Introduction to machine learning-part2

Last time

- What is Machine learning
- Machine learning applications
- Regression vs Classification
- Machine learning evaluation (validation, cross validation)
- Machine learning components

- Does anyone play chatbot?

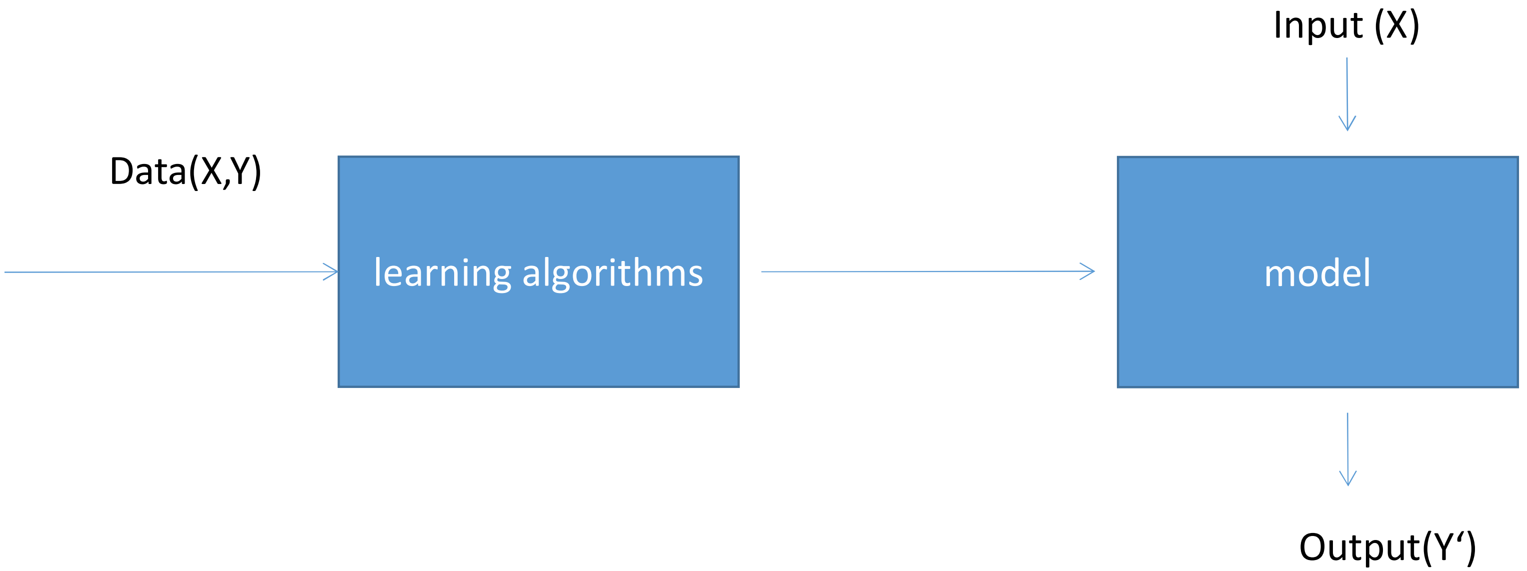
Learning

- Supervised learning
- Unsupervised learning
- Other learning

Supervised learning

- The training data set is a labeled data set.
  - In other words, the training data set contains the input value (X) and target value (Y).
- The learning algorithm generates a model.
- Then, new data set consisting of only the input value is fed.
- The model then generates the target value based on its learning.

Supervised learning



How to evaluate ?

Example of Y, Y’?

whats the range of X?

Whats the form of Y?

Applications

Spam classification

A simple dataset consisting of email and labels

Emails	Spam
email 1	Y
email 2	N
…	….

In this training data set, emails categorized as spam or not are done by a expert’s knowledge.

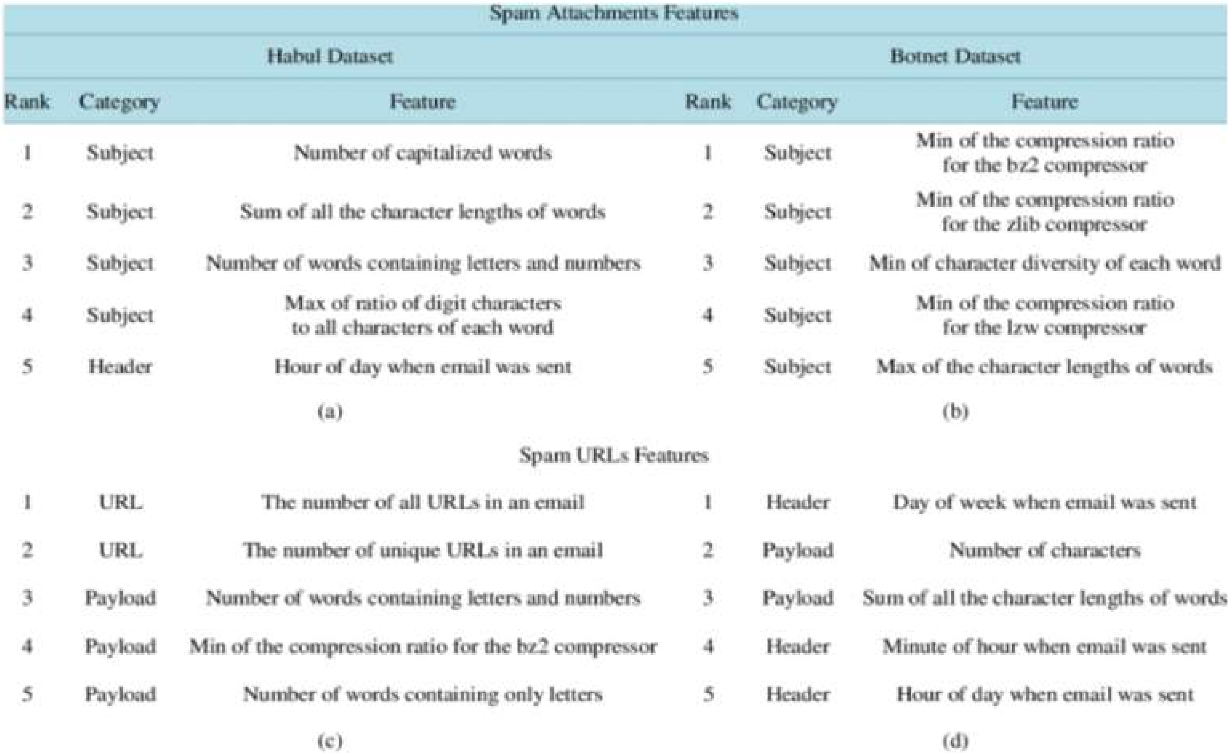
So, it is supervised learning algorithm

what are the features of emails ?

can you use email directly for training?

Spam

Improving Knowledge Based Spam Detection Methods: The Effect of Malicious Related Features in Imbalance Data Distribution



Spam Attachments Features					
		Habul Dataset			Botnet Dataset
Rank	Category	Feature	Rank	Category	Feature
1	Subject	Number of capitalized words	1	Subject	Min of the compression ratio for the bz2 compressor
2	Subject	Sum of all the character lengths of words	2	Subject	Min of the compression ratio for the zlib compressor
3	Subject	Number of words containing letters and numbers	3	Subject	Min of character diversity of each word
4	Subject	Max of ratio of digit characters to all characters of each word	4	Subject	Min of the compression ratio for the lzw compressor
5	Header	Hour of day when email was sent	5	Subject	Max of the character lengths of words
		(a)			(b)
		Spam URLs Features			
1	URL	The number of all URLs in an email	1	Header	Day of week when email was sent
2	URL	The number of unique URLs in an email	2	Payload	Number of characters
3	Payload	Number of words containing letters and numbers	3	Payload	Sum of all the character lengths of words
4	Payload	Min of the compression ratio for the bz2 compressor	4	Header	Minute of hour when email was sent
5	Payload	Number of words containing only letters	5	Header	Hour of day when email was sent
		(c)			(d)

Top 5 spam detection features [19].

Other applications

- Cancer detection
- House Price Prediction
- Credit Scoring (high risk or a low risk customer while lending loans by the banks)
- Face Recognition etc

What are the X and Y for these applications?

Computer-aided diagnosis/detection

“In the digital imaging community the term annotation is commonly used for visible metadata superimposed on an image without changing the underlying master image, such as sticky notes, virtual laser pointers, circles, arrows…” (**From Wikipedia**)

In CAD research: An annotation is metadata related to an imaging exam holding the location and description of a lesion

Purpose of image annotations

- Measurement of CAD performance
- Development of algorithms
- Training of CAD systems
- Design of observer studies
- Development of teaching tools

Annotation

- Patient level

- Image level

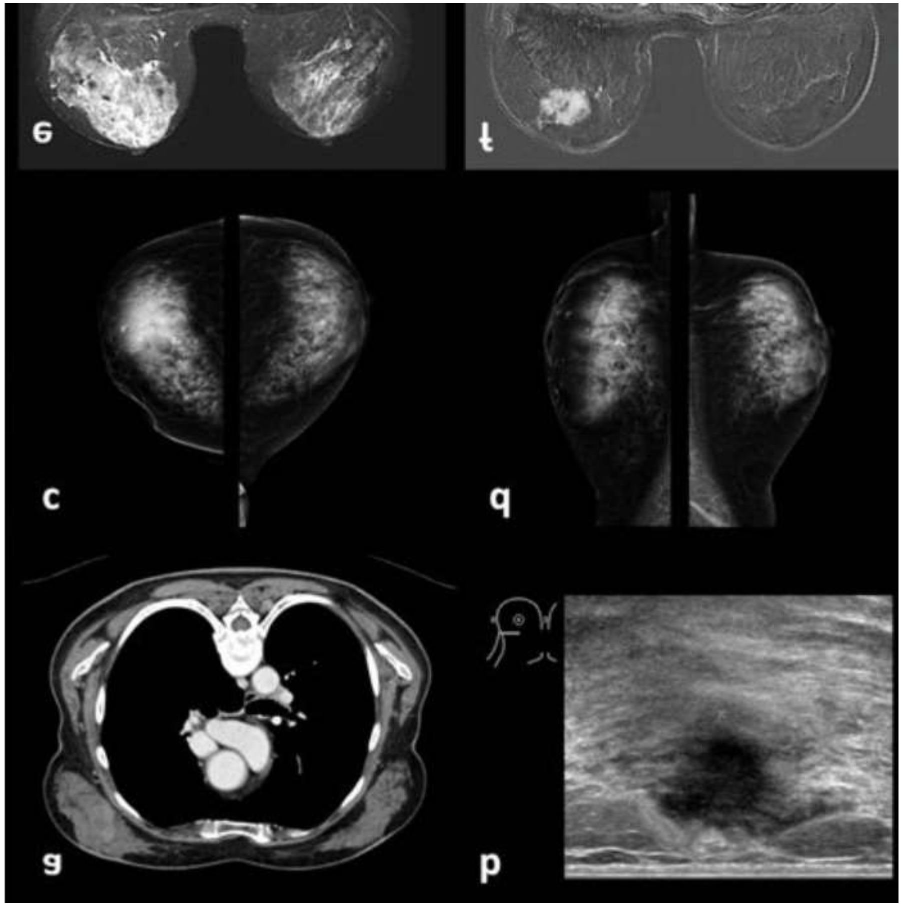
- Location level

Annotation issues: Location

- Can simply be the coordinates of a lesion
  - World coordinates are preferable
  - In 2D imaging image coordinates are also used. Scale or resolution should be included
- Lesion extent
  - Visual size estimate may be included
  - Segmentation
    - Tedious and time consuming in 3D
    - Subjective, but not crucial for many CAD research projects

Annotation issues: Location linking

- Multiple views of a lesion
  - Multiple views of organ in one exam
    - Anterior and posterior chest exams
    - MLO and CC views in a mammogram
  - Multiple exams, series
    - MRI sequences (T1, T2, DTI, Spectroscopy)
    - Retakes (e.g. screening and diagnostic exams)
    - Prior examinations
  - Multimodal studies
    - Example: Mammography, Tomo, Ultrasound and MRI



