i.e. minimizing *J* first with respect to w_{ik} and keep μ_k fixed. Then, minimizing *J* with respect to μ_k and keep w_{ik} fixed.

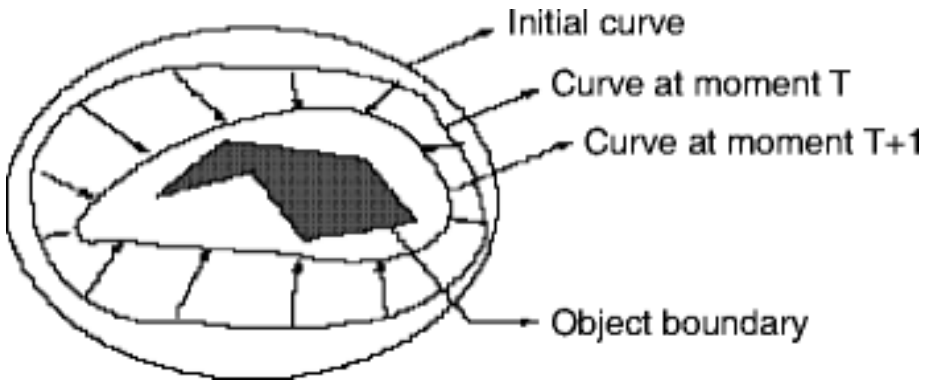
Clustering methods: the *K*-means algorithm

- A critical issue of *K*-means is the definition of the optimal number of clusters *K*
- The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation, or total **within-cluster sum of square (WSS**), is minimized.
- WSS measures the compactness of the clustering and we want it to be as small as possible
- Optimization criteria can be defined, e.g. through:
  - the Elbow method, where the k-means is run for K=[1, …10] and WSS is computed. The Elbow point in the curve is considered as the optimal number of clusters.



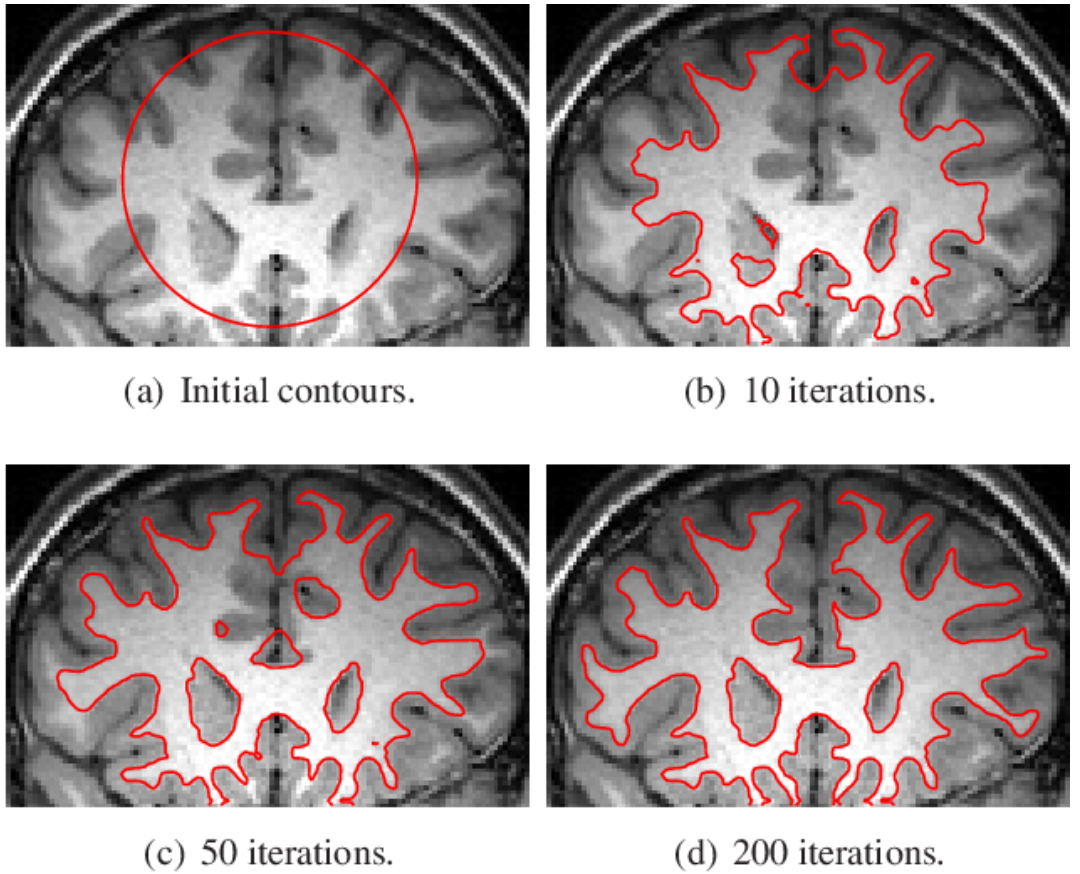
Active contour models/snakes

- In active contour models a contour algorithm, also called *snake*, moves to find object boundaries.
- The *snake* is an energy minimizing spline* guided by external constraint forces and influenced by image forces that pull it toward features such as lines and edges.



- It locks onto nearby edges and localizes them accurately

* A spline is a function, consisting of a set of polynomials connected together, whose purpose is to interpolate in a range a set of points, so that the function is continuous at least up to a given order of derivatives at any point in the interval.



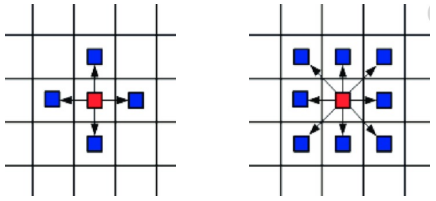
(a) Initial contours. (b) 10 iterations.

(c) 50 iterations. (d) 200 iterations.

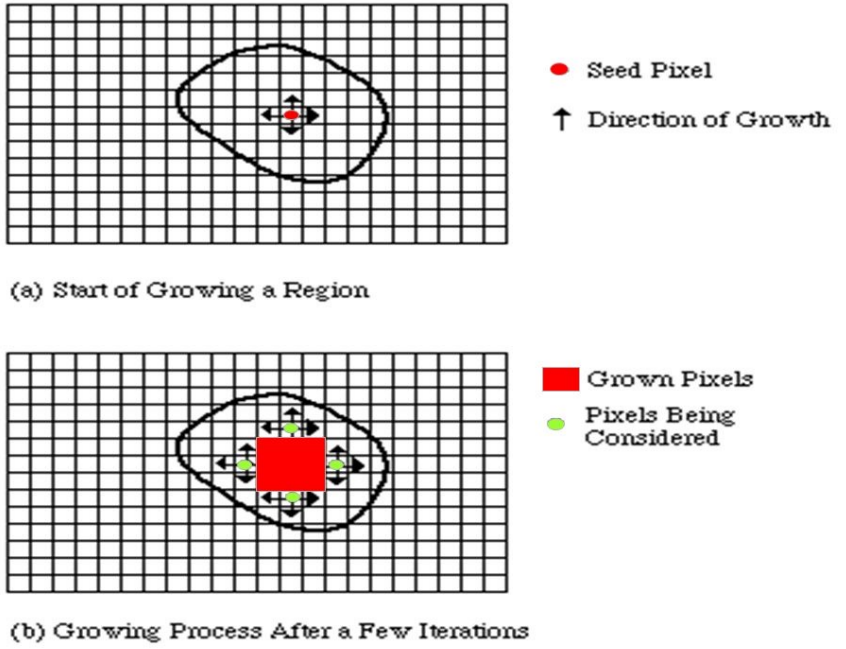
Chunming Li, Chiu-Yen Kao, John C. Gore, Zhaohua Ding
Implicit Active Contours Driven by Local Binary Fitting Energy, 2007
IEEE Conference on Computer Vision and Pattern Recognition

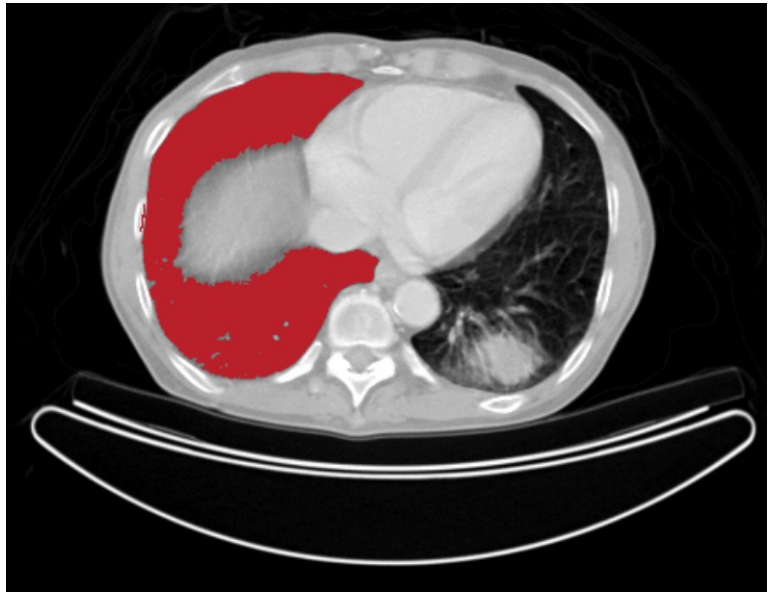
Region growing methods

- It requires the selection of an initial seed point.
- Region is ***grown*** from the seed pixel by adding in neighboring pixels that are *similar*, increasing the size of the region.
- When the growth of one region stops, another seed pixel is chosen, which does not yet belong to any region and the procedure starts again.
- *Similarity conditions* for pixels of the same regions have to be defined, e.g. pixel intensity should be in a certain range, or pixel intensity variance in a neighborhood should be in a certain range (to define the pixel neighborhood a connectivity measure should be defined, e.g. 4- or 8-connectivity in 2D, etc.)



