Biomedical Informatics Research

Horacio Gómez-Acevedo, PhD
Associate Professor
Department of Biomedical Informatics
University of Arkansas for Medical Sciences

April 21, 2023

UAMS

Plan

1. Biomedical Informatics
2. Machine Learning
3. Deep Learning
4. TCIA
5. Imaging Research
6. Final pitch for collaborating with DBMI

What is Biomedical Informatics (BMI)

The American Medical Informatics Association (AMIA) gives a broad definition of the subject.
Biomedical and health informatics applies principles of computer and information science to the advancement of life sciences research, health professions education, public health, and patient care.

- BMI develops, studies and applies theories, methods and processes for the generation, storage, retrieval, use, and sharing of biomedical data, information, and knowledge.
- BMI builds on computing, communication and information sciences and technologies and their application in biomedicine.
- BMI investigates and supports reasoning, modeling, simulation, experimentation and translation across the spectrum from molecules to populations, dealing with a variety of biological systems, bridging basic and clinical research and practice, and the healthcare enterprise

Areas of Research at DBMI

As of today, the Department of Biomedical Informatics (DBMI) has four speciality areas

- Translational Bioinformatics
- Clinical Research Informatics
- Imaging Informatics
- Clinical Informatics

We also have a fellowship in Biomedical Informatics.
But that is not all! Our faculty has very specific areas of research

- Radiomics, connectomics, neuro-imaging, etc.
- Natural Language Processing
- Ontologies
- Genomics, metagenomics, microbial genomics, cancer genomics, etc.

Omics in medicine

Since the molecular function is not restricted to the genome, other modalities are also informative

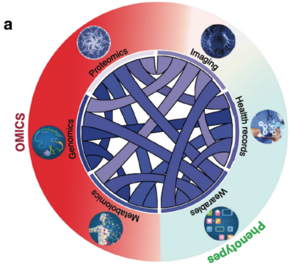
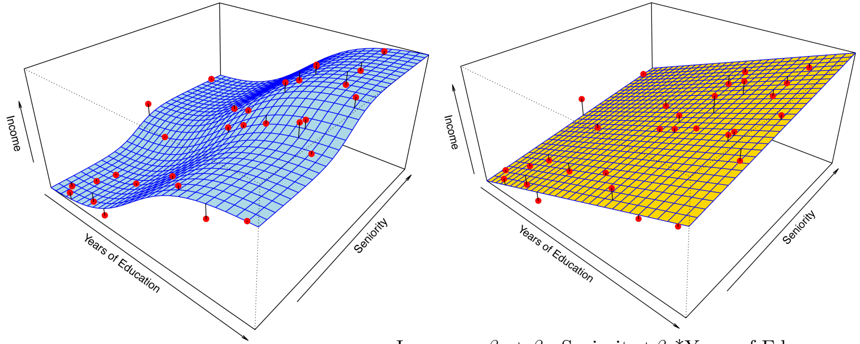


Fig 8: Other Omics in healthcare

Statistics and traditional parametric approach

Income=f(Seniority,Years of Ed)+ϵ



Income $\approx \beta_0 + \beta_1$∗Seniority+$\beta_2$*Years of Ed

Multi-linear regression

Statistics and traditional parametric approach

The philosophy of *classical* parametric paradigm is based on three beliefs

1. The functional dependency from the data can be approximated with a linear function with an small number of parameters
2. Random errors in real life problems follow a normal distribution.
3. Those parameters in the model can be calculated via the maximum likelihood method.

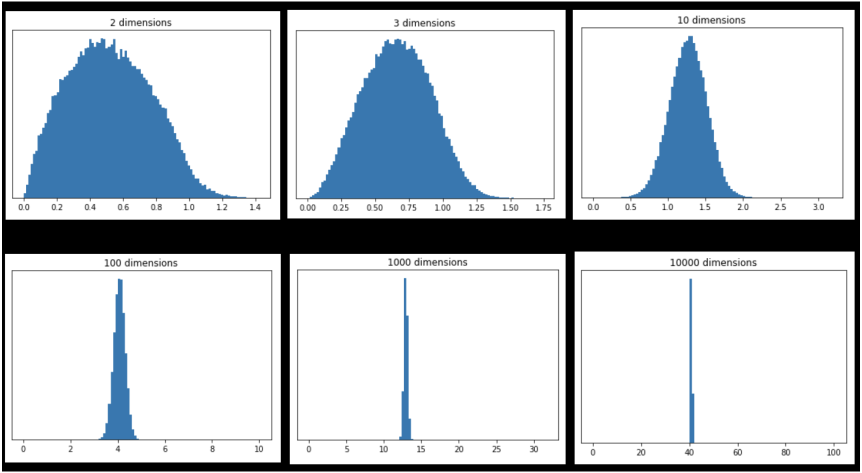
Parametric approach for big data problems

