Fall 2017

INTRODUCTION TO MACHINE LEARNING FOR MEDICINE

Carla E. Brodley
Professor & Dean
College of Computer and Information Science
Northeastern University

440 Huntington Avenue . 202 West Village H . Boston, MA 02115 . T 617.373.2462 . ccis.northeastern.edu

WHAT IS MACHINE LEARNING/DATA MINING?

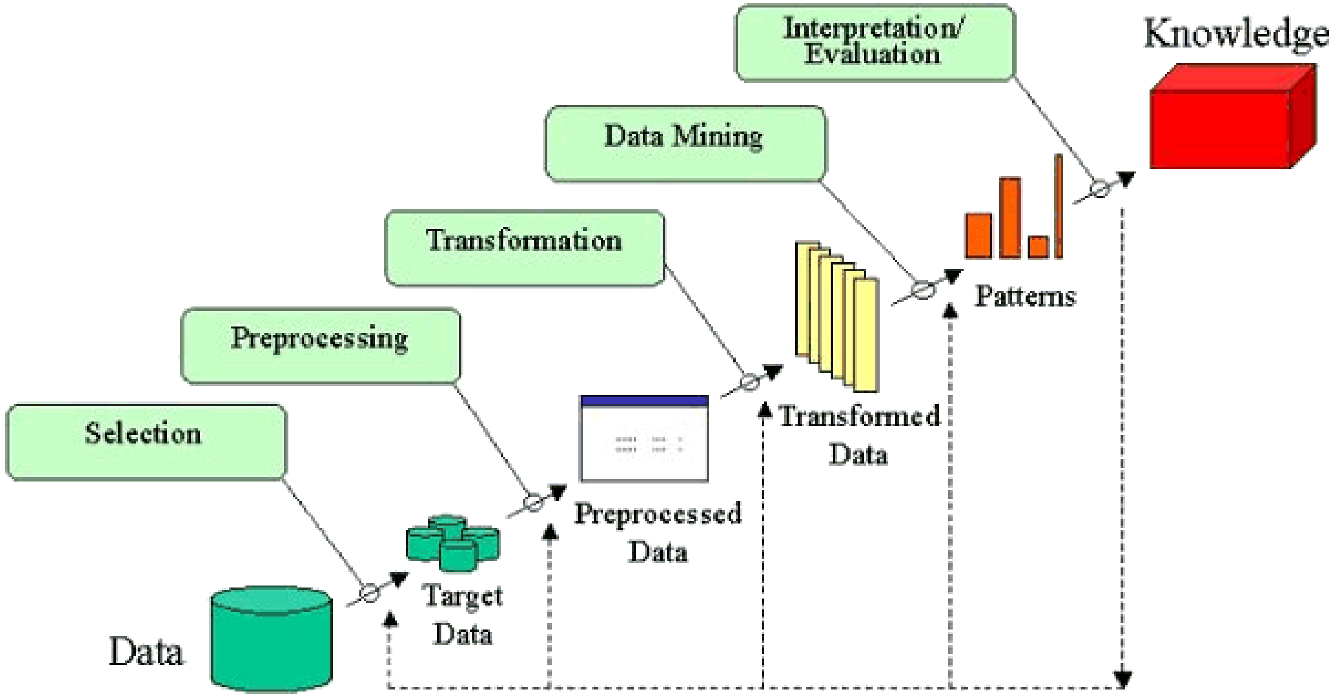


Figure is from Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy. Advances in Knowledge Discovery and Data Mining, 1996; image found at: www2.cs.uregina.ca/~dbd/cs831/notes/kdd/kdd.gif

Fall 2017

SUPERVISED LEARNING

SUPERVISED LEARNING

Given: example – $< x_1, x_2, \ldots x_n, f(x_1, x_2, \ldots x_n) >$ for some unknown function f

Find: A good approximation to f

Goal: Apply f to previously unseen data

Example Applications:

- **Regression:** f is a continuous variable (e.g., predicting EDSS for MS patients)
- **Classification:** f is a discrete variable (e.g., predicting whether a patient has unilateral or bilateral Meniere's)

CLASSIFICATION EXAMPLE:
CITATION SCREENING FOR SYSTEMATIC REVIEWS

- **Systematic review:** an exhaustive assessment of all the published medical evidence regarding a precise clinical question

  - e.g., "Is aspirin better than leeches in inducing more than 50% relief in patients with tension headaches?"