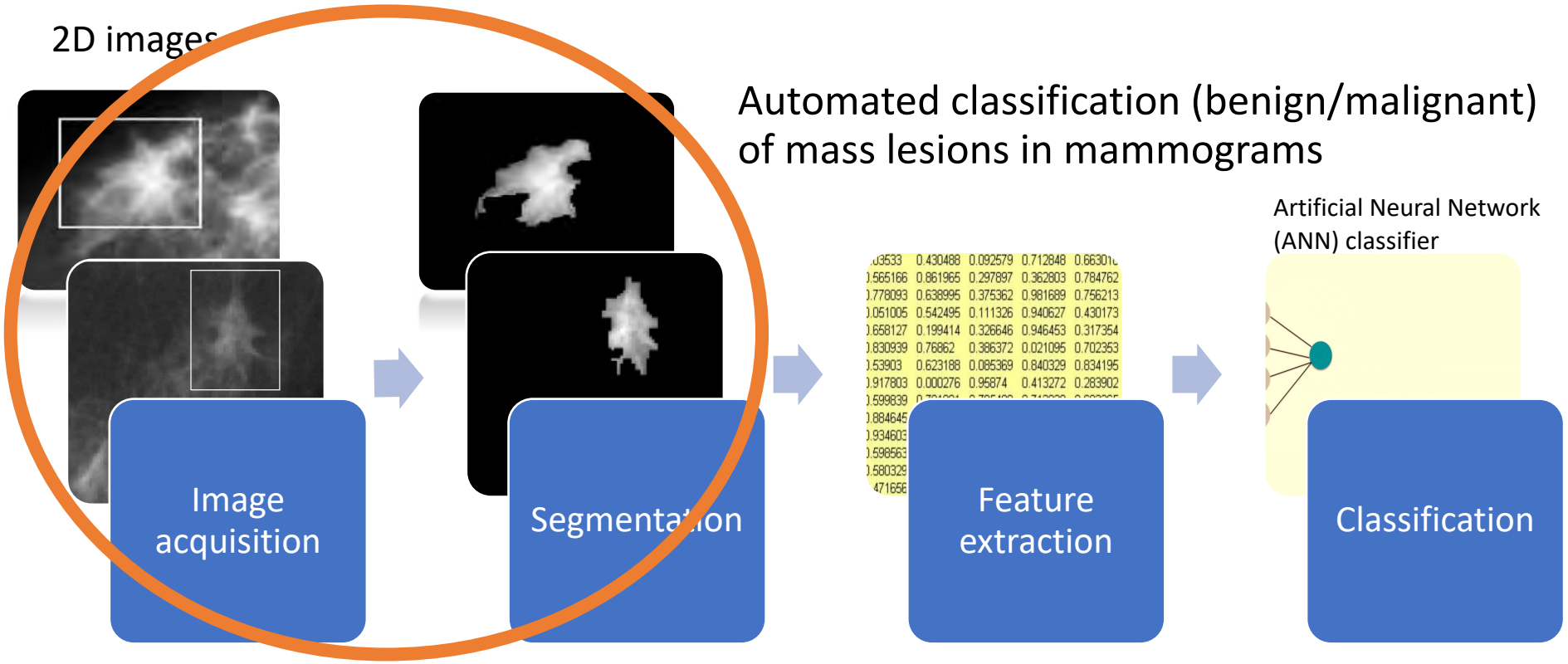
4-connectivity 8-connectivity





Typical image analysis pipeline for assisted diagnosis

Example: 1) Object segmentation; 2) Hand-crafted feature extraction; 3) Machine Learning classification



[P Delogu, ME Fantacci, P Kasae, A Retico, Computers in Biology and Medicine 37 (2007) 1479 – 1491]

See demo code: L5_code mass_segmentation; Data available on https://pandora.infn.it/public/cmepda/DATASETS and on https://drive.google.com/drive/folders/1YqK7ZkM-P2IrqfD7Pj-SCmjz-GWd_1-Y

Image types in MATLAB (Image Processing Toolbox)

- Binary images: *m*-by-*n* logical array. Array values of 0 and 1 are interpreted as black and white.
- Indexed images: *m*-by-*n* numeric matrix whose elements are direct indices into a color map. Each row of the color map specifies the red, green, and blue components of a single color.
- Greyscale images (intensity images): *m*-by-*n* numeric array whose elements specify intensity values.
- Truecolor images (RGB images): *m*-by-*n*-by-3 numeric array whose elements specify the intensity values of one of the three color channels. For RGB images, the three channels represent the red, green, and blue signals of the image

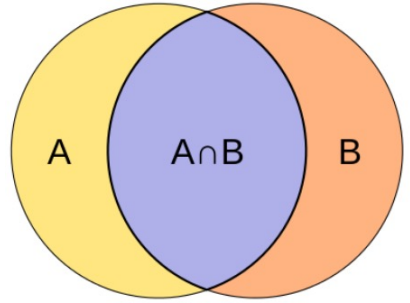
Multidimensional arrays

- The MATLAB `cat` function can be used to concatenate images:
  - 2D images can be concatenated in the 3rd dimension to create a 3-D array of size *m*-by-*n*-by-*p*. Each of the *p* images has size *m*-by-*n*, `cat(3,Im1,Im2)`
  - 3D images can be concatenated in the 4rd dimension, `cat(4,Vol1,Vol2)`
  - a sequence of 2D RGB images, then concatenate the images along the fourth dimension to create a 4-D array of size *m*-by-*n*-by-3-by-*p*. Each of the *p* images has size *m*-by-*n*-by-3, `cat(4,Im1,Im2)`
  - …

Segmentation similarity measures

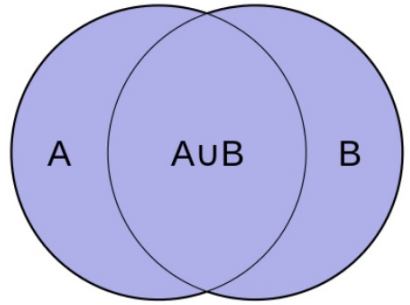
- **Jaccard similarity coefficient: Intersection over Union**

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad 0 \leq J(A,B) \leq 1.$$



- **Sørensen-Dice similarity coefficient**

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$



where $|X|$ and $|Y|$ are the cardinalities of the two sets (i.e. the number of elements in each set).
The Sørensen index equals twice the number of elements common to both sets divided by the sum of the number of elements in each set.
When applied to boolean data, using the definition of true positive (TP), false positive (FP), and false negative (FN), it can be written as

$$DSC = \frac{2TP}{2TP + FP + FN}.$$

It is different from Jaccard index which only counts true positives once in both the numerator and denominator.