- It was until the computers were introduced that more challenge problems were attempted.
- For large multivariate problems, it was observed that increasing the number of factors required more and more computational resources.
- R. Bellman called this phenomena the **curse of dimensionality**

New techniques were developed to make informal inferences of data instead of relying in purely statistical techniques.

Curse of Dimensionality

In higher dimensions our intuition is severely impaired



Random vectors in $\mathbb{R}^n$ (Johh Urbanic, Pittsburgh Supercomputing Center)

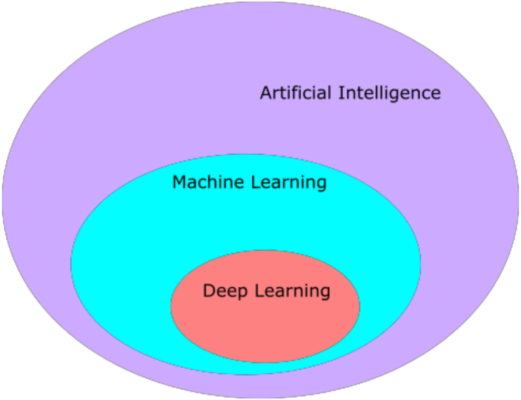
What is Machine Learning?

Machine learning (ML) is a vast field and here are some definitions.

> *(ML is the) field of study that gives computers the ability to learn without being explicitly programmed.*
> *Arthur Samuel 1959*

> *A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T as measured by P, improves with experience E.*
> *Tom Mitchell 1997*

Machine Learning $\neq$ Artificial Intelligence



When do you use ML?

- Problems for which existing solutions require a lot of hand-tuning or long lists of rules.
- Complex problems for which there is no good solution at all using a traditional approaches.
- Fluctuating environments as Machine Learning systems can adapt to not previously seen data.
- **Getting insights about complex problems and large amounts of data.**

Types of ML algorithms

ML algorithms can be classified according to the amount and type of supervision is required during training. There are fundamentally four major categories:

1. supervised
2. unsupervised
3. semisupervised
4. reinforced learning

Supervised Learning

In Supervised Learning, the training data you feed the algorithm includes the desired solutions (called **labels**).

- **Classification**. Algorithm is trained with samples along with their class, and we select parameters that discerns best the labels.
- **Regression**. Algorithm is trained and predict a target numeric value given a set of features (covariates or predictors).

Depending on the data set or problem, we divide our sample in three classes:

- Training set. A subset in which the algorithm will fit the best model.
- Testing set. A subset in which our measures of performance will be used.
- Validation set. Ideally, and independent set that will see how our model performs in not-previously seen data.

Supervised Learning (cont)

Other supervised learning algorithms include

- Multilinear (logistic) regression
- Suport Vector Machines (SVMs)
- Decision Trees and Random Forest
- Linear Discriminant Analysis

Basic Classification: Separation by planes

A very common approach for classification of vectors is to find a (hyper)plane that separates the data