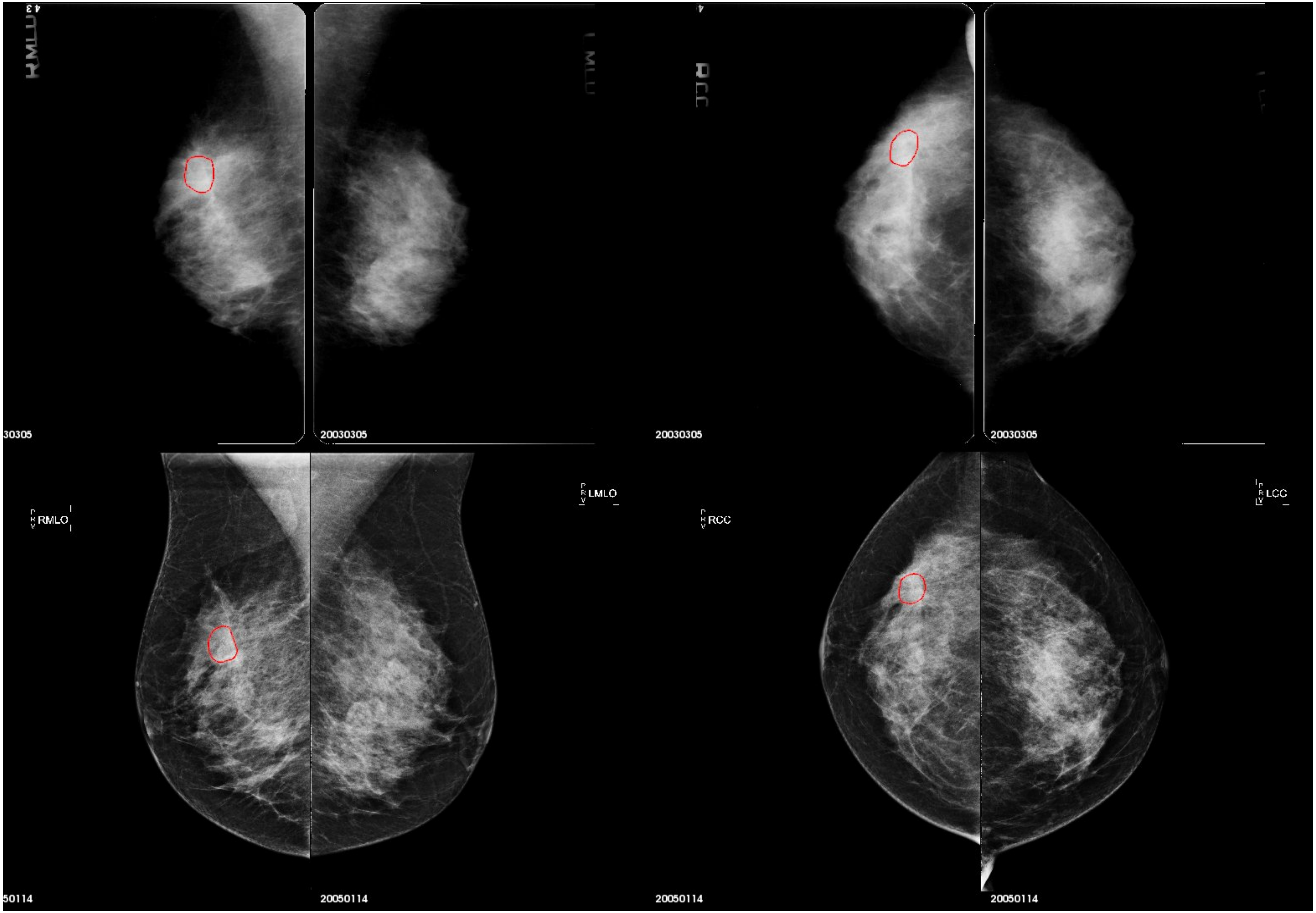
Annotation: Lesion description

- Pathology
  - Benign / malignant
  - Multi category classification
  - Verification by biopsy or follow-up
- Visual appearance
  - Type of pattern
  - Subtlety
  - Appearance may vary across views, series and exams

Annotation: Standardization

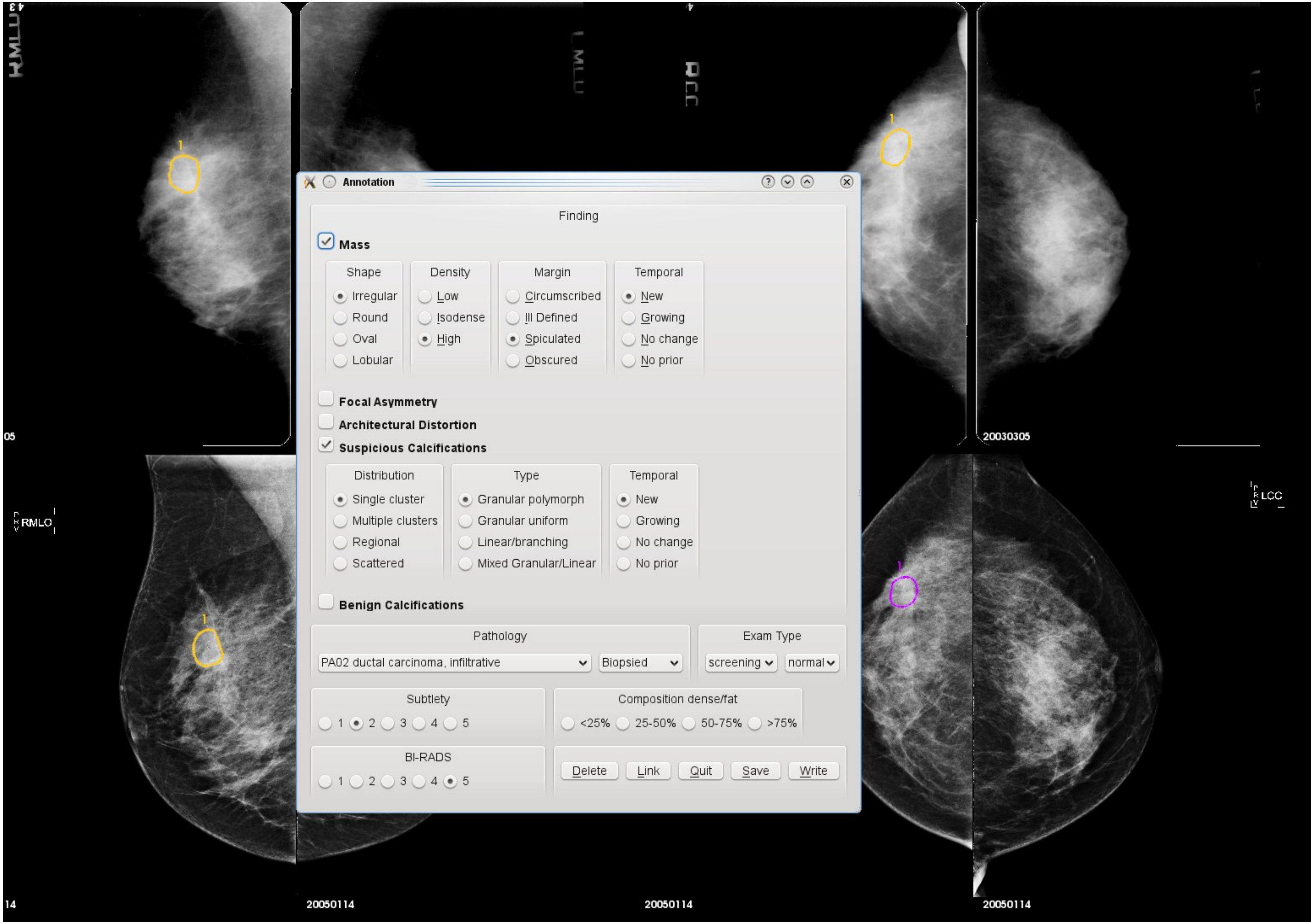
- Initiatives to develop standard: Annotation of Image Markup (AIM)
- Basis could be structured reporting in clinical workflow
  - BI-RADS in breast imaging is an example
- DICOM Structured Reports

An



43
RMLO
MLO
RCC
30305
20030305
20030305
20030305
RMLO
LMLO
RCC
LCC
50114
20050114
20050114
20050114

An



Annotation

Finding

☑ **Mass**

Shape	Density	Margin	Temporal
● Irregular	○ Low	○ Circumscribed	● New
○ Round	○ Isodense	○ Ill Defined	○ Growing
○ Oval	● High	● Spiculated	○ No change
○ Lobular		○ Obscured	○ No prior

□ **Focal Asymmetry**

□ **Architectural Distortion**

☑ **Suspicious Calcifications**

Distribution	Type	Temporal
● Single cluster	● Granular polymorph	● New
○ Multiple clusters	○ Granular uniform	○ Growing
○ Regional	○ Linear/branching	○ No change
○ Scattered	○ Mixed Granular/Linear	○ No prior

□ **Benign Calcifications**

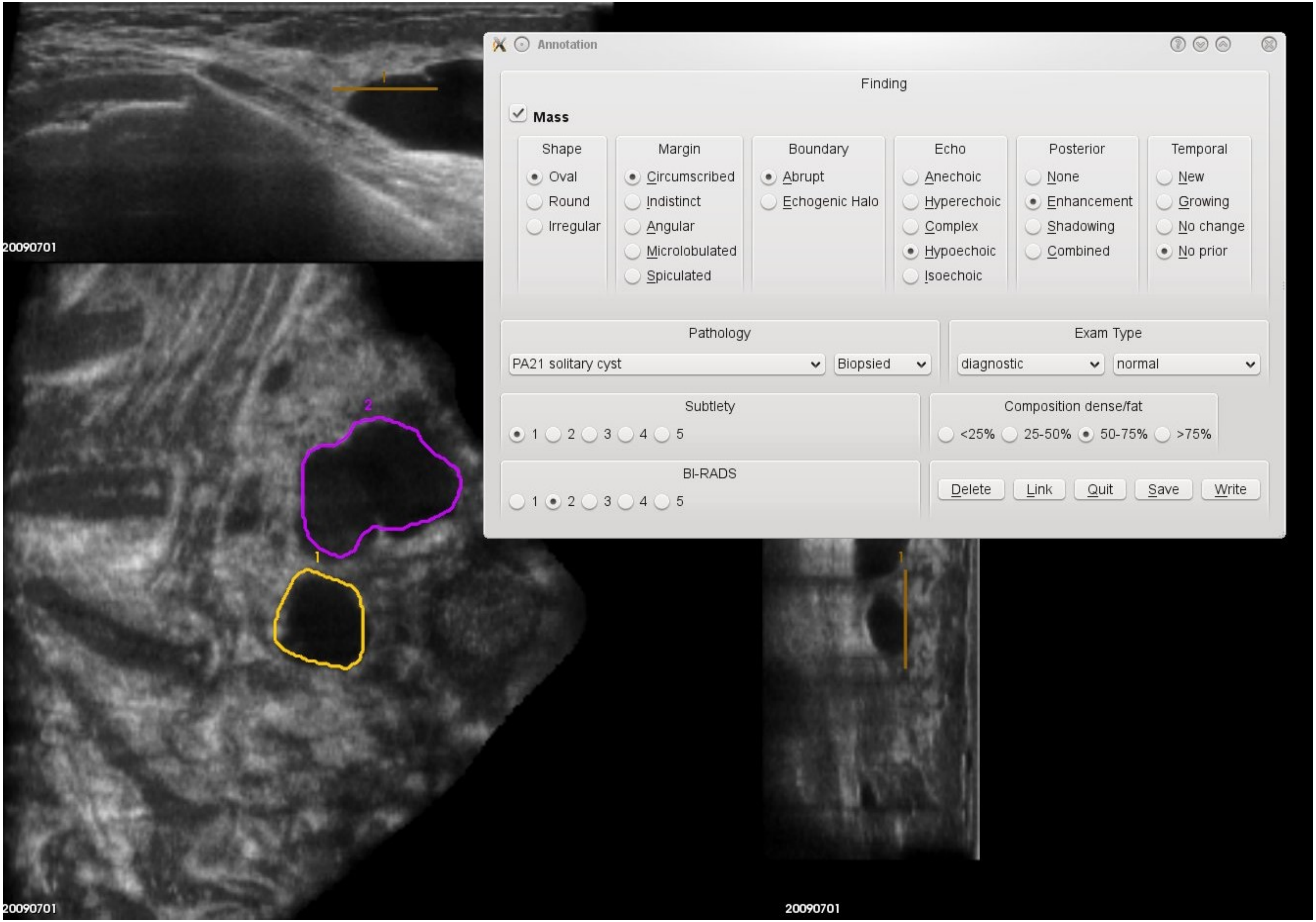
Pathology: PA02 ductal carcinoma, infiltrative | Biopsied