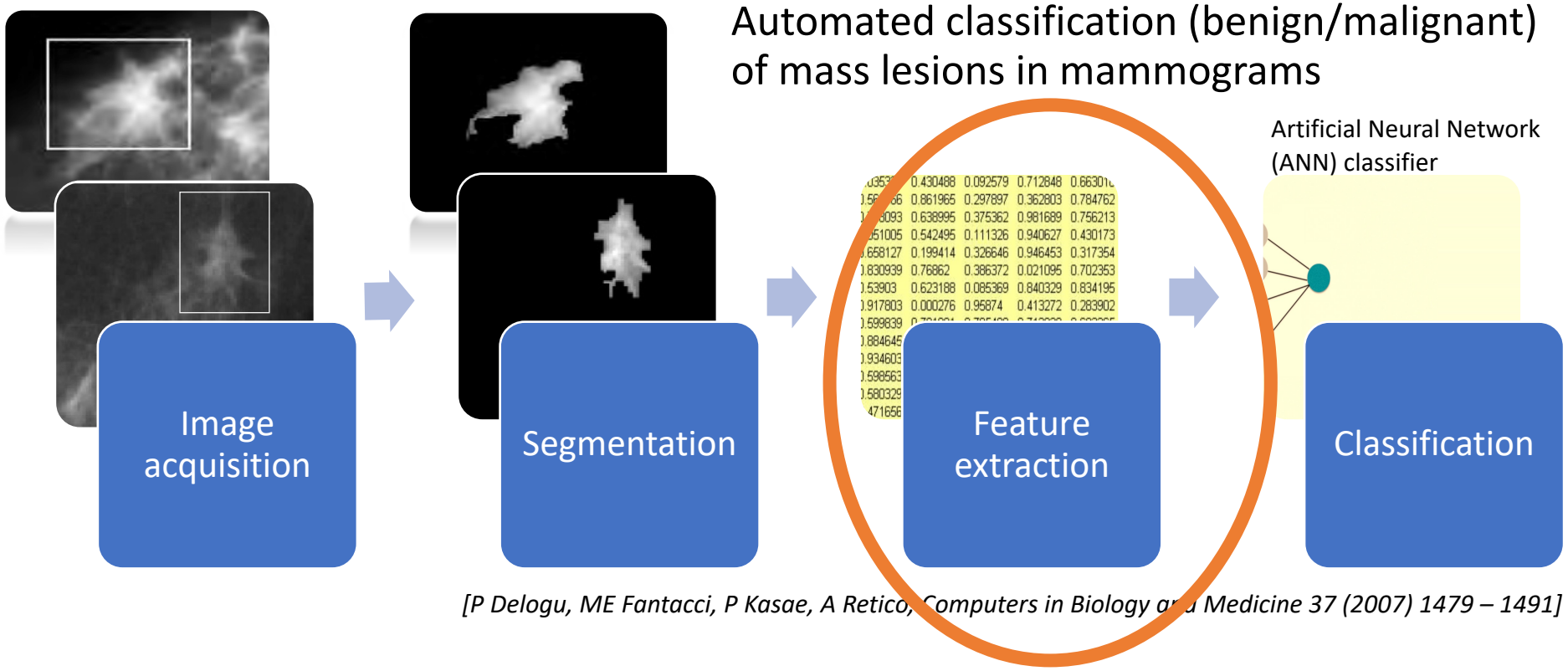
Assignment: mass_segment.m

- Start from script_mass_segment.m (https://github.com/retico/cmepda_medphys/tree/master/L5_code)
- It contains the main steps of the segmentation algorithm developed in the paper by *P Delogu, ME Fantacci, P Kasae, A Retico, Computers in Biology and Medicine 37 (2007) 1479 – 1491*
- It calls the two custom functions draw_radial_lines.m and max_var_points_interp.m
- Try how the script_mass_segment.m works on more than one benign/malignant mass of the dataset available on https://pandora.infn.it/public/cmepda/DATASETS or on https://drive.google.com/drive/folders/1YqK7ZkM-P2IrqfD7Pj-SCmjz-GWd_1-Y (malignant masses are named as xxxx_1.png, whereas benign ones as xxxx_2.png)
- Modify the script to obtain a function (following the instructions at the end of the script) and obtain the mass_segment.m. Set it up so to pass all free parameters to the function to enable an optimization study.
- Use the mass_segment.m function in a new script following the instructions provided in Lecture5_exercise1.pdf with the aim of testing the reproducibility of the segmentation algorithm
- (I have also shared a run segmentation script (run_segmentation.m), which allows to run the mass_segment function on *.png files of a directory. It implements also the possibility to repeat the segmentation if the result is not satisfactory.)

Typical image analysis pipeline for assisted diagnosis

Example: 1) Object segmentation; 2) Hand-crafted feature extraction; 3) Machine Learning classification

2D images



[P Delogu, ME Fantacci, P Kasae, A Retico, Computers in Biology and Medicine 37 (2007) 1479 – 1491]

See demo code: Lecture5_demo1_extract_features.m

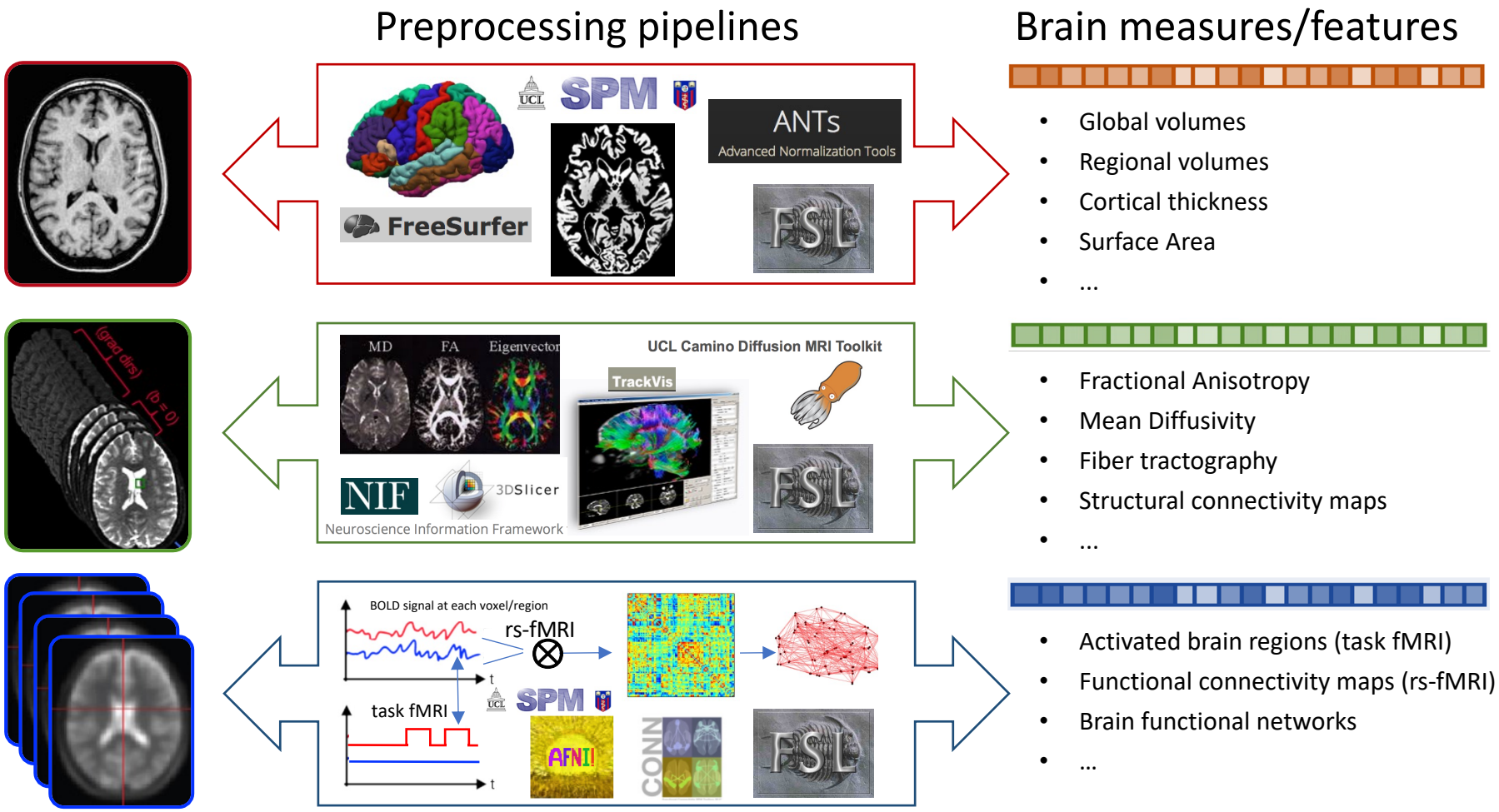
Image/ROI features

Several descriptive features of segmented ROIs are provided by the regionprops MATLAB function, which can return:

- Shape Measurements:
  - 'Area','Centroid', and 'BoundingBox'
  - 'ConvexArea', 'ConvexHull','ConvexImage', 'Circularity', 'EulerNumber', 'Filled Area', 'FilledImage', 'MaxFeretProperties', 'MinFeretProperties'and 'Solidity'
- Pixel Value Measurements:
  - 'MaxIntensity', 'MeanIntensity', 'MinIntensity', 'PixelValues', 'WeightedCentroid'