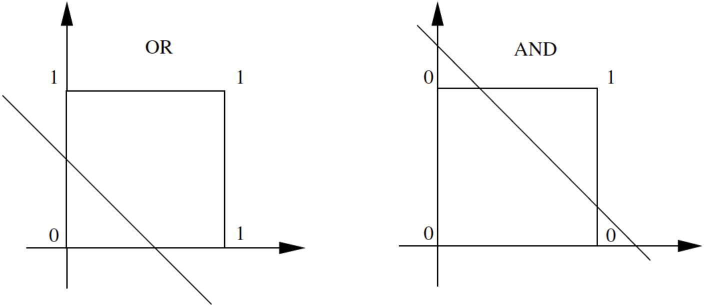
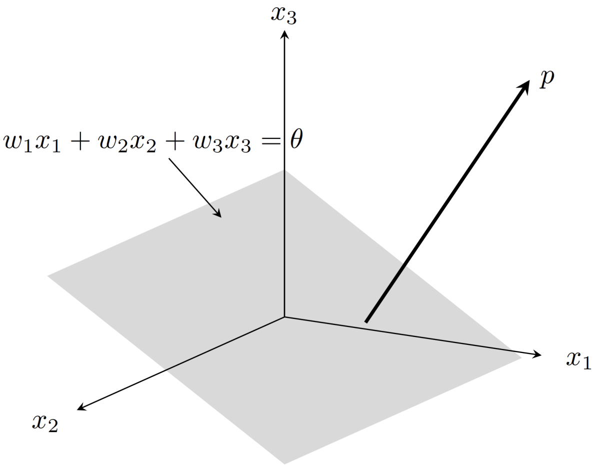


Fig. 3.6. Separations of input space corresponding to OR and AND

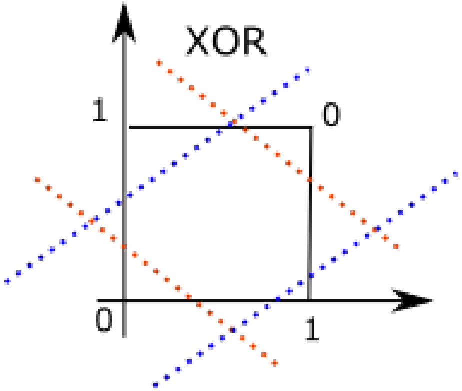
Separation by hyperplanes

More generally, we separate points (vectors) in higher dimensions with hyperplanes



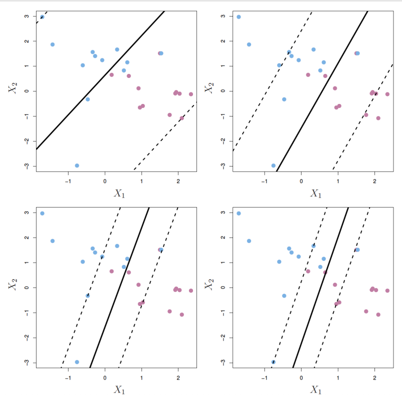
XOR function

However, even when we have a relatively simple function XOR this separation is not possible in $\mathbb{R}^2$.



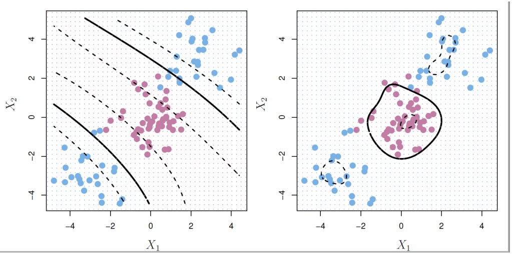
Support Vector Machines

Using an optimization procedure over the number of misclassified elements close to the hyperplane, we develop another separation algorithm based on the support vectors.



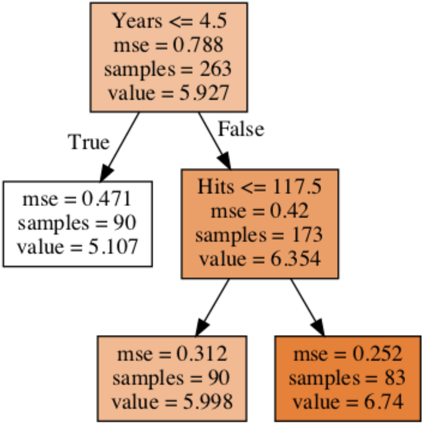
Support Vector Machines

This methodology allows us to use **kernels** to better capture the classification.