Exam Type: screening | normal

Subtlety: ○ 1 ● 2 ○ 3 ○ 4 ○ 5

Composition dense/fat: ○ <25% ○ 25-50% ○ 50-75% ○ >75%

BI-RADS: ○ 1 ○ 2 ○ 3 ○ 4 ● 5



Annotation

Finding

☑ **Mass**

Shape	Margin	Boundary	Echo	Posterior	Temporal
● Oval	● Circumscribed	● Abrupt	○ Anechoic	○ None	○ New
○ Round	○ Indistinct	○ Echogenic Halo	○ Hyperechoic	● Enhancement	○ Growing
○ Irregular	○ Angular		○ Complex	○ Shadowing	○ No change
	○ Microlobulated		● Hypoechoic	○ Combined	● No prior
	○ Spiculated		○ Isoechoic		

Pathology: PA21 solitary cyst — Biopsied

Exam Type: diagnostic — normal

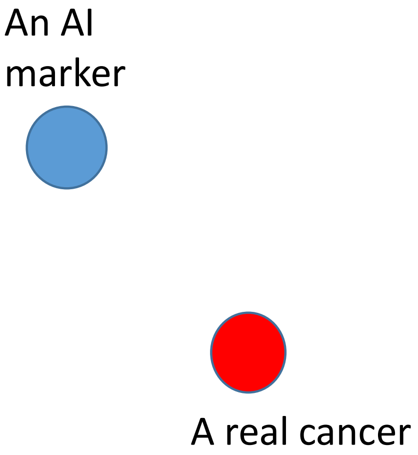
Subtlety: ● 1 ○ 2 ○ 3 ○ 4 ○ 5

Composition dense/fat: ○ <25% ○ 25-50% ● 50-75% ○ >75%

BI-RADS: ○ 1 ● 2 ○ 3 ○ 4 ○ 5

Ground Truth and Scoring Methods

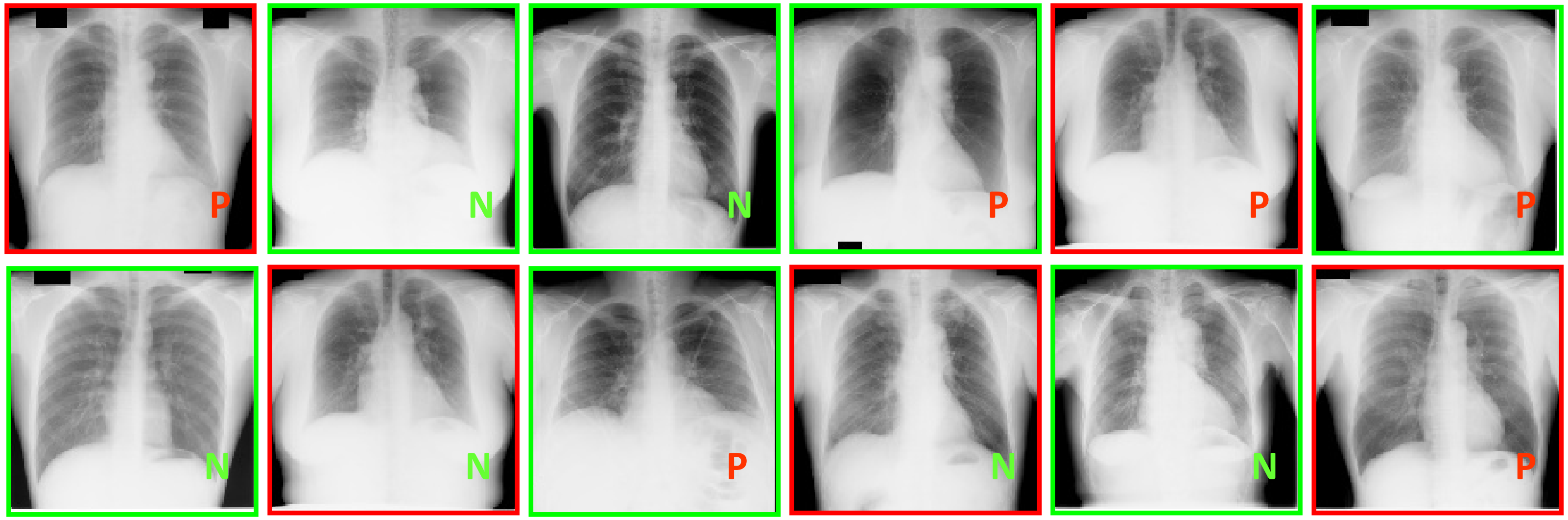
- Determine when an annotated lesion is detected using one of the following criteria:
  - Distance of true lesion to CAD finding < threshold
    - Center of mass
    - Distance to boundary
  - Center of CAD finding in annotated lesion segmentation
  - Use overlap between annotated lesion segmentation and CAD finding
- A false positive is a CAD finding which does not correspond to a true lesion



An AI marker

A real cancer

Evaluating CAD algorithms



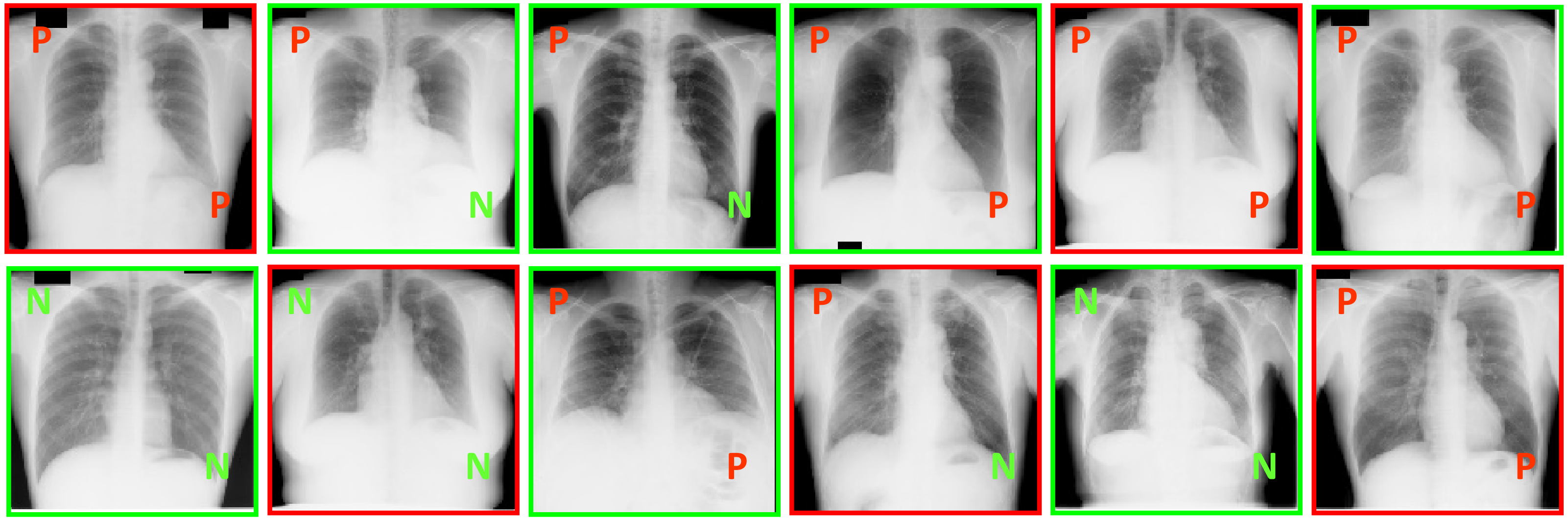
- P true positives (TP) = 3
- P false positives (FP) = 3
- N false negatives (FN) = 2
- N true negatives (TN) = 4

sensitivity = TP / (TP + FN)
= 3 / 5 = 0.6

specificity = TN / (TN + FP)
= 4 / 7 = 0.57

accuracy = (TP + TN) / (TP + TN + FP + FN)
= 7 / 12 = 0.58

Comparing 2 algorithms



algorithm 1

🟥 P = 3 🟥 N = 2

🟩 P = 3 🟩 N = 4

sensitivity = 3 / 5 = 0.6
specificity = 4 / 7 = 0.57
accuracy = 7 / 12 = 0.58

algorithm 2

🟥 P = 4 🟥 N = 1

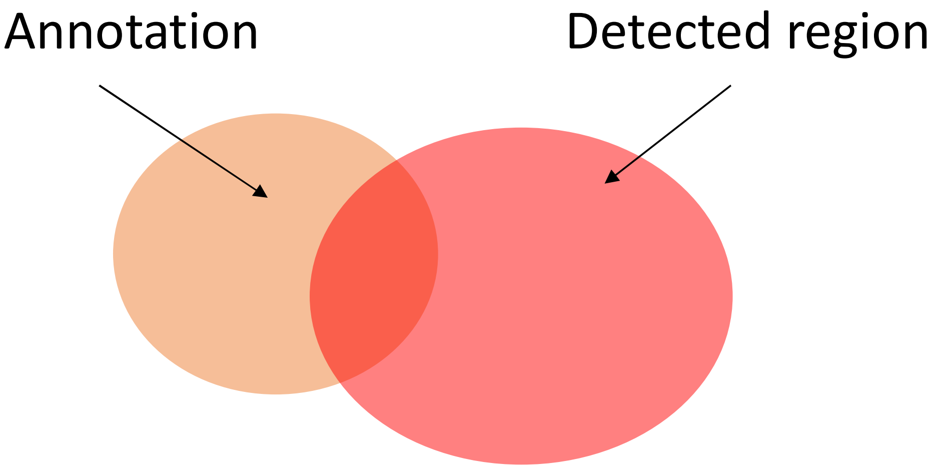
🟩 P = 5 🟩 N = 2

sensitivity = 4 / 5 = 0.8
specificity = 2 / 7 = 0.29
accuracy = 6 / 12 = 0.5

Which system is better?

Overlap measures

- Overlap = intersection / union = TP/(TP+FP+FN)
- Dice = 2TP / ((FN+TP)+(FP+TP))

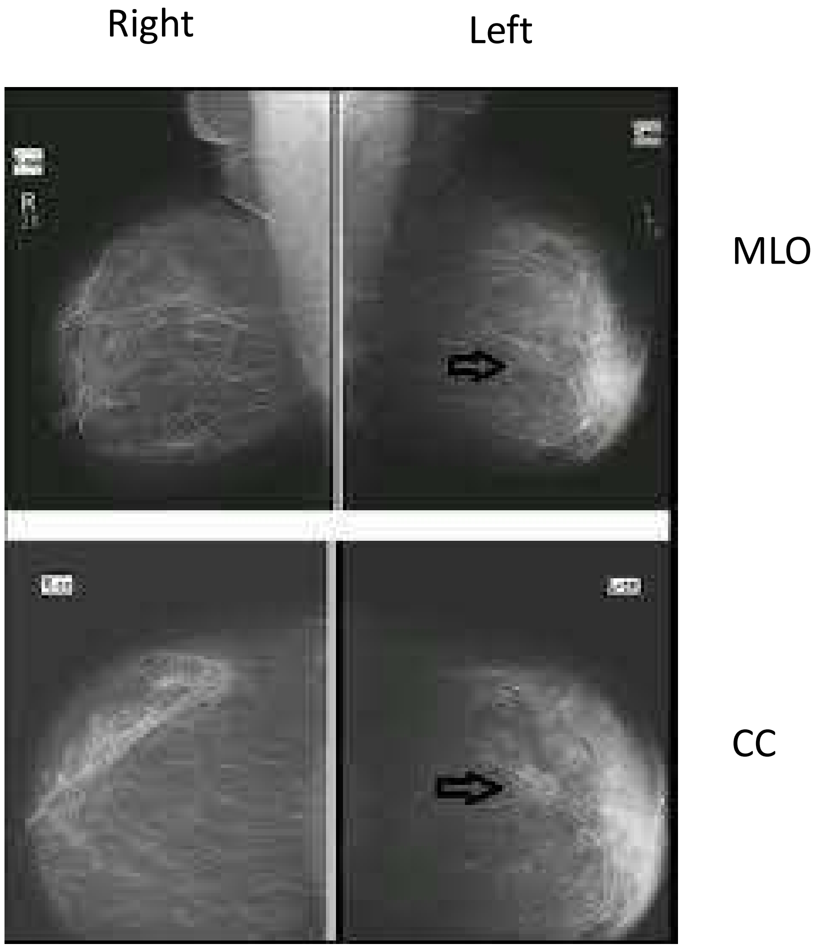


Overlap

- Overlap ranges from 0 (no overlap) to 1 (complete overlap)
- The background (TN) is disregarded in the overlap measure (useful for applications where most background is easy to classify)
- Small objects with irregular borders have lower overlap values than big compact objects