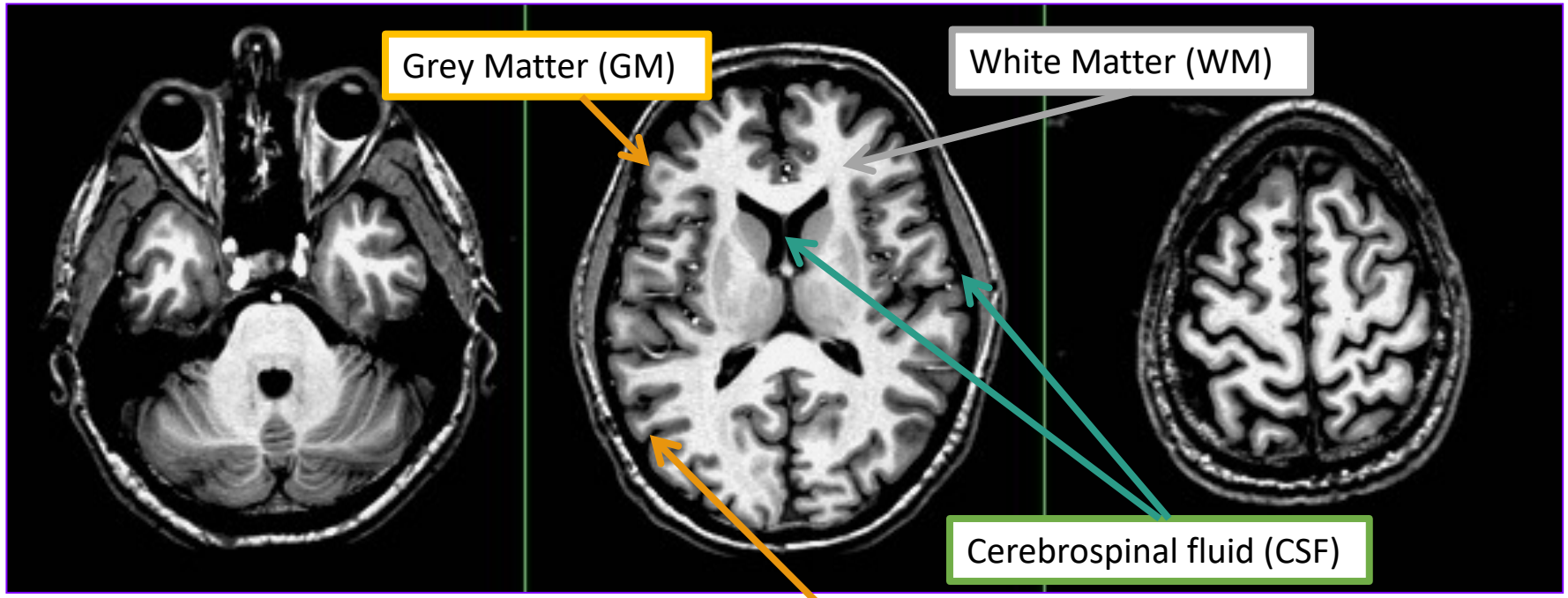
The function regionprops3 provides analogous features for 3D ROIs

Feature extraction from brain images (few examples)



MRI T_1-weighted brain images

- Axial slices of a human head with spatial resolution of 1 mm^3



Grey Matter (GM) cortex can by followed and cortical thickness can be evaluated to investigate GM involvement in pathological conditions.

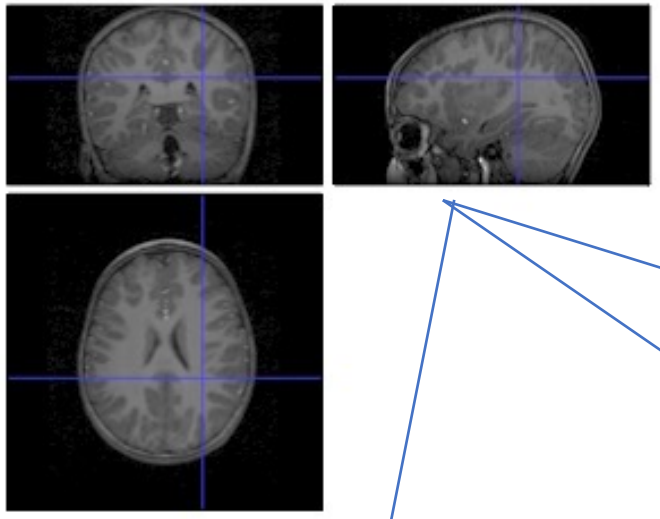
The need for tissue segmentation in MRI

- MRI intensity is usually not quantitatively meaningful (as opposed to e.g. computed tomography images)

- Regional volumes of the three main tissue types: gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), are well-defined and potentially very interesting

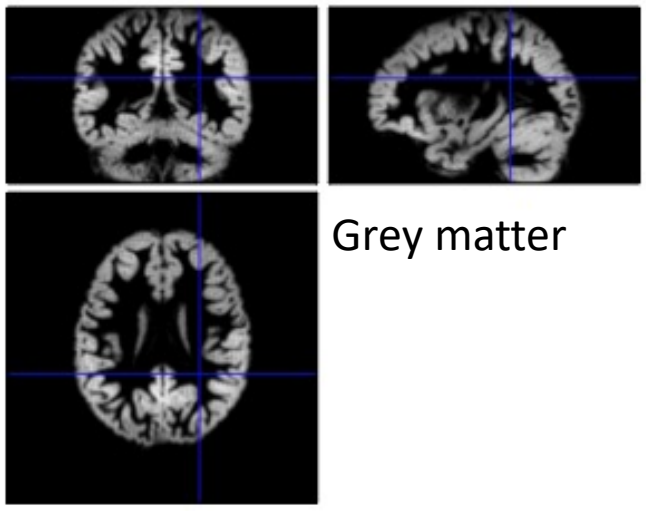
➡ **Quantification**

Segmentation of brain components

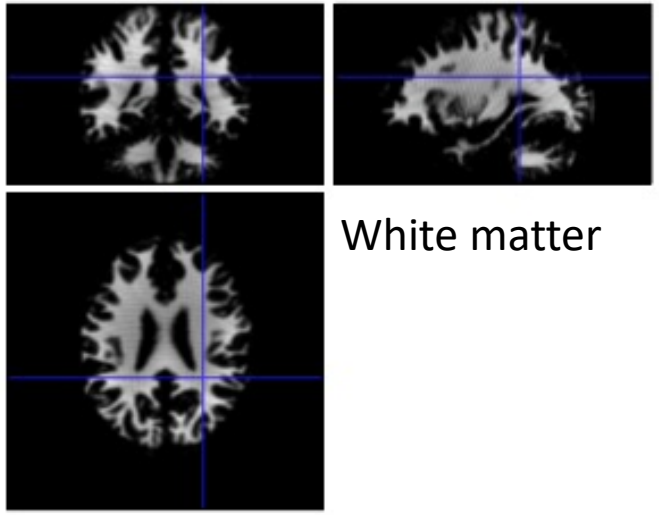


Tissue segmentation in SPM (Statistical Parametric Mapping)
Wellcome Trust Centre for Neuroimaging, Functional Imaging Laboratory (FIL), University College of London (UCL)