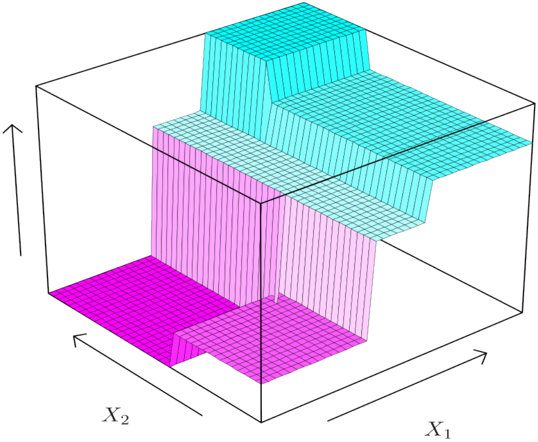
Tree-based methods

One of the most intuitive methods for classification is decision trees.



Partitions

Regression tree based methods have a very simple interpretation .



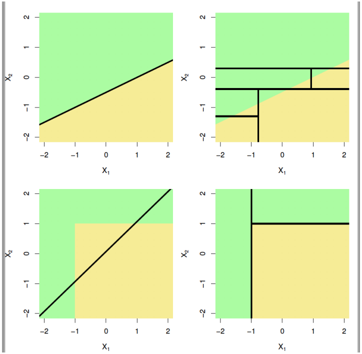
Random Forest

In the setup of classification, we conducted this procedure in two steps

1. Construct binary decision trees using bootstrapped training sets (without pruning)
2. Predict the value based on the majority vote. That is the class most commonly occurring will be selected.

Trees vs OLS

There are some setups where tree methods are more appropriate



Random forest (boosting/bagging) are very powerful methodologies that compete with regression, their interpretation is complicated.

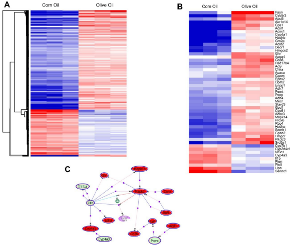
Unsupervised Learning

In unsupervised learning the training data is unlabeled and the system tries to learn without programmers intervention.
Some of the most important types of unsupervised learning are

- Clustering (K-means)
- Hierarchical Cluster Analysis (HCA)
- Visualization and dimensionality reduction
  1. Principal Component Analysis (PCA)
  2. Manifold Learning (t-SNE, MDS)

Clustering

The goal of clustering is to detect groups with similar characteristics. If you use hierarchical clustering algorithm it divides each group into smaller subgroups based on certain similarities.

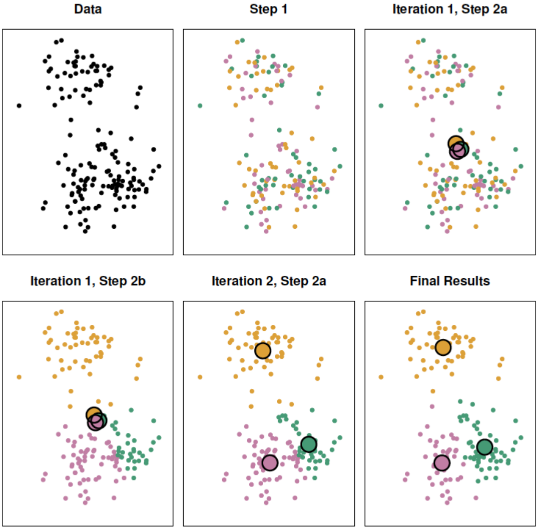


Classification with k-Means Clustering

The algorithm goes like this

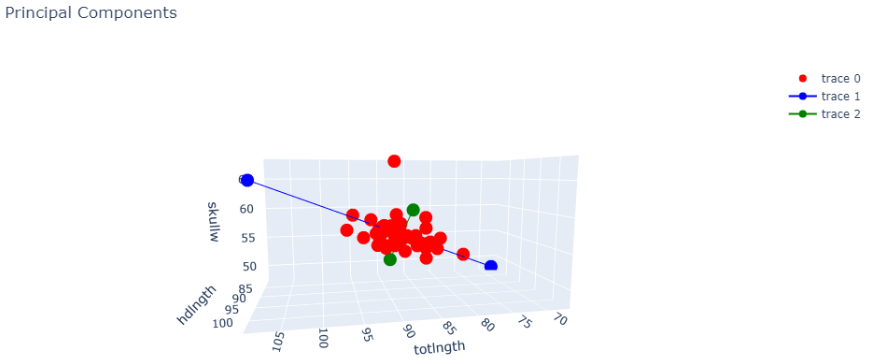
1. Assign randomly a value between 1 and K to each data point.
2. Iterate the following procedure until the clusters assignments stop changing
    1. Find the centroid for each of the K clusters.
    2. Each point will be assigned to the cluster K whose distance is the smallest. If two or more are equidistant, select randomly the cluster among the equidistant clusters.