Scoring Methods and Multiple Views

- Lesion based analysis
  - Treat all ground truth regions independently
  - Number of true positives larger than number of lesions
- Case based analysis
  - Count a true positive if a lesion is detected in at least one view
  - Or count a true positive if a lesion is detected in all views



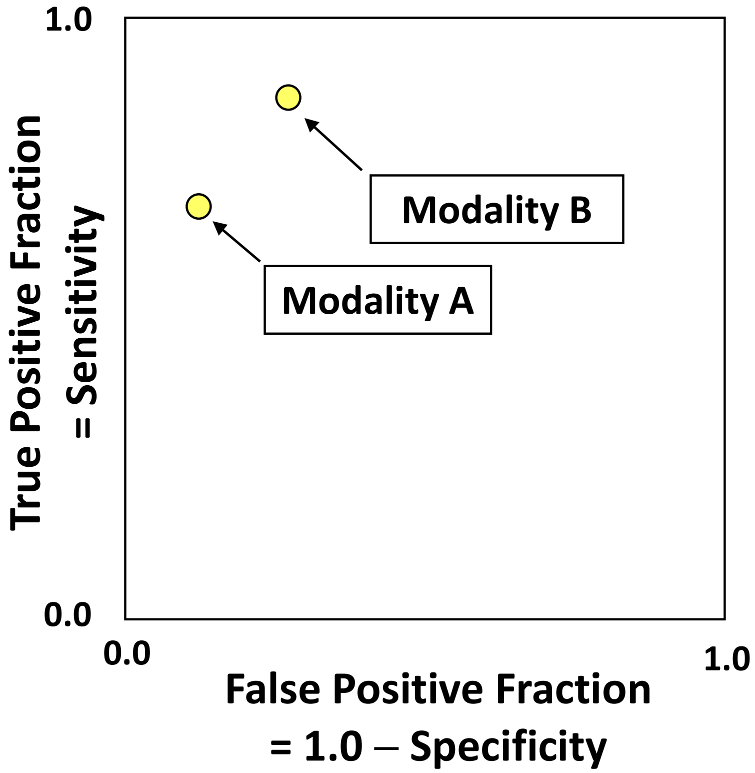
[Diagnostic Accuracy of Digital Mammography in the Detection of Breast Cancer]

ROC Analysis for binary classification

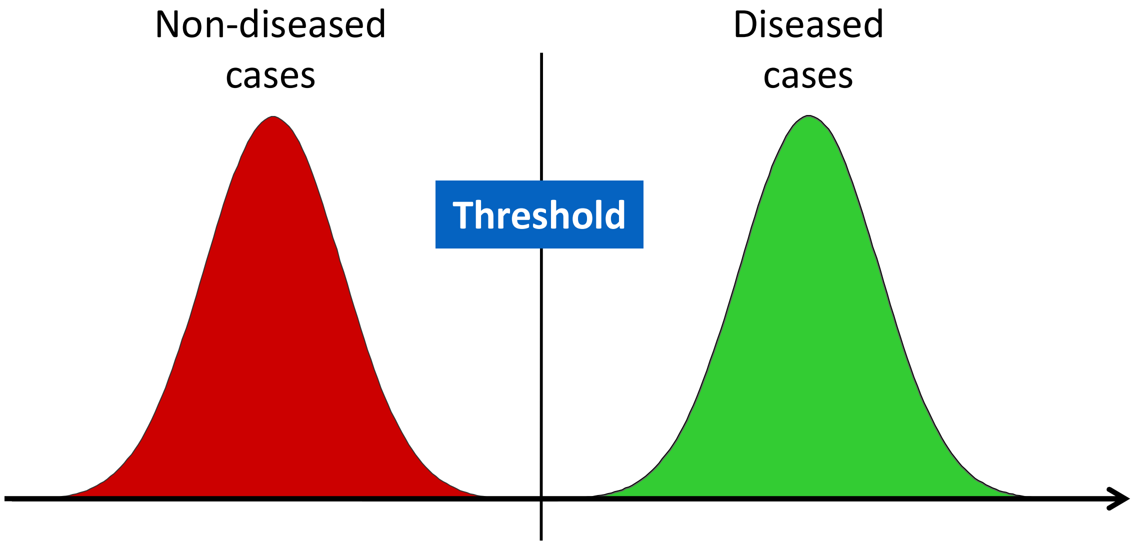
- Receiver Operating Characteristic (historic name from *radar* studies)
- Relative Operating Characteristic (psychology, psychophysics)
- Operating Characteristic (preferred by some)

Some of the following slides were obtained from a presentation by Bob Wagner

Dilemma: Which modality is better?



ROC Analysis



Non-diseased cases

Diseased cases

Threshold

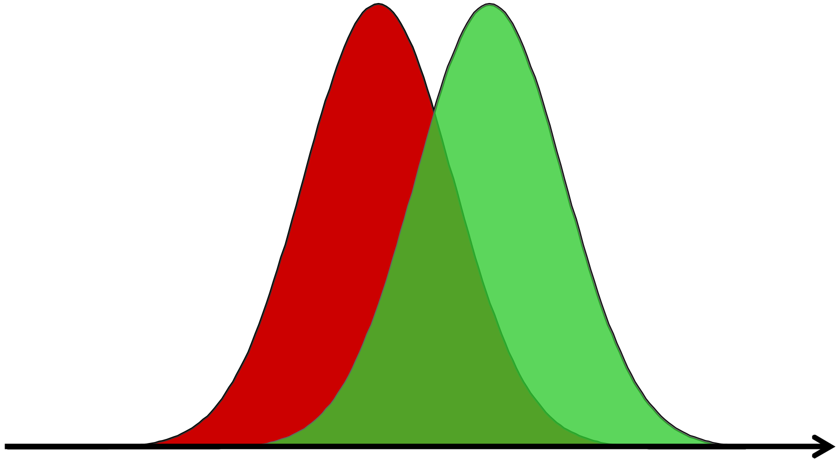
Test result value
or
subjective judgement of likelihood that case is diseased

ROC Analysis

more typically:

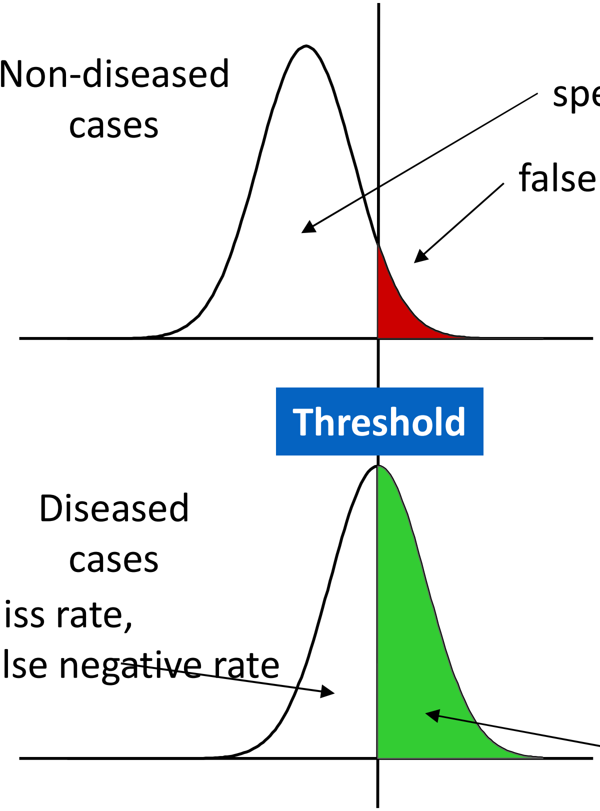
Non-diseased cases

Diseased cases



Test result value
or
subjective judgement of likelihood that case is diseased

ROC Analysis


Non-diseased cases

specificity

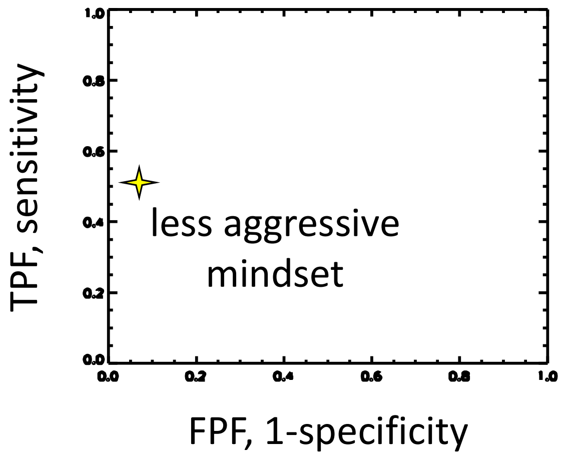
false alarm/positive rate, 1 - specificity