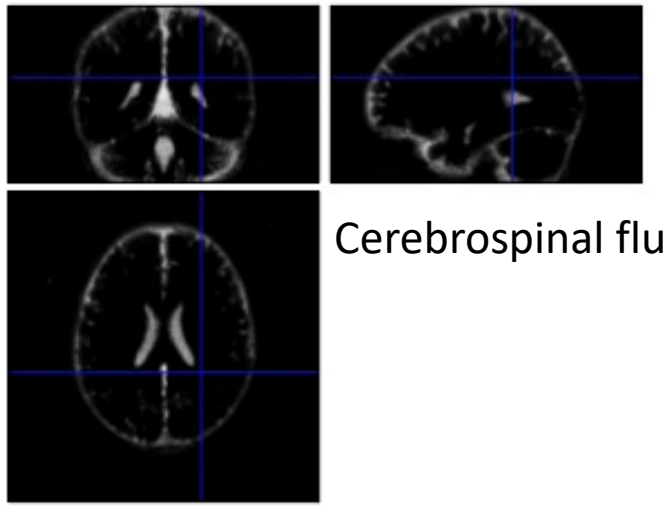
www.fil.ion.ucl.ac.uk/spm/



Grey matter

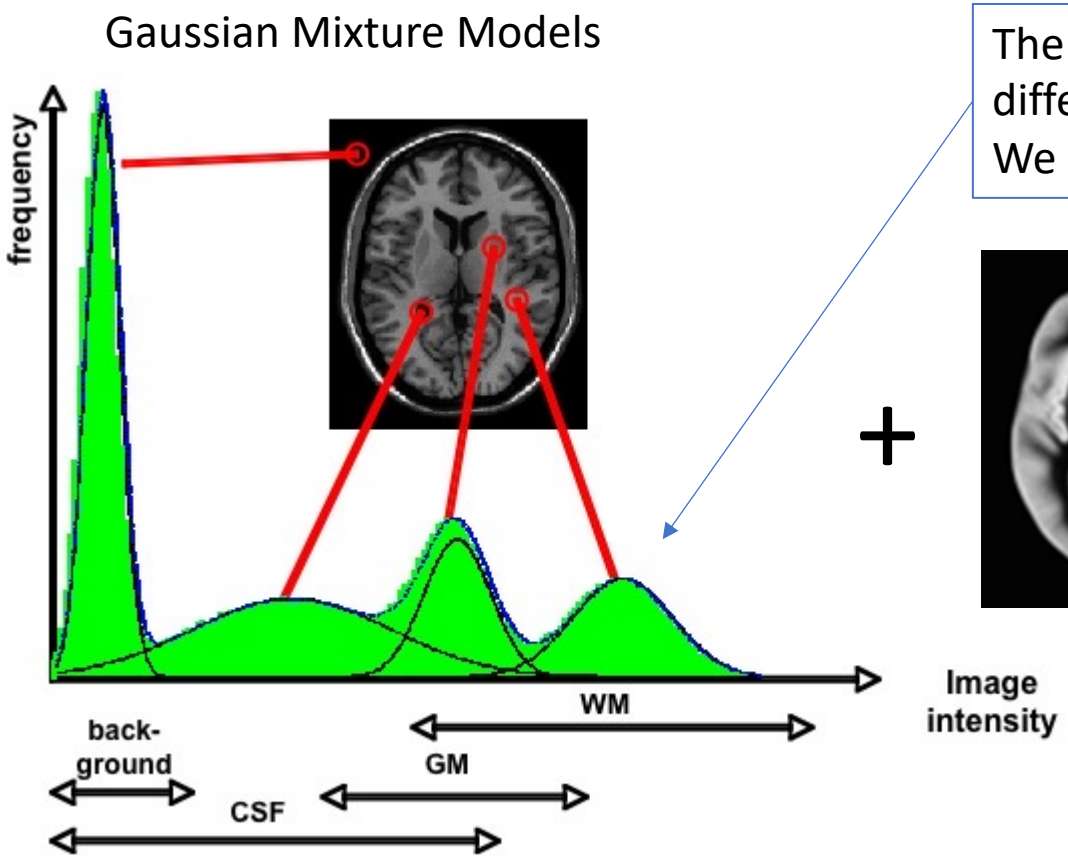


White matter

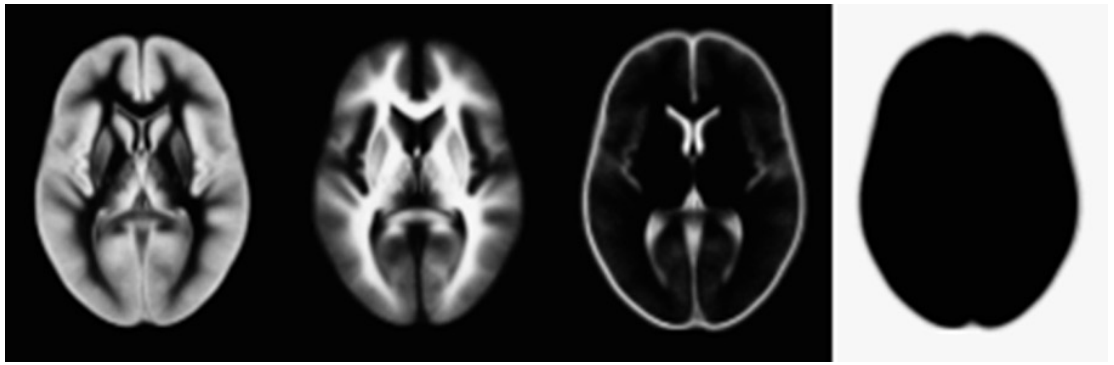


Cerebrospinal fluid

Gaussian Mixture Models (MOG)



The central values of these Gaussians are variable across different images/subjects.
We can exploit the spatial information provided by TPMs.



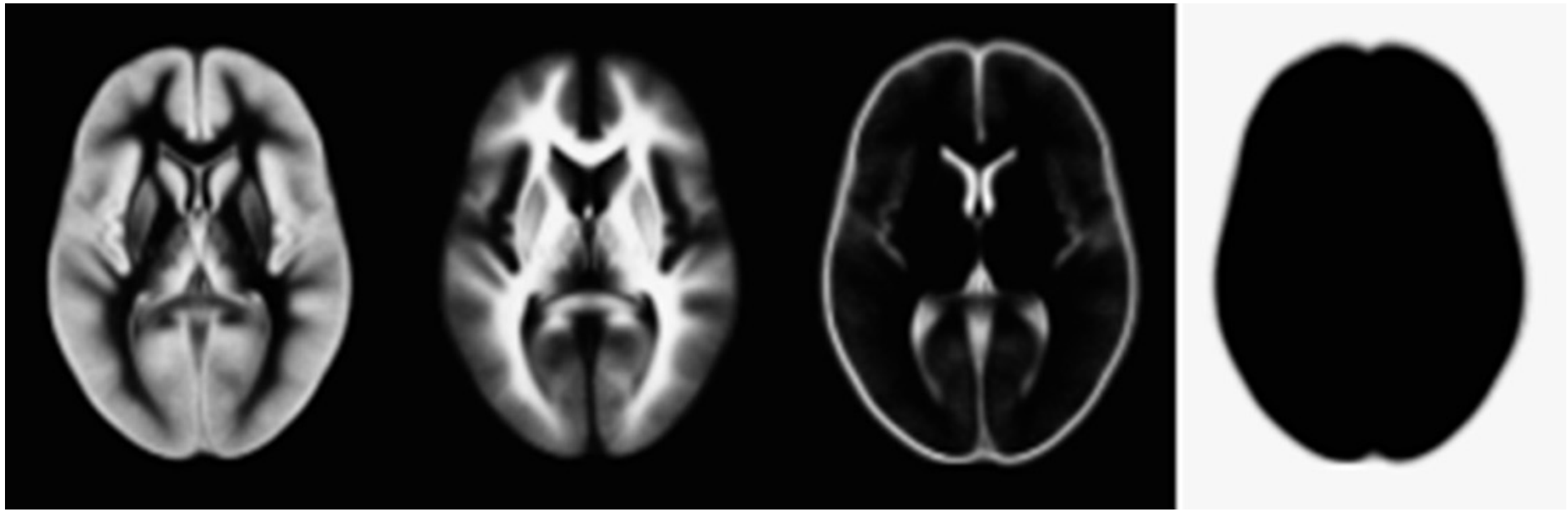
Tissue Probability Maps

- Take each voxel and look at both its intensity and prior spatial knowledge, to make segmentation an easier problem to solve.

Tissue Probability Maps

- Tissue Probability Maps (TPM): standard space structural estimates of the location of GM, WM, CSF, etc…
- Based on multiple previous segmentations

TPM,1 TPM,2 TPM,3 TPM,4



Segmentation of brain components

- In a simple MOG, the probability of obtaining a voxel with intensity y_i given that it belongs to the k^{th} cluster, i.e. Gaussian ($c_i = k$), and that the k^{th} Gaussian is parameterized by μ_k and σ_k^2 is:

$$P(y_i|c_i = k, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\right)$$

- The prior probability of any voxel, irrespective of its intensity, belonging to the k^{th} is:

$$P(c_i = k|\gamma_k) = \gamma_k \qquad \sum_{k=1}^{K} \gamma_k = 1$$

- Using Bayes rule, the joint probability of cluster k and intensity y_i is:

$$P(y_i, c_i = k|\mu_k, \sigma_k, \gamma_k)$$
$$= P(y_i|c_i = k, \mu_k, \sigma_k)P(c_i = k|\gamma_k)$$