K-Means Clustering (cont.)

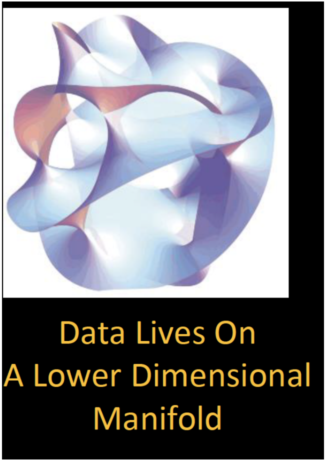


PCA

The idea of principal component analysis is to reduce the number of dimensions while preserving the data variability.



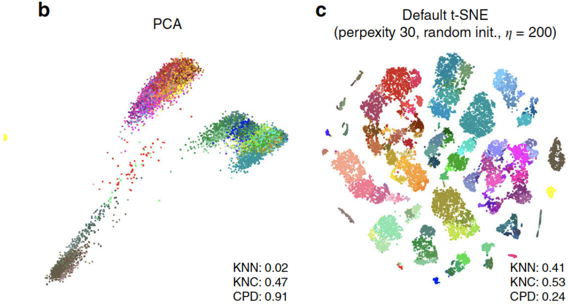
Manifold Hypothesis



John Urbanic, Pittsburgh Supercomputing Center

t-Stochastic Neighbor Embedding

This is another methodology that uses the data and their corresponding embedding (projection) and minimizes the Kullback-Leibler divergence of the normalized distribution of the data and their corresponding embedding.



Semisupervised Learning

Algorithms that handle a mixture of labeled and unlabeled data are called semisupervised. As the name suggests, it uses a combination of supervised and unsupervised algorithms. For instance deep belief networks (DBNs) are based on unsupervised components called restricted Boltzmann machines that are trained sequentially in an unsupervised manner, and then the whole system is fine-tuned using supervised learning techniques.

Reinforced Learning

The learning system is called an agent in this context, and can observe the environment, select and perform actions and get rewards or penalties. It must then learn by itself what is the best strategy, called a policy, to the most reward over time. A policy defines what action the agent should choose when it is in a given situation. DeepMind's Alpha Go program used reinforced learning to beat world champion Ke Jie at the game of Go. It learned its winning policy by analyzing millions of games, and then playing many games against itself.

Challenges in ML

There are either bad algorithms or bad data that can derail any serious ML effort.

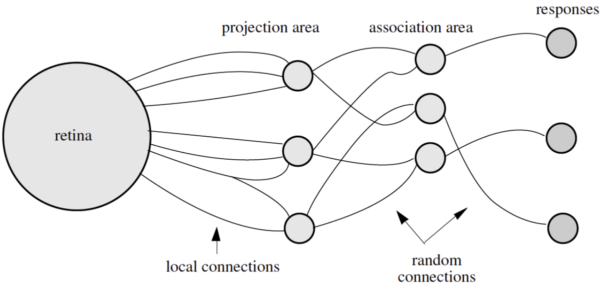
- Insuficient quality of training data. For very simple problems you typically need thousands of samples, and for complicated such as image or speech recognition you may need millions of samples.
- Nonrepresentative training data To generalize a ML algorithm you need that your training data be representative of the new cases you want to generalize. If the sample is too small, you will have sampling noise, but even very large samples can be nonrepresentative if the sampling method is flawed. This is called sampling bias

Challenges in ML (cont)

- Poor quality data. If your training data is full of errors, outliers and noise, it will make the ML methodology to under perform. Data quality is a field on its own right. Missing features can happen at random or being systematic. Spend enough time with your data to decide quality or features that are missing loads of information.
- Overfitting. This occurs when the model is too complex relative to the amount and noisiness of the training data. Regularization if the process of reducing overfitting by making the model simpler by introducing a hyperparameter (this is a parameter not for the model but for the algorithm).