Threshold

Diseased cases

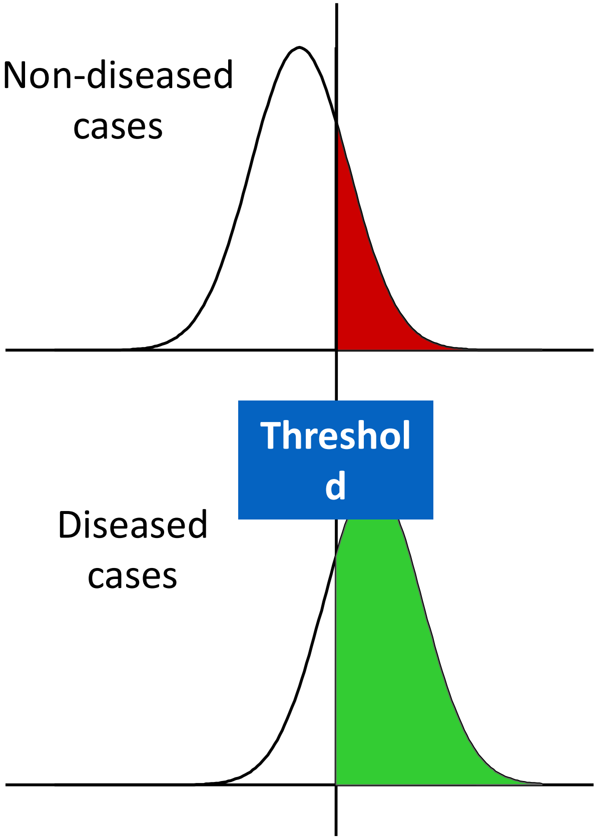
miss rate,
false negative rate

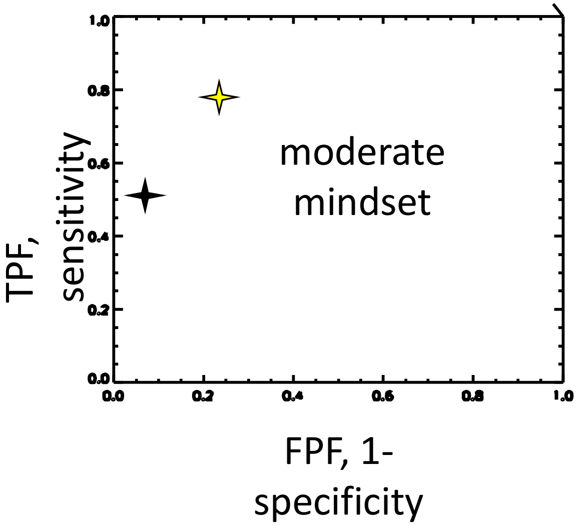
sensitivity, how much do you pick up



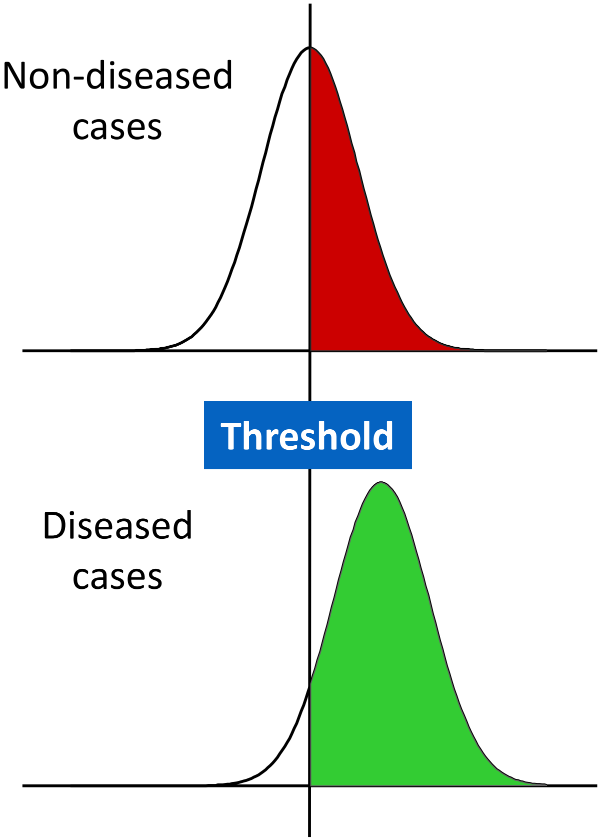
less aggressive mindset

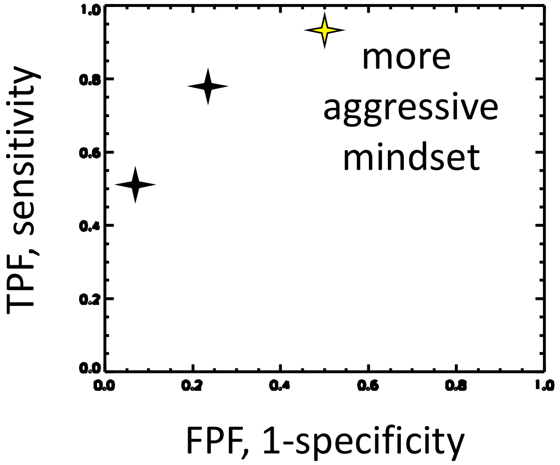
ROC Analysis



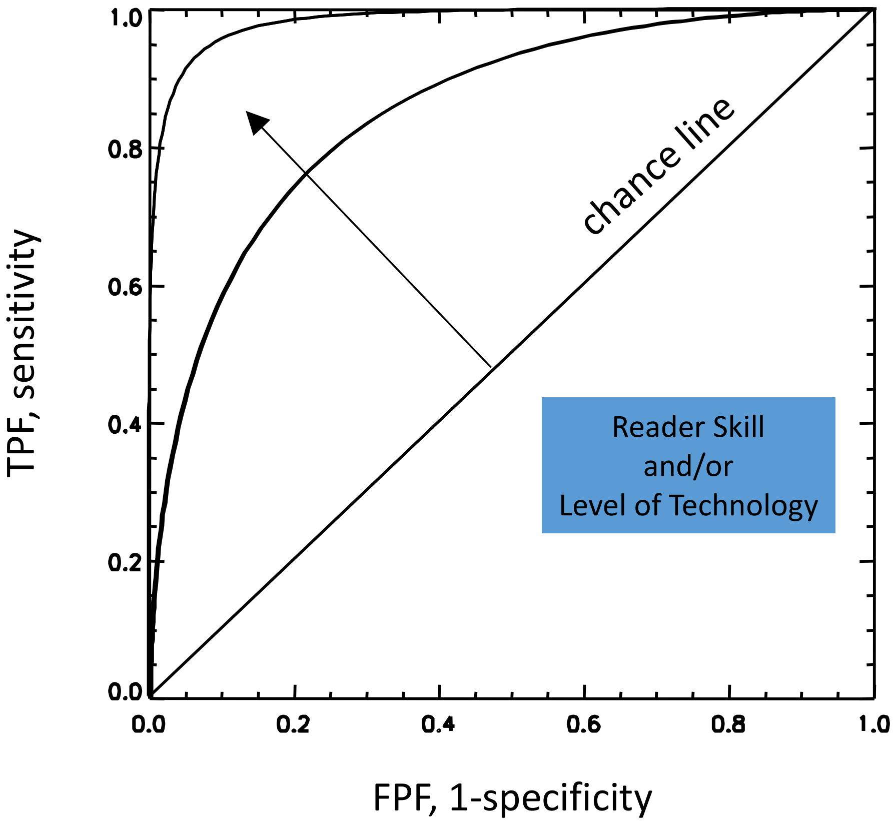


ROC Analysis





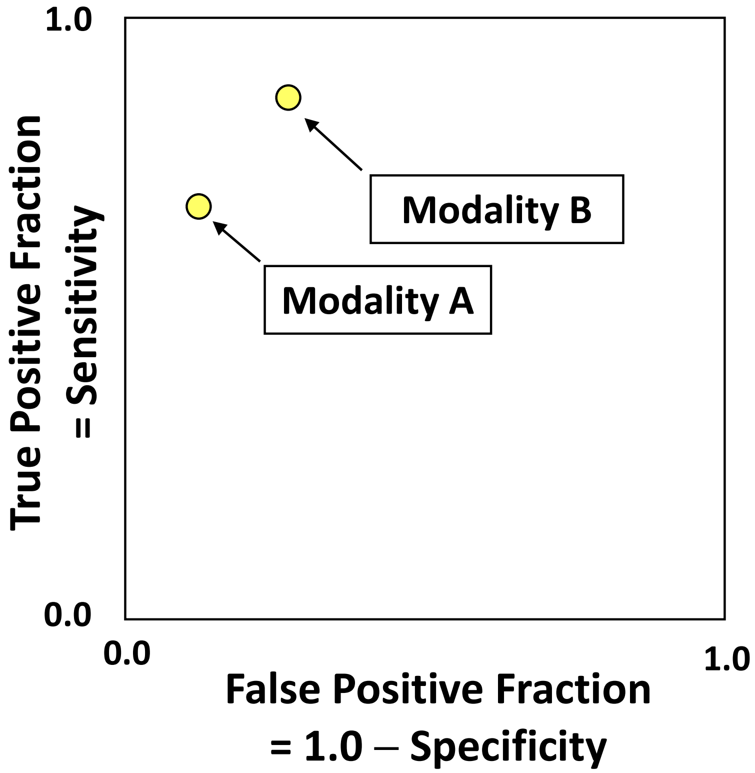
ROC Analysis



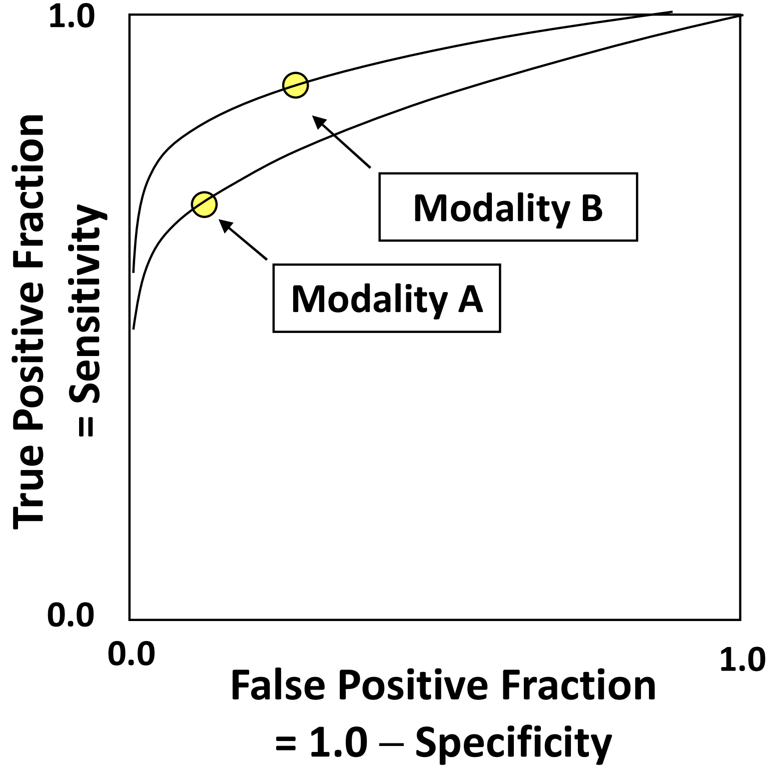
Test Statistic: Area under the ROC curve

Symbol: Az or AUC

Dilemma: Which modality is better?



The dilemma is resolved after ROCs are determined (<u>one</u> scenario):



Conclusion:

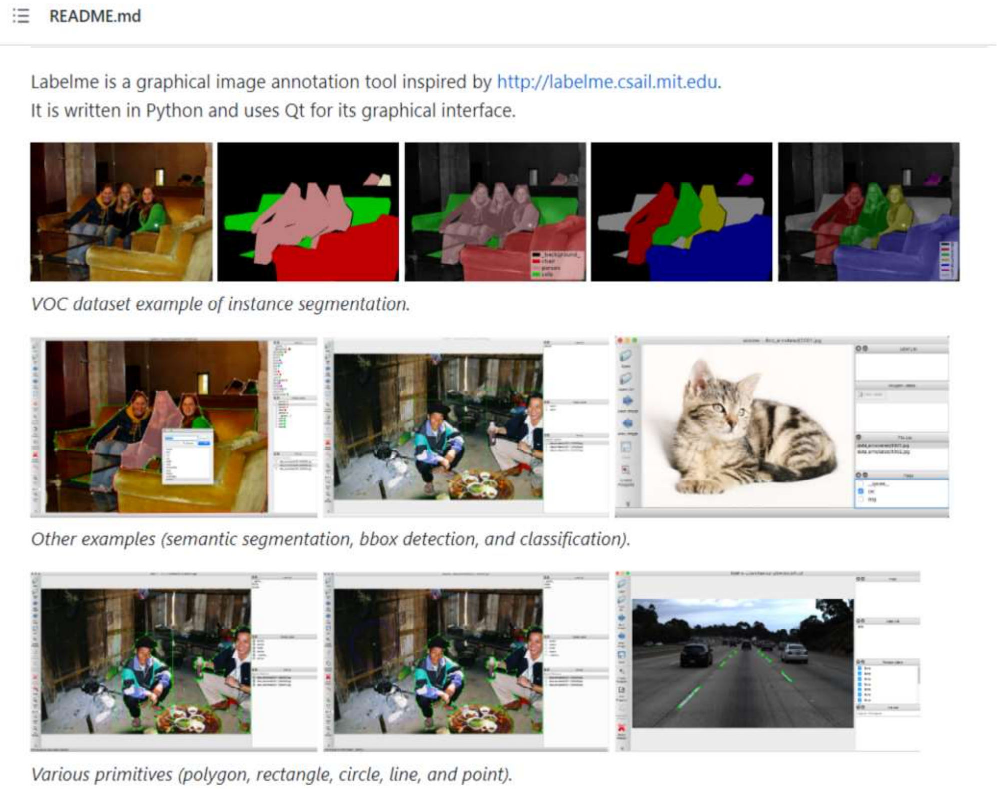
Modality B is better:

higher TPF at same FPF, or

lower FPF at same TPF

Label me

- https://github.com/wkentaro/labelme



README.md

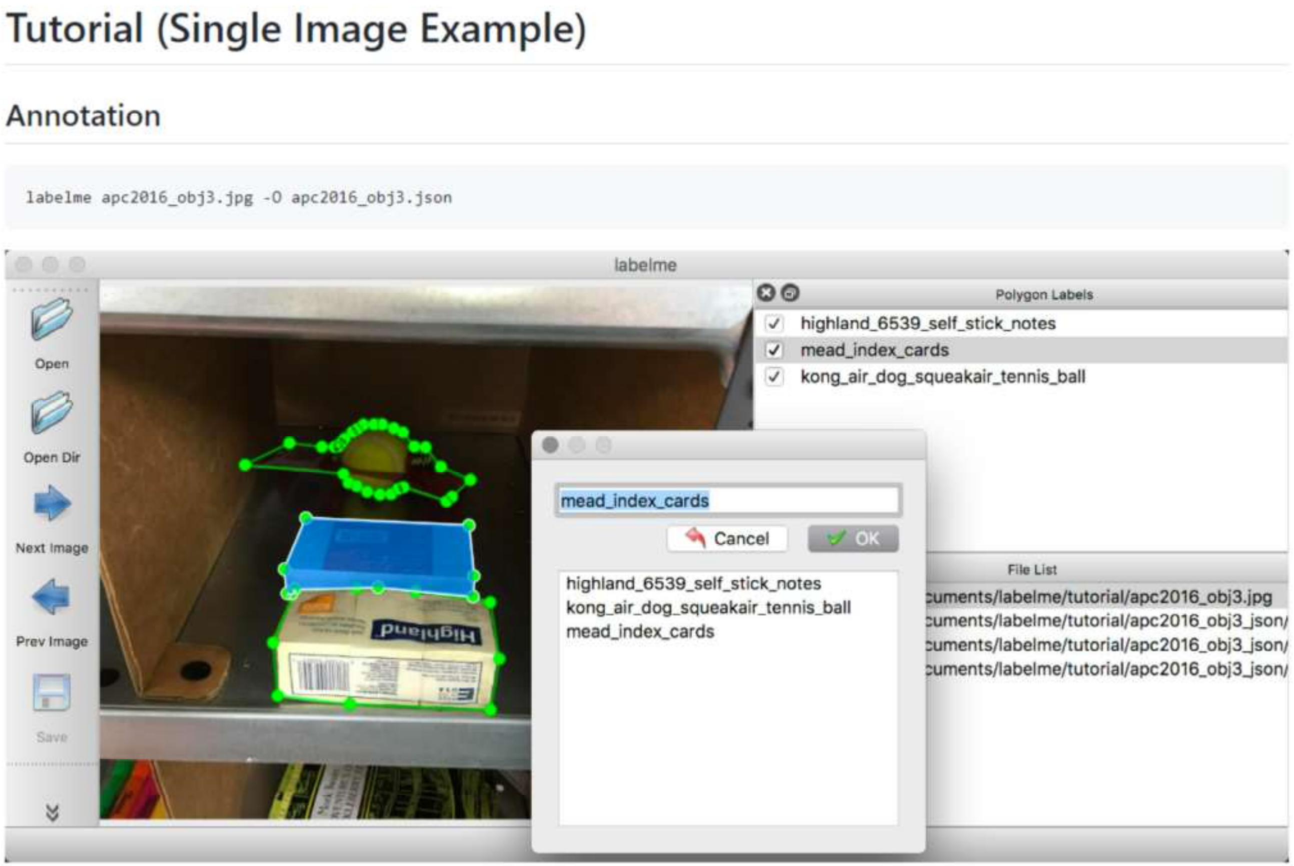
Labelme is a graphical image annotation tool inspired by http://labelme.csail.mit.edu.
It is written in Python and uses Qt for its graphical interface.

VOC dataset example of instance segmentation.

Other examples (semantic segmentation, bbox detection, and classification).

Various primitives (polygon, rectangle, circle, line, and point).

Labelme



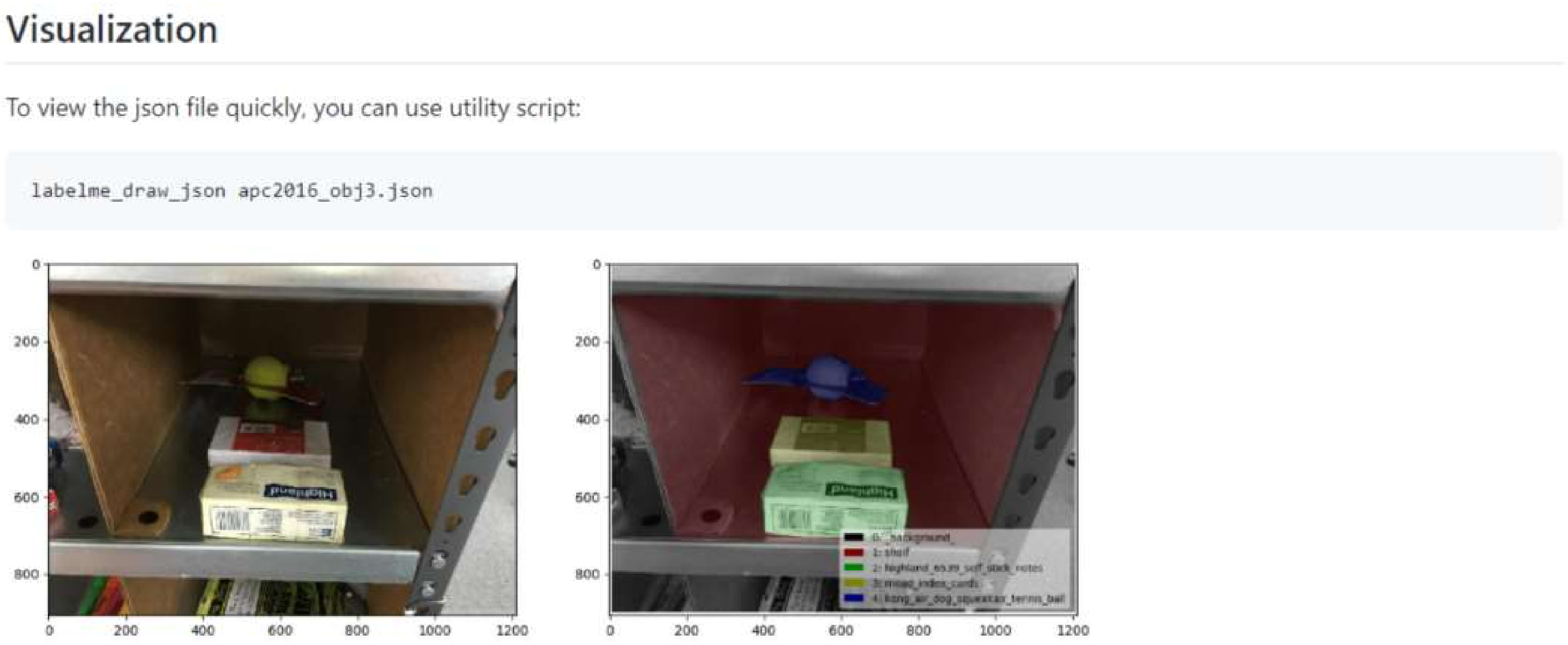
Tutorial (Single Image Example)

Annotation

```
labelme apc2016_obj3.jpg -O apc2016_obj3.json
```

labelme
Open
Open Dir
Next Image
Prev Image
Save
Polygon Labels
☑ highland_6539_self_stick_notes
☑ mead_index_cards
☑ kong_air_dog_squeakair_tennis_ball
mead_index_cards
Cancel
OK
highland_6539_self_stick_notes
kong_air_dog_squeakair_tennis_ball
mead_index_cards
File List
cuments/labelme/tutorial/apc2016_obj3.jpg
cuments/labelme/tutorial/apc2016_obj3_json/
cuments/labelme/tutorial/apc2016_obj3_json/
cuments/labelme/tutorial/apc2016_obj3_json/
Highland

Labelme

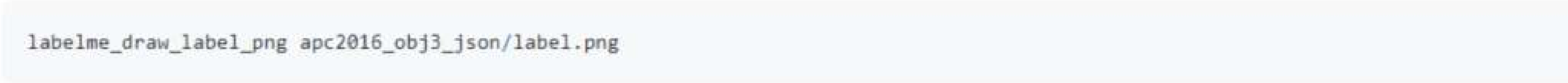


Visualization

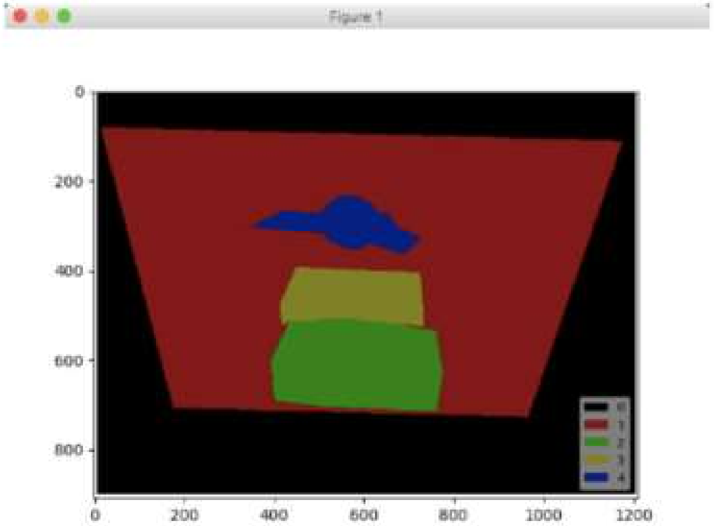
To view the json file quickly, you can use utility script:

```
labelme_draw_json apc2016_obj3.json
```

Labelme



```
labelme_draw_label_png apc2016_obj3_json/label.png
```



Supervised learning

- Regression: the target variable (Y) has continuous value.
  - Example- house price prediction

- Classification: the target variable (Y) has discrete values such as Yes or No, 0 or 1 and many more.
  - Example- Credit Scoring, Spam Filtering

What about cancer detection?

Break

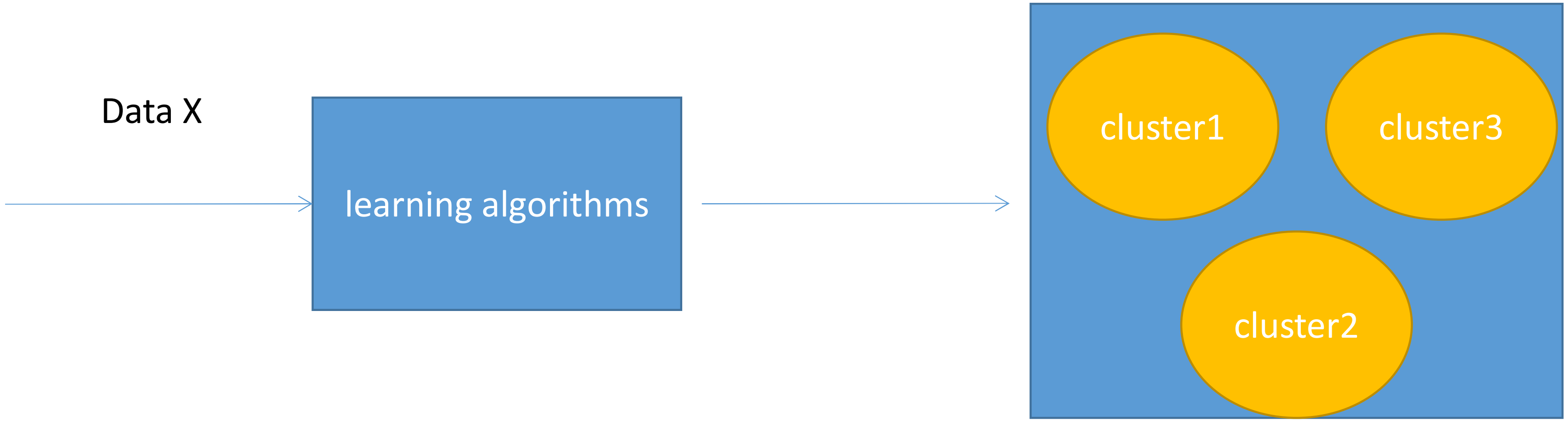
Unsupervised Learning

- Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don’t necessarily know the effect of the variables.
- We can derive this structure by clustering the data based on relationships among the variables in the data.
- With unsupervised learning, there is **no feedback** based on the prediction results.

Unsupervised learning

- Take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on.
- Non-clustering: The "Cocktail Party Algorithm", allows you to find structure in a chaotic environment. (i.e. identifying individual voices and music from a mesh of sounds at a cocktail party).

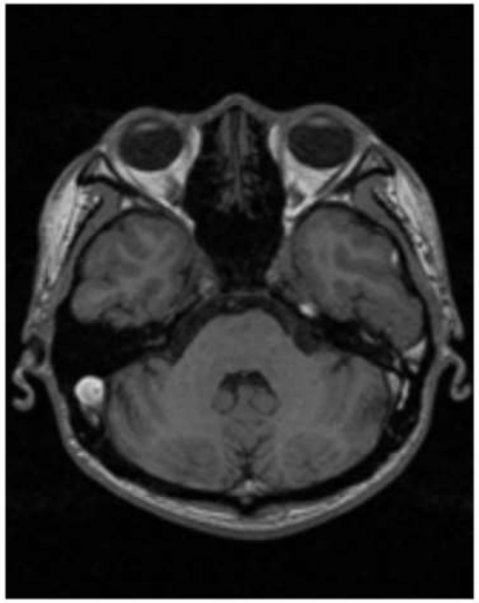
Unsupervised learning



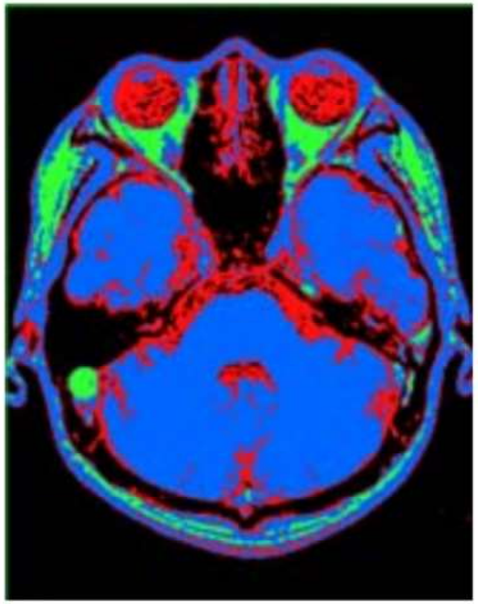
Unsupervised learning

- Build a model to predict the quality of my dish based on the quantity of sugar, milk, salt. Unfortunately, that didn’t go well. But Still, hope I will build an effective model on that.