⟹ Sum over K clusters (e.g. GM, WM, CSF, other components), accounting for each voxel and maximize with respect to the unknowns (μ, σ)

www.fil.ion.ucl.ac.uk/spm/

The freesurfer brain segmentation software

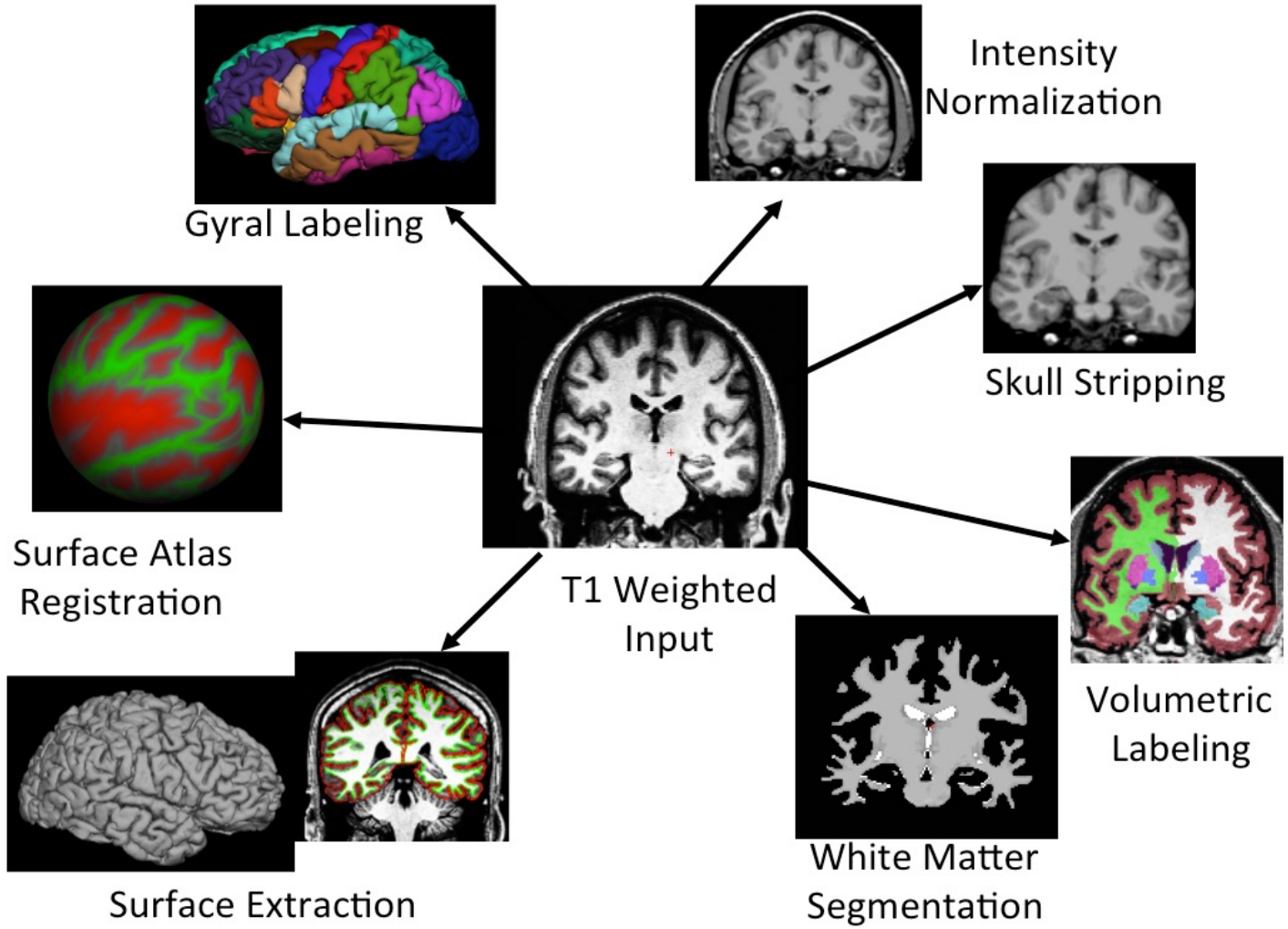
FreeSurfer v6.0
http://freesurfer.net/

FreeSurfer Software Suite

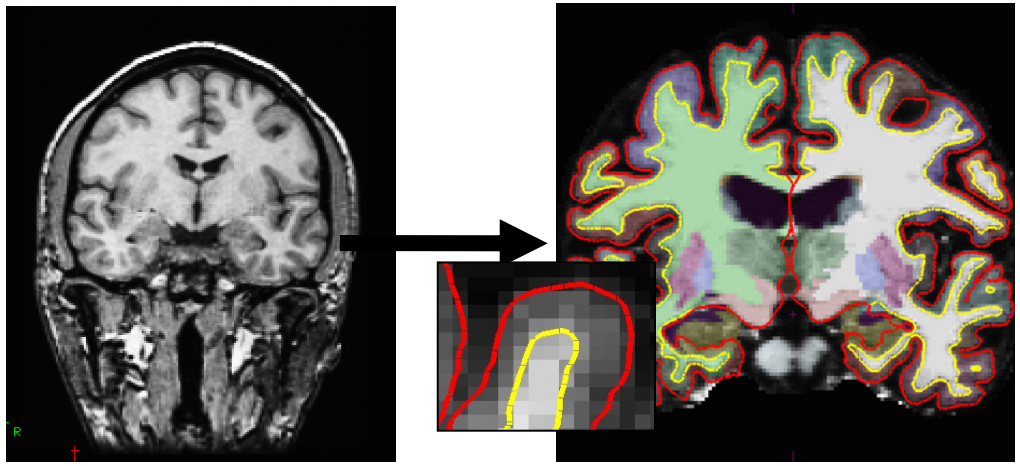
An open source software suite for processing and analyzing (human) brain MRI images.

- Skullstripping
- Image Registration
- Subcortical Segmentation
- Cortical Surface Reconstruction
- Cortical Segmentation
- Cortical Thickness Estimation
- Longitudinal Processing
- fMRI Analysis
- Tractography
- FreeView Visualization GUI
- and much more...

to run all steps in the FreeSurfer preprocessing stream: recon-all -all

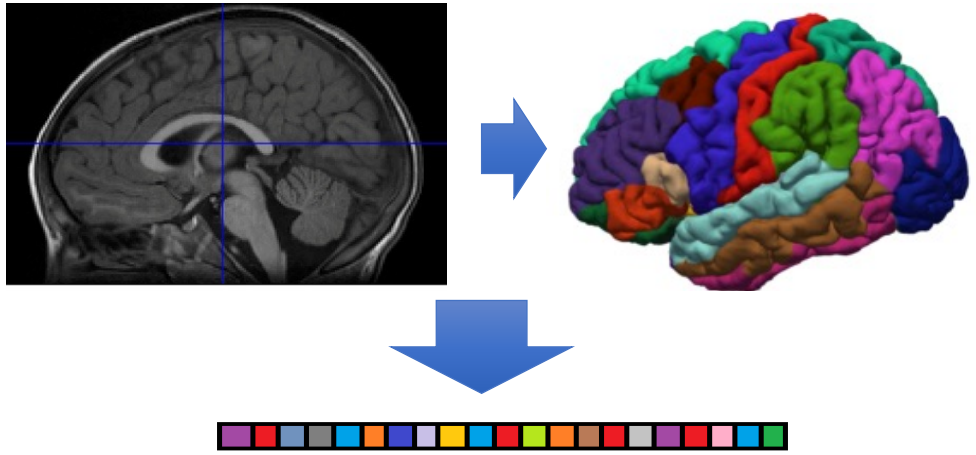


The freesurfer brain segmentation SW



Input: T1-weighted 1mm^3 resolution (.dcm)

Output: Segmented & parcellated volume (.mgz)



ROI-based features for each subject

http://surfer.nmr.mgh.harvard.edu/

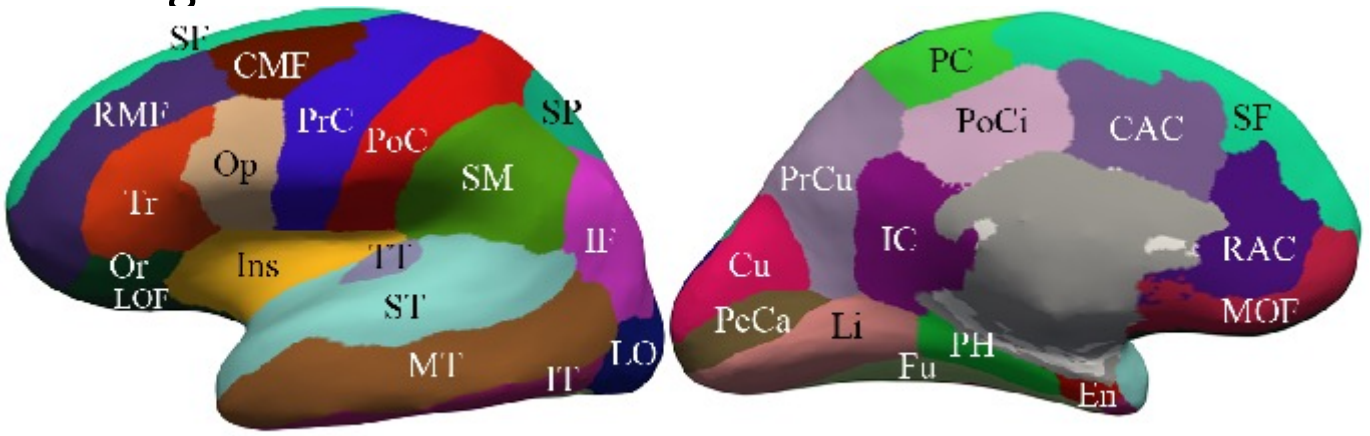
- The cerebral cortex is parcelled into a number of structures depending on the atlas we choose, e.g. **62 structures** according to Desikan-Killiany atlas.
- **Several Region-of-Interest (ROI)-based features** can be computed for each structure (**area**; **volume**; **average thickness**; **thickness standard deviation**; **mean curvature**)
- Also **Global features** can be considered, e.g.: **white surface total area** and **mean thickness of cerebral cortex** for both hemispheres.