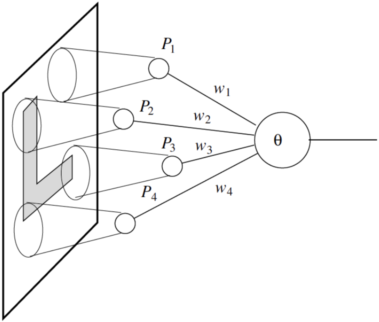
In the beginning...

In 1958 Frank Rosenblatt proposed the **perceptron** which is a more general computational model. The essential innovation was the introduction of numeral weights and a special interconnection pattern. In the original Rosenblatt model the computing units are threshold elements and the connectivity is determined stochastically. Learning takes place by adapting the weights of the network with a numerical algorithm. The so-called classical perceptron is depicted below

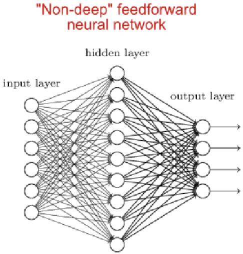


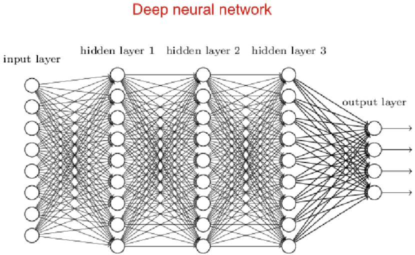
Threshold logic

The computational equivalent is a network with weights w_i and a number of input units P_i. If the combined value of the inputs is larger than θ, then it fires a 1, otherwise a 0.



Deep Neural Networks





Deep Learning

> *Deep learning achieves great power and flexibility by representing the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts and more abstract representations computed in terms of less abstract ones.*
> *Goodfellow et al.*

Deep learning is only possible because of the advances in computer power (GPUs in particular). The basic concepts date back to then end of 1990's with Y. LeCun and colleagues.

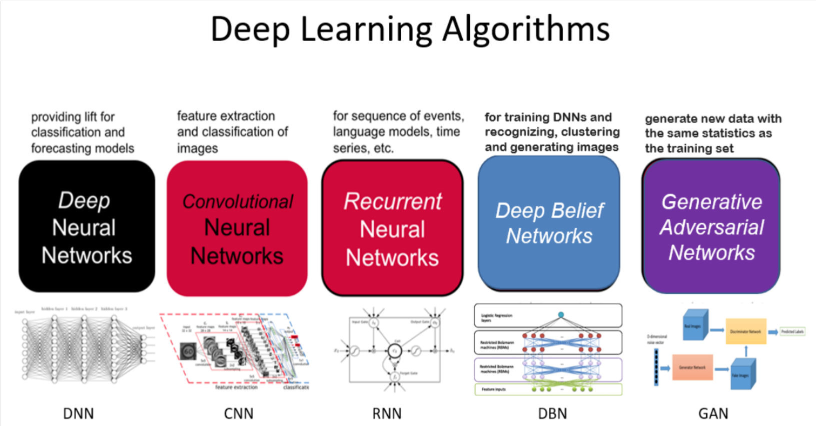
Deep Learning in real time

One impressive representation of deep learning applied to a classification problem. This is an extension of the famous MINST digit recognition problem.

`https://adamharley.com/nn_vis/mlp/3d.html`

Deep Learning Methods

Every DL method relies on a specific design (architecture) and purpose

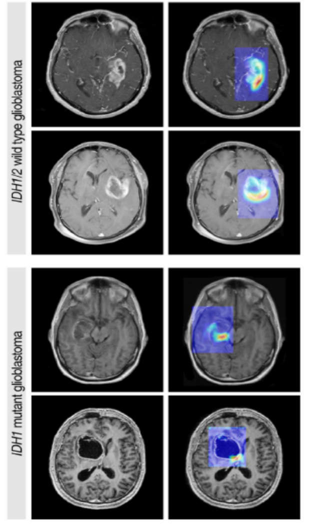


Understanding, Interpreting, Explaining

- Understanding a model ⇒ characterizing a model's behavior without elucidating its internal mechanisms
- Interpreting a model ⇒ rendering the algorithms decision-making process in terms a human expert can comprehend
- Explaining a model ⇒ defining the collection of features (usually in the input) that are most significant in producing the algorithm's output

Holzinger et al. arXiv:1712.09923 (2017).