Medical image segmentation



brain MRI



segmentation

Break up the brain MRI into meaningful or perceptually similar regions

why we need segmentation?

Clustering

- Clustering
  - Basic idea: Group together similar instances
    - Group emails or search results
    - Grouping the behaviors of animals

How to do clustering for coffees?

Your own examples?

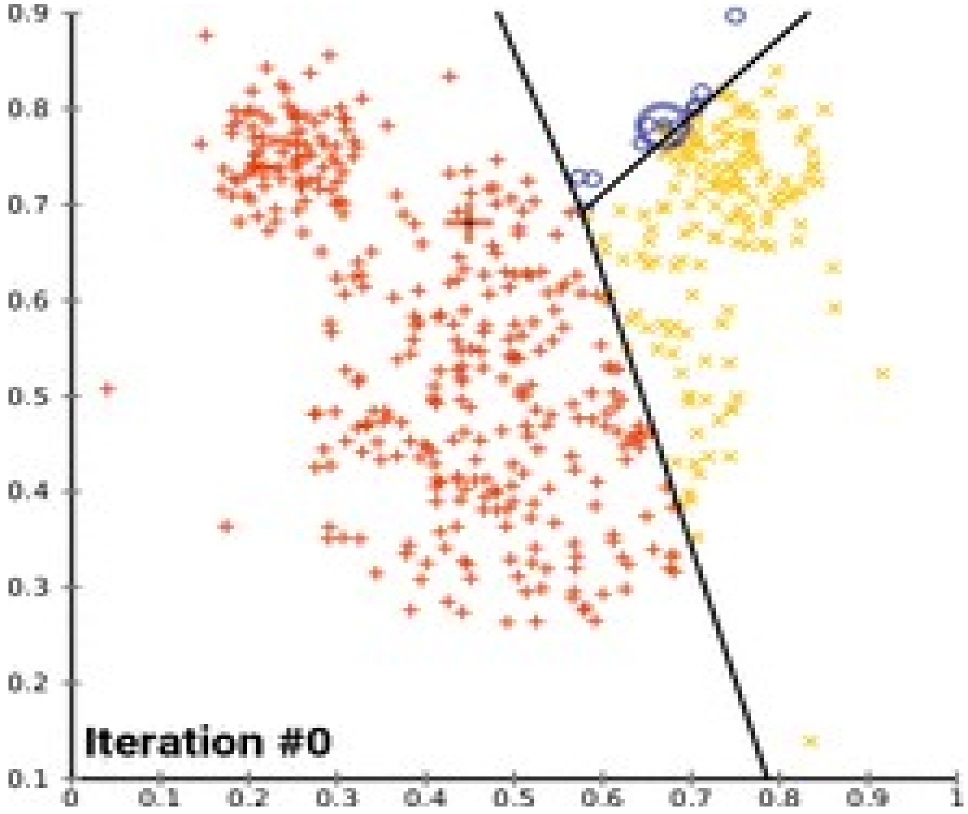
K means clustering

- Given a set of observations (x1, x2, ..., xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, ..., Sk}
- Minimize the within-cluster sum of squares (WCSS) (i.e. variance).

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

K means clustering

- An iterative clustering approach
  - Step 1: Pick K random points as cluster centers
  - Step 2: Assign data points to closest cluster center
  - Step 3: Recompute the cluster center its assigned points
  - Stop when no points' assignments change



Wikipedia

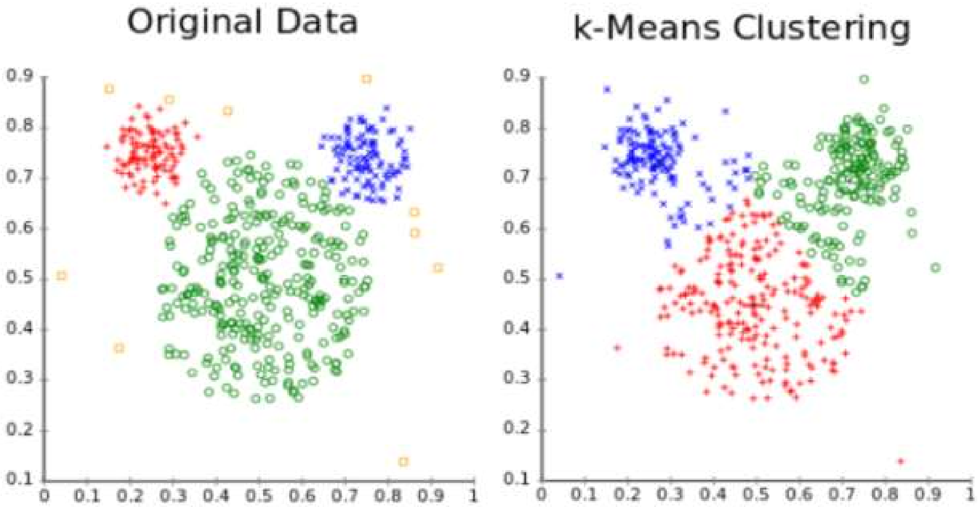
Drawbacks

- Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.

- The number of clusters k is an input parameter: an inappropriate choice of k may yield poor results.

- Convergence to a local minimum may produce counterintuitive ("wrong") results.

k-means to leads to bad results here



wikipedia

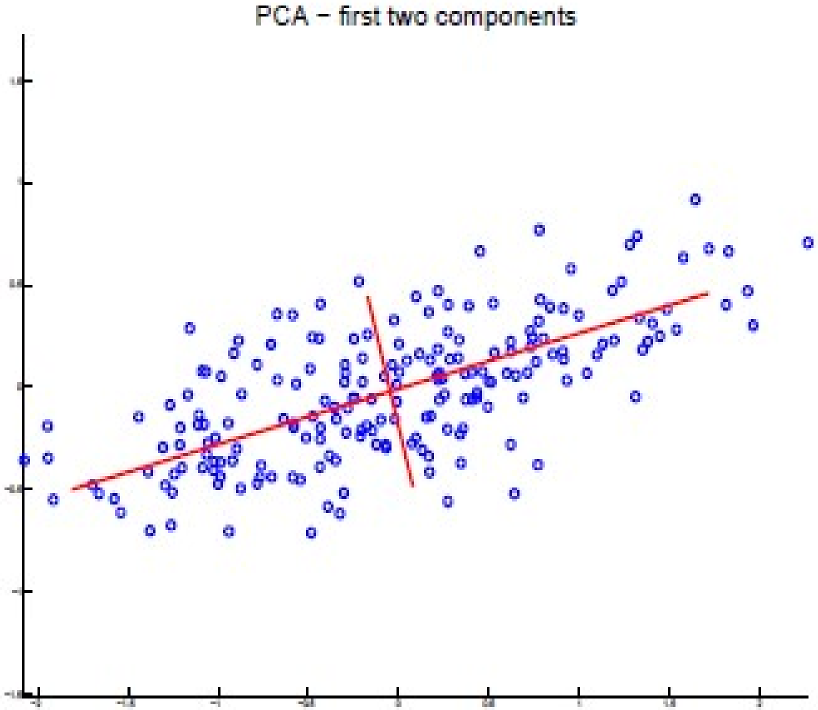
K means clustering

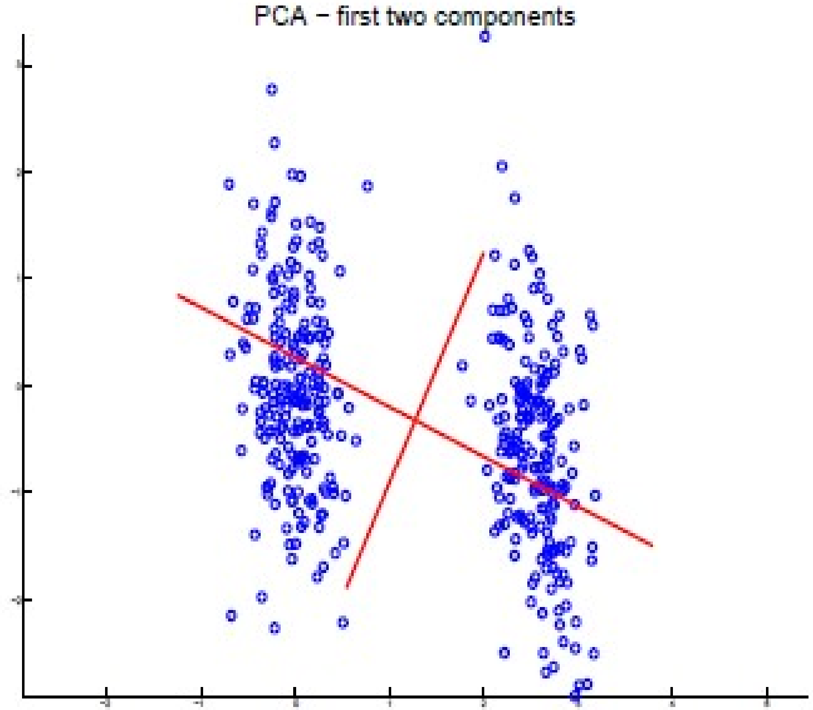
- Is Kmeans deterministic ?

- The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results.

Dimensionality reduction

Principle component analysis





Supervised vs Unsupervised

Difference?

Supervised learning	**Unsupervised learning**
Labeled data $R^{N \times D}$ -> $R^{N \times M}$	Just data -$R^{N \times D}$
Goal: learn a function to map X -> y	Goal: learn some underlying hidden structures
Examples: Classification Regression Segmentation	Examples: Clustering Dimensionality reduction

Break

Reinforcement learning

- Consider teaching a dog a new trick. You cannot tell it what to do, but you can reward/punish it if it does the right/wrong thing. It has to figure out what it did that made it get the reward/punishment.

- Teaching a game bot to perform better and better at a game by learning and adapting to the new situation of the game.

- We can use a similar method to train computers to do many tasks, such as playing backgammon or chess, scheduling jobs, and controlling robot limbs.

Reinforcement learning

- What does reinforcement mean?

- Reinforcement learning is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

Reinforcement Learning

- How can an agent learn behaviors when it doesn’t have a teacher to tell it how to perform?
  - The agent has a task to perform
  - It takes some actions in the world
  - At some later point, it gets feedback telling it how well it did on performing the task
  - The agent performs the same task over and over again
- This problem is called **reinforcement learning:**
  - The agent gets *positive reinforcement* for tasks done well
  - The agent gets *negative reinforcement* for tasks done poorly

Reinforcement Learning

- The goal is to get the agent to act in the world so as to maximize its rewards
- The agent has to figure out what it did that made it get the reward/punishment
  - This is known as the **credit assignment** problem
- Reinforcement learning approaches can be used to train computers to do many tasks
  - chess playing
  - shop scheduling
  - controlling robot limbs

Formalization

- Given:
  - a state space S
  - a set of actions a_1, …, a_k
  - reward value at the end of each trial (may be positive or negative)
- Output:
  - a mapping from states to actions

example: Alex (driving agent)
state: configuration of the car
learn a steering action for each state

Reactive Agent Algorithm

Accessible or
observable state

```
Repeat:
    s ← sensed state
    If s is terminal then exit
    a ← choose action (given s)
    Perform a
```

A policy is a complete mapping from states to actions

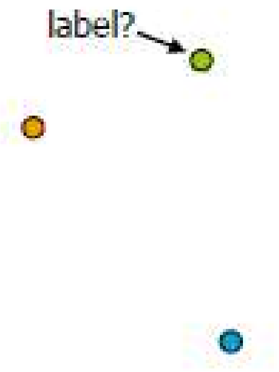
Learn policy directly– function mapping from states to actions