The surface-based pipeline consists of several stages described in detail in:

- Dale, A.M., Fischl, Bruce, Sereno, M.I., Cortical Surface-Based Analysis I: Segmentation and Surface Reconstruction. NeuroImage 9(2):179-194. 1999.
- Fischl, B.R., Sereno, M.I., Dale, A. M. Cortical Surface-Based Analysis II: Inflation, Flattening, and Surface-Based Coordinate System. NeuroImage 9, 195-207. 1999.

Cortical parcellation

- Cortical parcellation according to the Desikan-Killiany atlas. Both hemispheres contain the same regions.



Region (gyrus)	Abb.
Caudal anterior cingulate	CAC
Caudal middle frontal	CMF
Cuneus	Cu
Entorhinal	En
Fusiform	Fu
Inferior parietal	IP
Inferior temporal	IT
Isthmus cingulate	IC
Lateral occipital	LO
Lateral orbitofrontal	LOF
Lingual	Li

Region (gyrus)	Abb.
Medial orbitofrontal	MOF
Middle temporal	MT
Parahippocampal	PH
Paracentral	PC
Pars opercularis	Op
Pars orbitalis	Or
Pars triangularis	Tr
Pericalcarine	PeCa
Postcentral	PoC
Posterior cingulate	PoCi

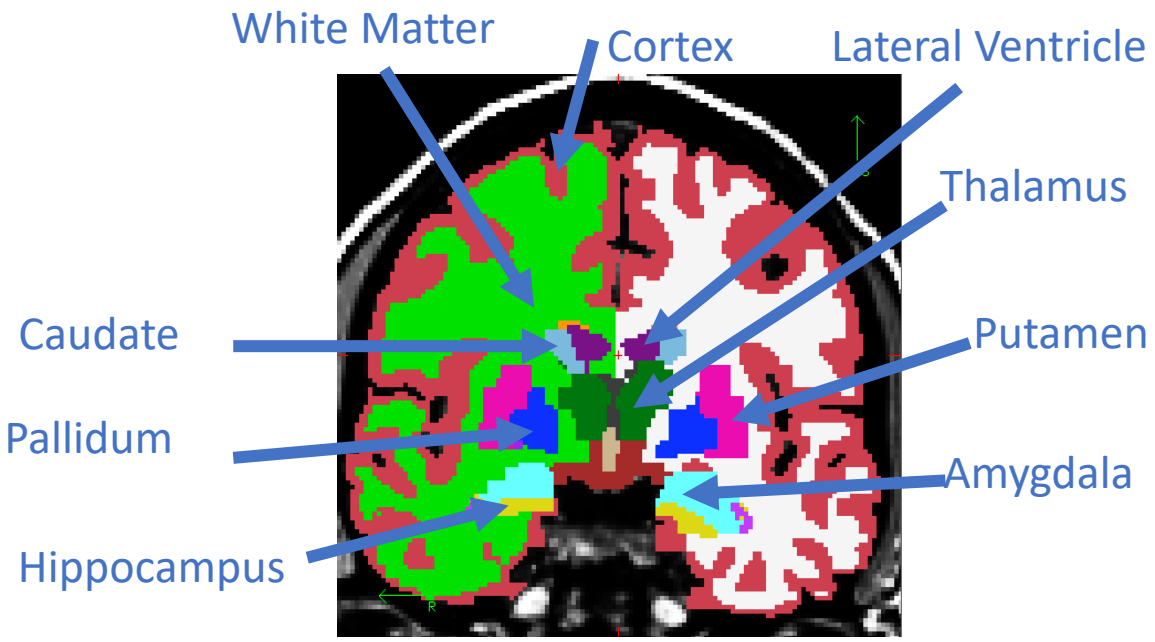
Region (gyrus)	Abb.
Precentral	PrC
Precuneus	PrCu
Rostral anterior cingulate	RAC
Rostral middle frontal	RMF
Superior frontal	SF
Superior parietal	SP
Superior temporal	ST
Supramarginal	SM
Transverse temporal	TT
Insula	Ins

R. S. Desikan, F. Sgonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest", *Neuroimage*, vol. 31, no. 3, pp. 968-980, 2006.

Subcortical volume segmentation

The volume-based stream to label subcortical tissue classes consists of five stages:

- affine registration with MNI305
- initial volumetric labeling*
- correction of variation in intensity due to inhomogeneities in the B1 magnetic field
- high dimensional nonlinear volumetric alignment to the MNI305 atlas
- final labeling



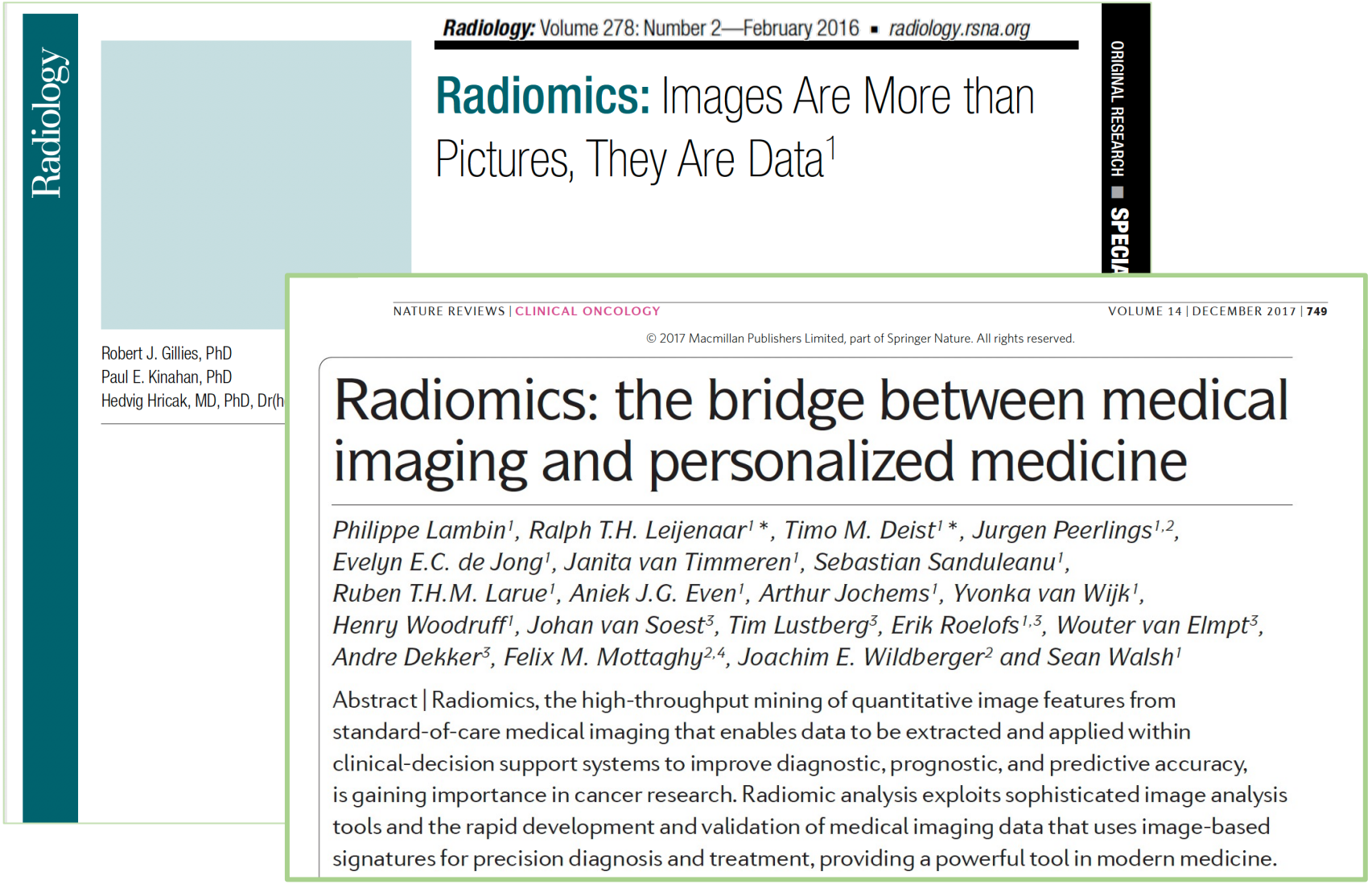
* The atlas used for labeling is built from a training set, i.e., a set of subjects whose brains (surfaces or volumes) have been labeled by hand.