Visual Explainability in CNN



Grad-CAM algorithm. Red regions weights on the discriminative regions for IDH status.

Bi et al. CA Cancer J Clin 2019

Deep Learning pitfalls

Applications of deep learning and architectures is abundant, and results really impressive. However, there are still some fundamental theoretical problems that have not been solved. For instance the **Black box problem** in which is not clear how the deep neural network makes decisions and how can it be modified.

The Cancer Imaging Archive

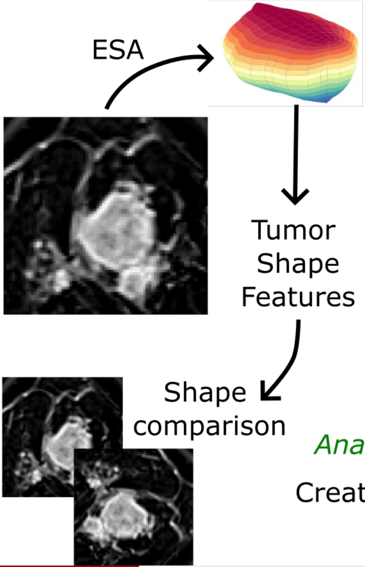
UAMS hosts the Cancer Imaging Archive

What do I do in DBMI?

I have research projects in:

- Statistical Machine Learning
- Functional and Elastic Shape Analysis
- Deep Learning in imaging
- Neuro-imaging
- Cancer genomics/imaging data integration
- Limb Development

Imaging Research



Elastic Shape Analysis

Differential Geometry
+
Functional Data Analysis

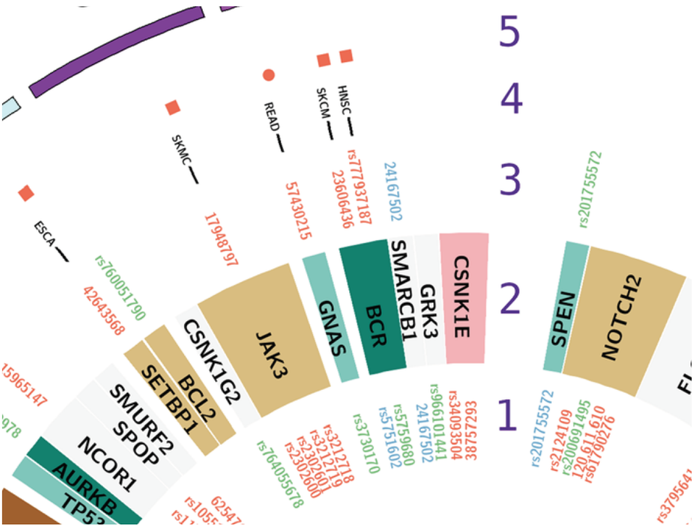
Deep Learning

Semi-automatic
segmentation

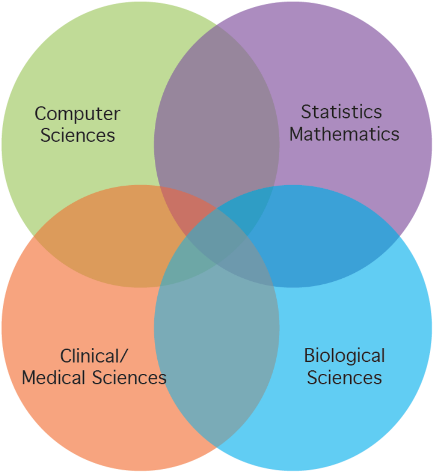
Analytic Data Augmentation

Creating sensible artificial
images

Bioinformatics



DBMI is highly interdisciplinary



What computer language do I need?

It really depends on the area that you want to specialize and your background. At DBMI, all our students must acquire certain competency in Python

Computer/Scripting Languages

Figure credits

Some figures come from James et al. An introduction to statistical learning, and G. Rojas, Neural Networks, both from Springer Verlag. Dr. Fred Prior kindly shared the image on Deep Learning Algorithms from our Machine Learning lecture series.