AlphaGo

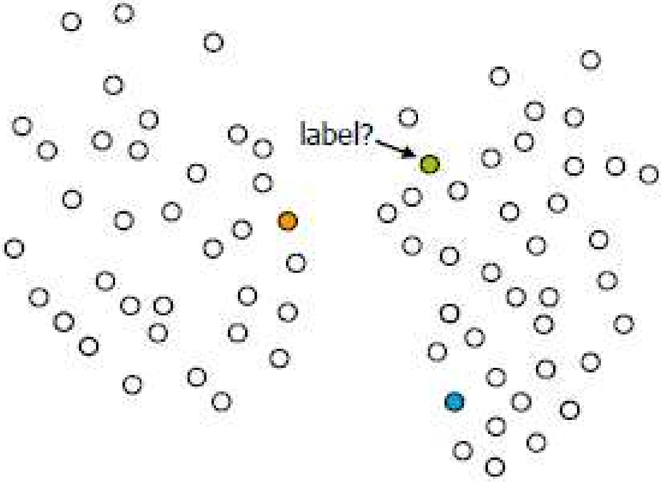
- Full story

- Trailer

Semi-supervised learning

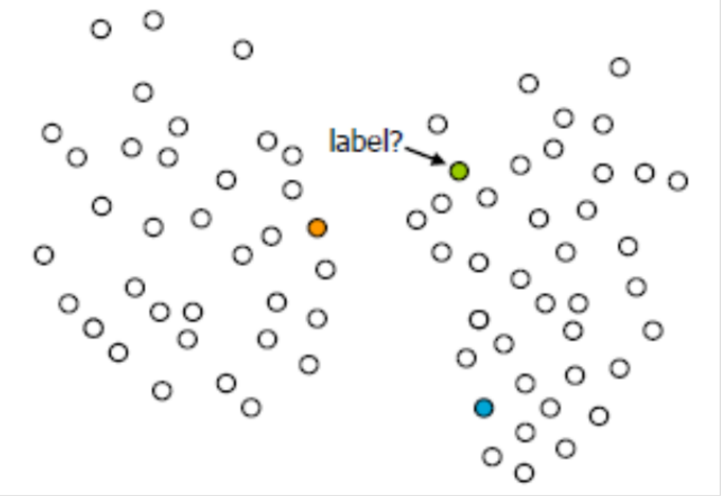


Semi-supervised learning



Semi-supervised learning

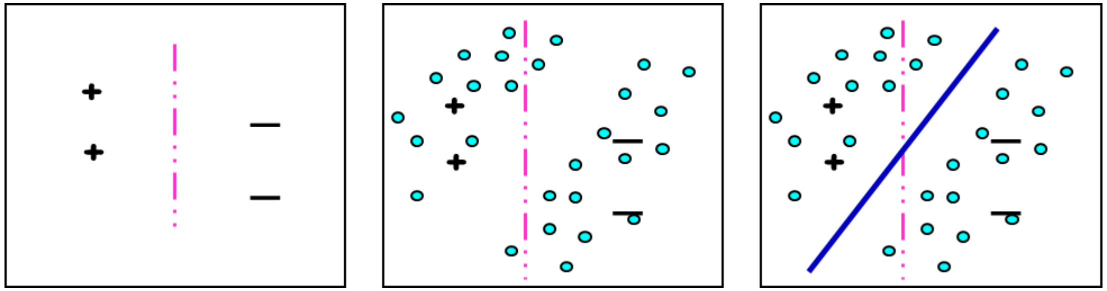
- How to improve decision rule by means of unlabeled data?



- Why?
    - Unlabeled data is easy to obtain …
    - … while labeled data is highly expensive

Semi-supervised learning

- **Methods:**
  - Self-training/self-learning:
    - Train on labeled data
    - Classify unlabeled data
    - Pick points classified e.g. above certain confidence
    - Include in training set and retrain.
  - Semi-supervised (transductive) SVM
    - Suppose we believe target separator goes through low density regions of the space/large margin.



Multiple instance learning

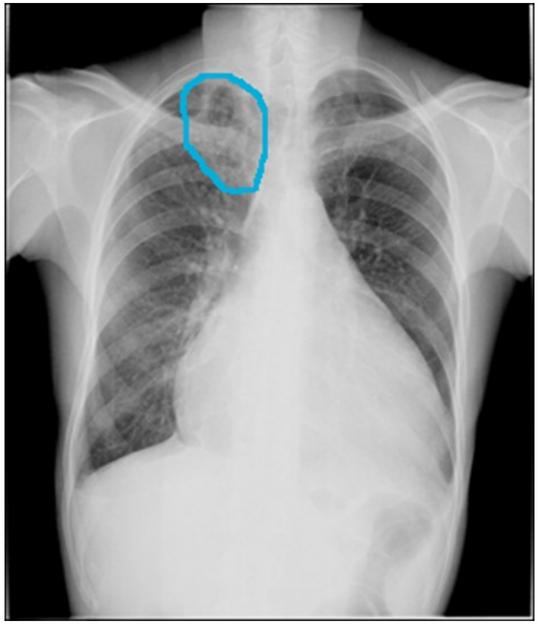
- Problem: Inaccurate/unreliable labels at the pixel level, but accurate/reliable labels at the image level

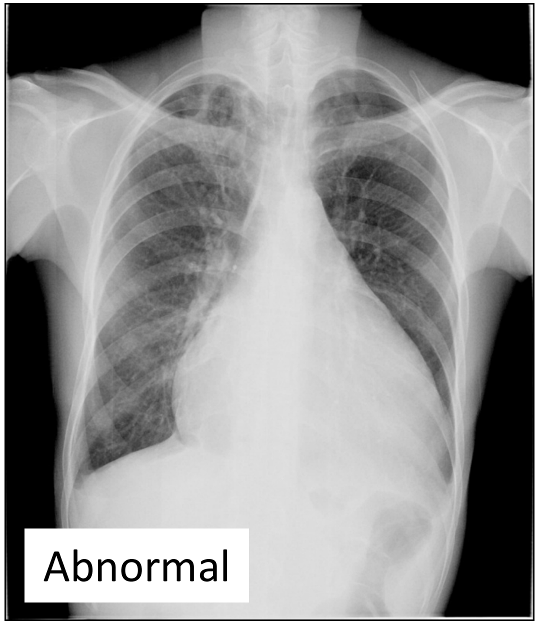
- Weakly supervised learning

- MIL definition
    - The labels of individual instances are unknown, but the labels of groups of instances (bags) are actually known

- MIL for medical imaging
    - Bags = images
    - Instances = regions, pixels or superpixels.

Multiple instance learning

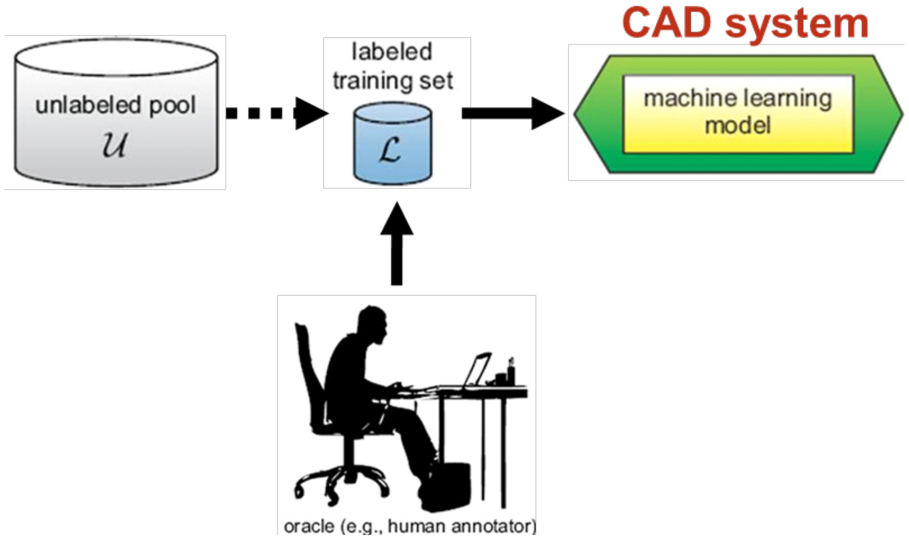




- Advantages:
    - Does not require detailed class label information. It is enough with case- or image-level labels
- Disadvantages:
    - The ambiguity is higher than with the previous approach

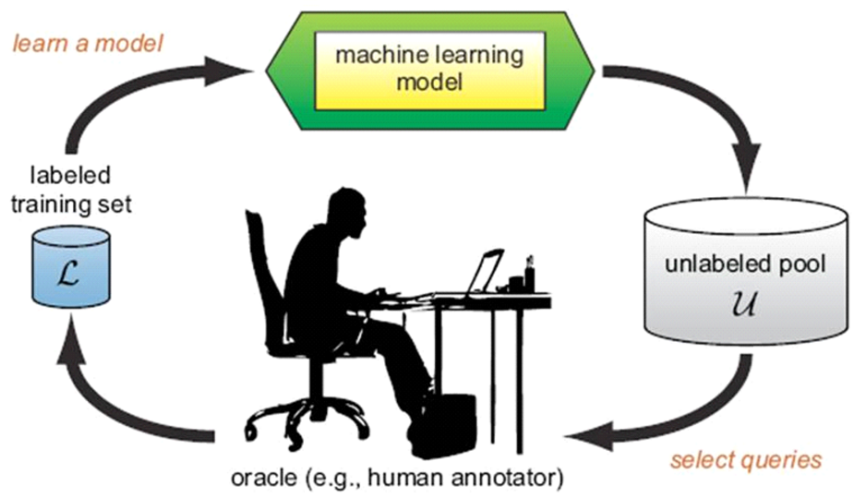
Active Learning

- How to create an efficient training set from unlabeled data?

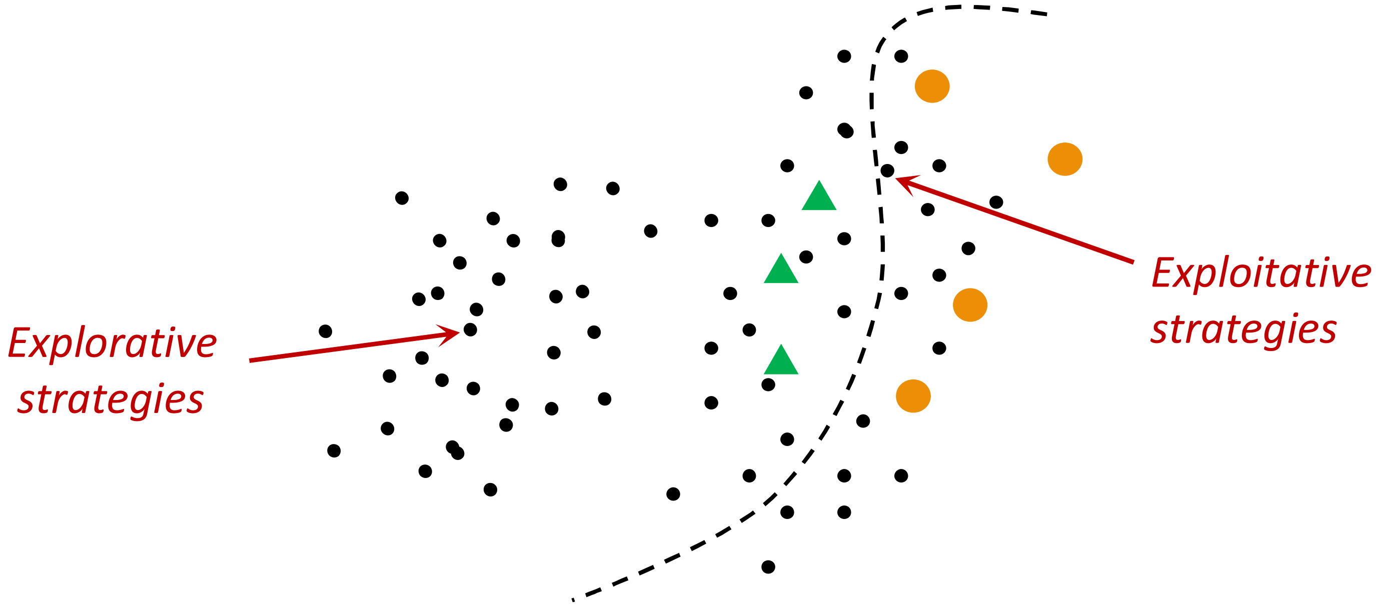


Active Learning

- Active learning allows dynamic selection of a small representative training set
- Only this training set is presented to the expert to be labeled
  - Optimum classification accuracy
  - Efficient use of the expert time



Active Learning



Practice

- Whats the difference between semi-supervised learning and weakly supervised learning?

- Find a dataset and perform clustering

  - For example: https://www.kaggle.com/code/heeraldedhia/kmeans-clustering-for-customer-data