The volume-based stream is designed to label subcortical tissue classes is described in detail in:

- Fischl B, Salat D, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale A. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron Jan 31;33(3):341-55. 2002.
- Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM. Automatically parcellating the human cerebral cortex. Cereb Cortex. 2004 Jan;14(1):11-22.

Radiomics

- 2016
- 2017



Radiology

Radiology: Volume 278: Number 2—February 2016 ▪ *radiology.rsna.org*

ORIGINAL RESEARCH ■ SPECIA

Radiomics: Images Are More than Pictures, They Are Data[1]

Robert J. Gillies, PhD
Paul E. Kinahan, PhD
Hedvig Hricak, MD, PhD, Dr(h

NATURE REVIEWS | CLINICAL ONCOLOGY VOLUME 14 | DECEMBER 2017 | 749

Radiomics: the bridge between medical imaging and personalized medicine

Philippe Lambin[1], Ralph T.H. Leijenaar[1], Timo M. Deist[1]*, Jurgen Peerlings[1,2], Evelyn E.C. de Jong[1], Janita van Timmeren[1], Sebastian Sanduleanu[1], Ruben T.H.M. Larue[1], Aniek J.G. Even[1], Arthur Jochems[1], Yvonka van Wijk[1], Henry Woodruff[1], Johan van Soest[3], Tim Lustberg[3], Erik Roelofs[1,3], Wouter van Elmpt[3], Andre Dekker[3], Felix M. Mottaghy[2,4], Joachim E. Wildberger[2] and Sean Walsh[1]*

Abstract | Radiomics, the high-throughput mining of quantitative image features from standard-of-care medical imaging that enables data to be extracted and applied within clinical-decision support systems to improve diagnostic, prognostic, and predictive accuracy, is gaining importance in cancer research. Radiomic analysis exploits sophisticated image analysis tools and the rapid development and validation of medical imaging data that uses image-based signatures for precision diagnosis and treatment, providing a powerful tool in modern medicine.

Radiomics

Radiology: Volume 278: Number 2—February 2016 ▪ *radiology.rsna.org*

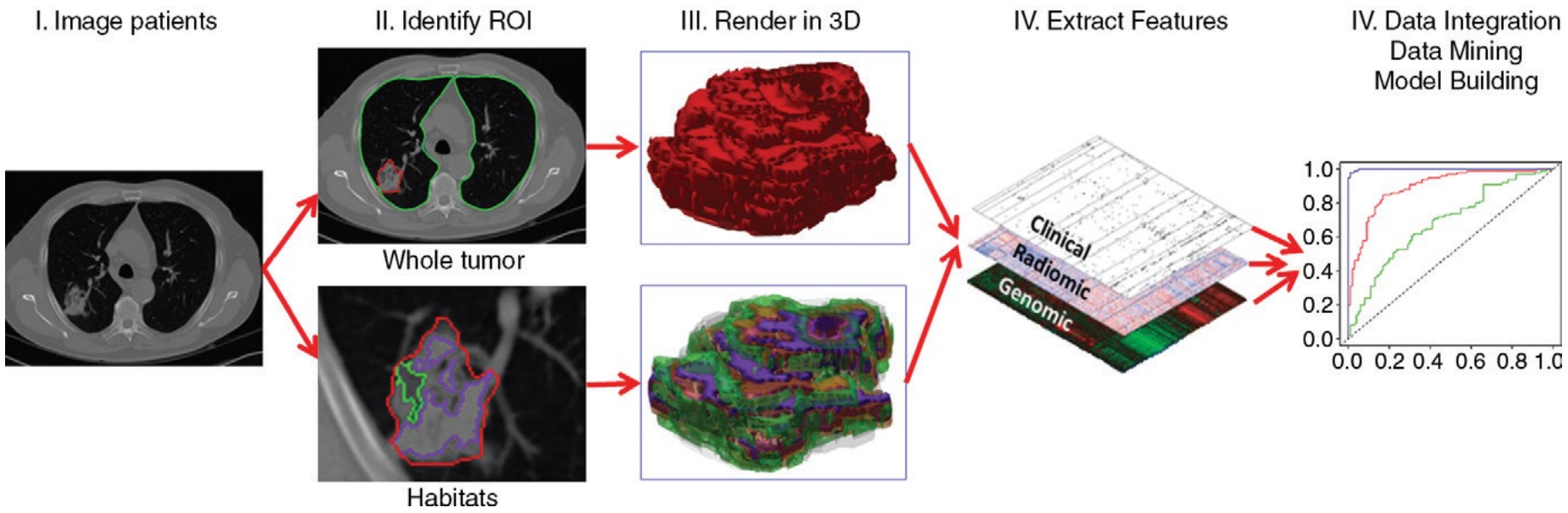


Figure 1: Flowchart shows the process of radiomics and the use of radiomics in decision support. Patient work-up requires information from disparate sources to be combined into a coherent model to describe where the lesion is, what it is, and what it is doing. Radiomics begins with acquisition of high-quality images. From these images, a region of interest *(ROI)* that contains either the whole tumor or subregions (ie, habitats) within the tumor can be identified. These are segmented with operator edits and are eventually rendered in three dimensions *(3D)*. Quantitative features are extracted from these rendered volumes to generate a report, which is placed in a database along with other data, such as clinical and genomic data. These data are then mined to develop diagnostic, predictive, or prognostic models for outcomes of interest.

What's new ?

- Radiological features -> quantitative measures to be reported in image annotations
- Integration with clinical and genomic data

PyRadiomics: open-source python package for the extraction of Radiomics features from medical images

- https://pyradiomics.readthedocs.io/en/latest/

- Start the python interactive session:
  - `python`
- Import the necessary classes:

```
from radiomics import firstorder, glcm, imageoperations, shape, glrlm, glszm, getTestCase
import SimpleITK as sitk
import six
import sys, os
```

- Set up a data directory variable:

```
dataDir = '/path/to/pyradiomics/data'
```

- You will find sample data files brain1_image.nrrd and brain1_label.nrrd in that directory.
- Use SimpleITK to read a the brain image and mask:

```
imageName, maskName = getTestCase('brain1', dataDir)
image = sitk.ReadImage(imageName)
mask = sitk.ReadImage(maskName)
```

- Calculate the first order features:

```
firstOrderFeatures = firstorder.RadiomicsFirstOrder(image,mask)
firstOrderFeatures.enableAllFeatures()  # On the feature class level, all features are disa
firstOrderFeatures.calculateFeatures()
for (key,val) in six.iteritems(firstOrderFeatures.featureValues):
  print("\t%s: %s" % (key, val))
```

Radiomic Features

This section contains the definitions of the various features that can be extracted using PyRadiomics. They are subdivided into the following classes:

- `First Order Statistics` (19 features)
- `Shape-based (3D)` (16 features)
- `Shape-based (2D)` (10 features)
- `Gray Level Cooccurence Matrix` (24 features)
- `Gray Level Run Length Matrix` (16 features)
- `Gray Level Size Zone Matrix` (16 features)
- `Neighbouring Gray Tone Difference Matrix` (5 features)
- `Gray Level Dependence Matrix` (14 features)

All feature classes, with the exception of shape can be calculated on either the original image and/or a derived image, obtained by applying one of several filters. The shape descriptors are independent of gray value, and are extracted from the label mask. If enabled, they are calculated separately of enabled input image types, and listed in the result as if calculated on the original image.

Most features defined below are in compliance with feature definitions as described by the Imaging Biomarker Standardization Initiative (IBSI), which are available in a separate document by Zwanenburg et al. (2016) [1]. Where features differ, a note has been added specifying the difference.

See demo code:
- Lecture5_demo1_extract_features_pyradiomics.ipynb

A list of free SW for processing medical images

- PET, MRI, US, RX and CT images
  - 3DSlicer, https://www.slicer.org
  - ImageJ, https://imagej.nih.gov/ij/index.html
  - Itk-SNAP, http://www.itksnap.org/pmwiki/pmwiki.php
  - LIFEx, www.lifexsoft.org
  - OsiriX (only for Mac), https://www.osirix-viewer.com

- Brain MRI:
  - AFNI, https://afni.nimh.nih.gov
  - Freesurfer, http://freesurfer.net
  - FSL, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
  - Mango, http://rii.uthscsa.edu/mango/mango.html
  - SPM (requires MATLAB, standalone versions also available), https://www.fil.ion.ucl.ac.uk/spm/software/