- **Must find all relevant studies**

TYPICAL WORKFLOW

26M — PubMed

SEARCH

10,000 — Potentially eligible

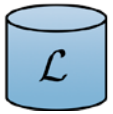
SCREEN

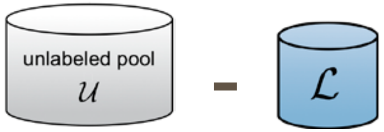
500 — Relevant

CITATION SCREENING

Doctors read these. They’d rather be doing something else.

GENERATING TRAINING DATA FOR SUPERVISED LEARNING

Induce (train) a classifier *C* over 

Apply *C* to unlabeled examples 

A DETOUR INTO TEXT ENCODING

- Classification algorithms operate on vectors
- Feature space: an n-dimensional representation

A ‘bag-of-words’ example:

S_1= “Boston drivers are frequently aggressive”

S_2= “The Boston Red Sox frequently hit line drives”

TEXT ENCODING: STOP WORDS

S_1= “Boston drivers ~~are~~ frequently aggressive”
S_2= “~~The~~ Boston Red Sox frequently hit line drives”

TEXT ENCODING: LOWERCASING

S_1= "boston drivers ~~are~~ frequently aggressive"
S_2= "~~The~~ boston red sox frequently hit line drives"

TEXT ENCODING: STEMMING

S_1= "boston drive ~~are~~ frequent aggressive"
S_2= "~~The~~ boston red sox frequent hit line drive"

TEXT ENCODING: VOILA

	hit	red	sox	line	boston	frequent	drive	aggressive
$S_1 =$	0	0	0	0	1	1	1	1
$S_2 =$	1	1	1	1	1	1	1	0

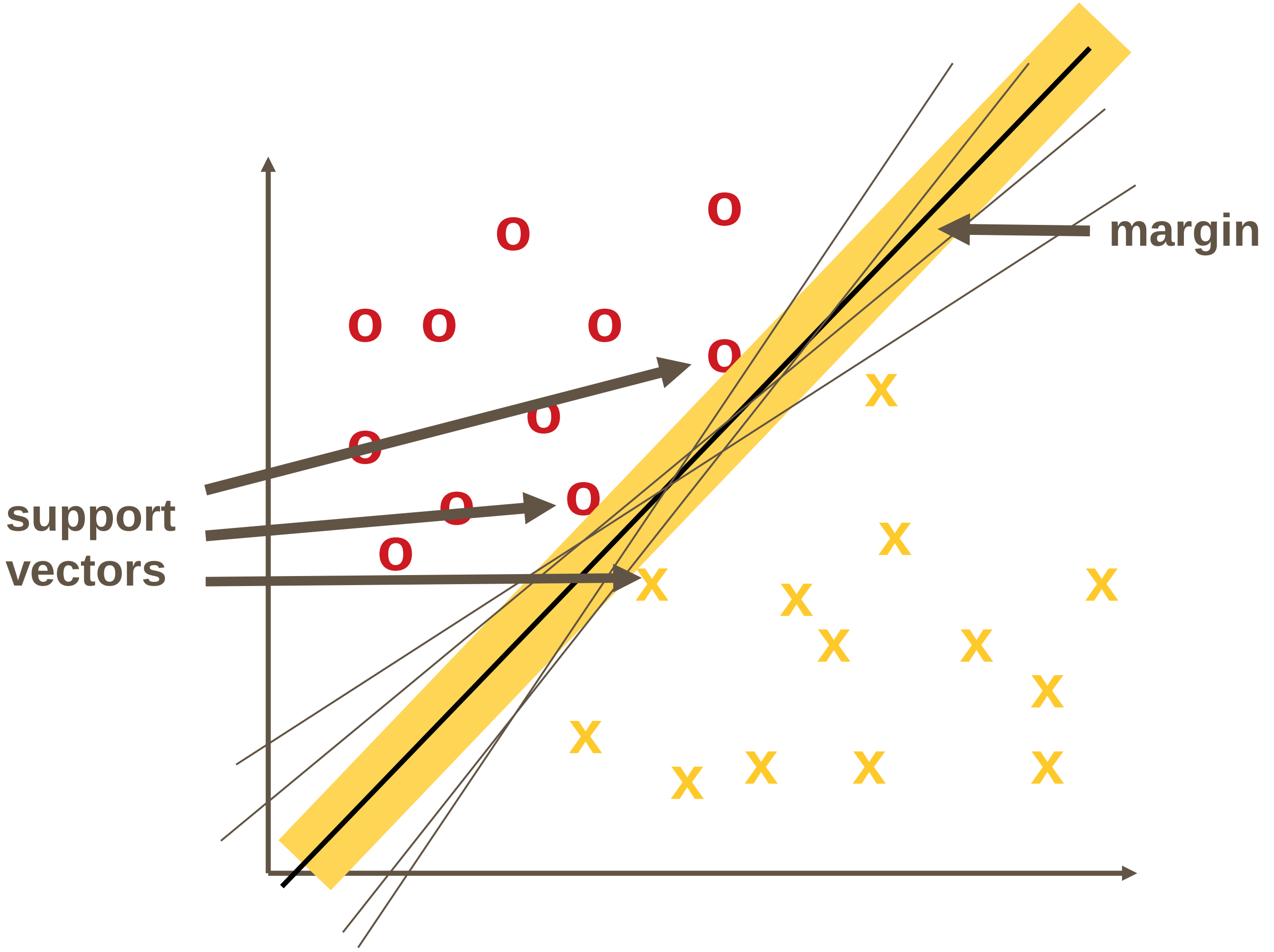
A new sentence, S_3, comes along:

"I hate the red sox."

Which sentence is it most similar to?

$S_3 =$	0	1	1	0	0	0	0	0

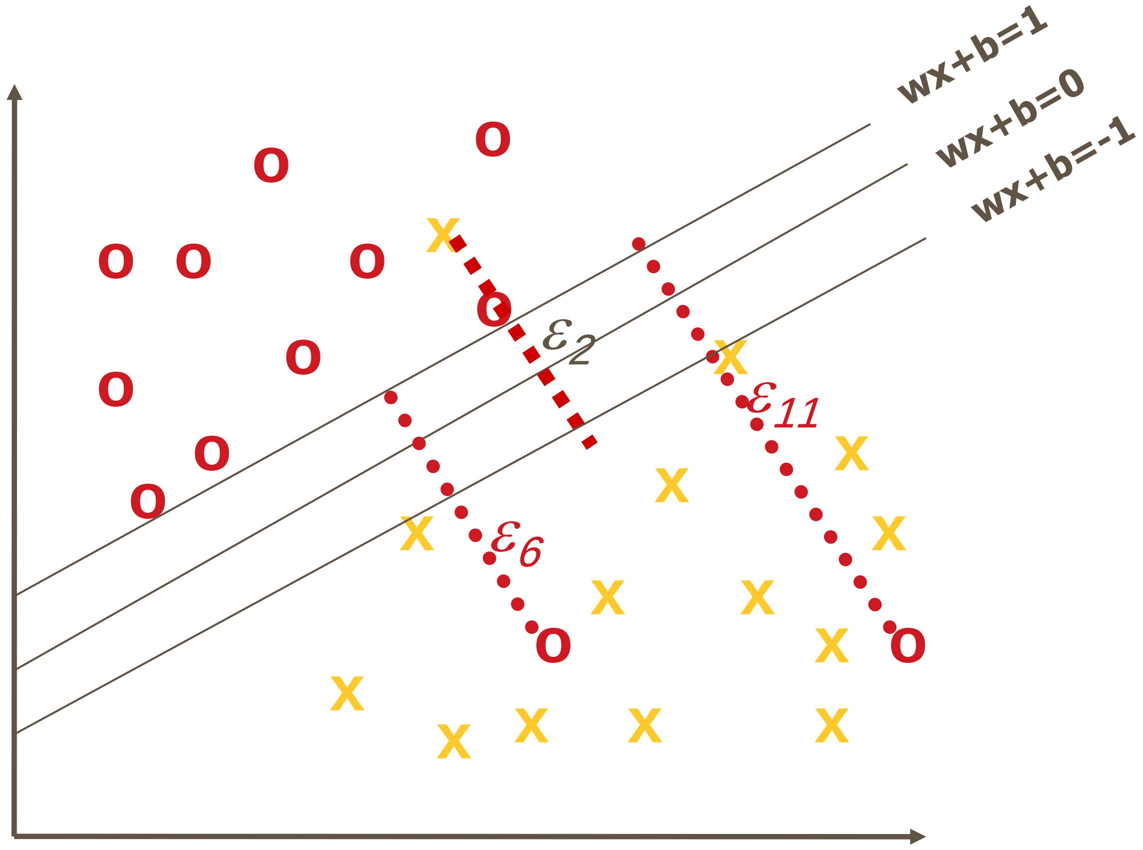
SUPPORT VECTOR MACHINES:
A HAND-WAVING EXPLANATION



Minimize:

$$\frac{1}{2}\,\boldsymbol{w}.\boldsymbol{w}$$

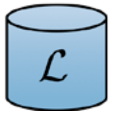
SUPPORT VECTOR MACHINES: THE NON-LINEARLY SEPARABLE CASE

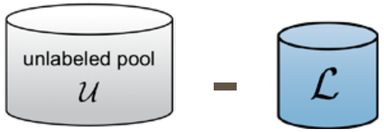


Minimize:

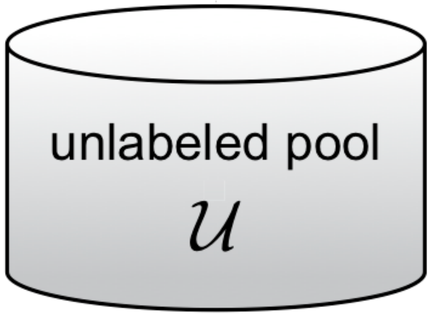
$$\frac{1}{2} w.w + C\sum_{k=1}^{R} \varepsilon_k$$

SUPERVISED LEARNING

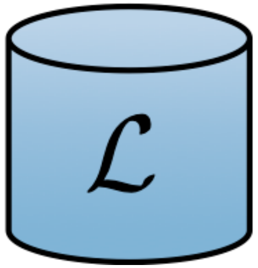
Induce (train) a classifier *C* over 

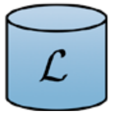
Apply *C* to unlabeled examples 

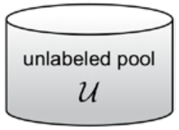
SUPERVISED LEARNING



What if we are clever in what examples we label?



Induce (train) a classifier *C* over 

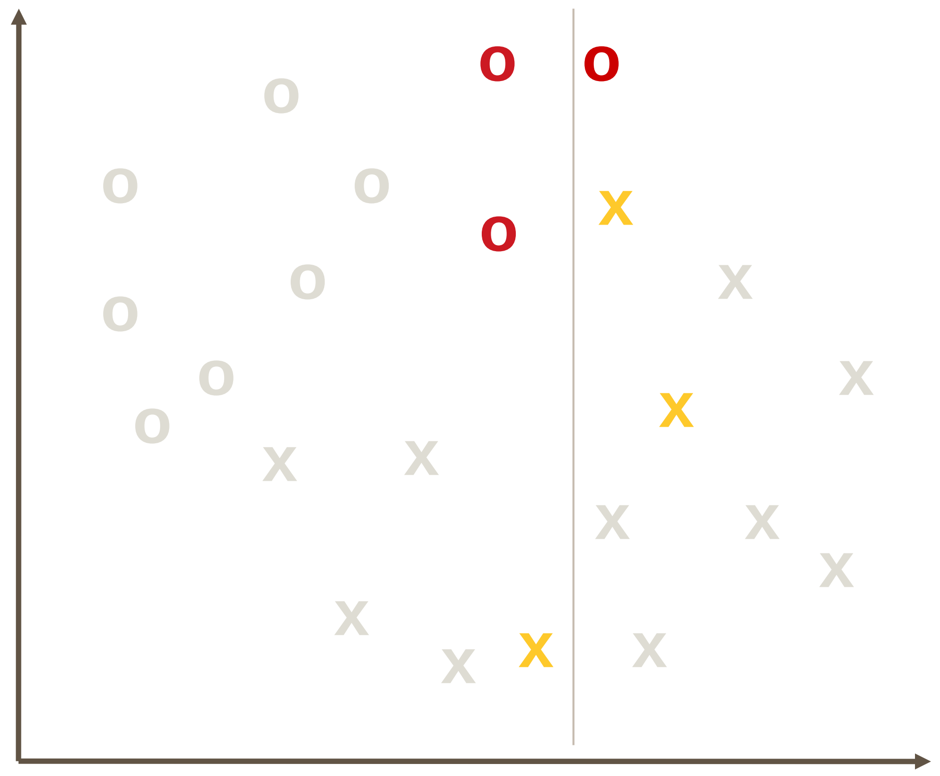
Apply *C* to unlabeled examples 

-


ACTIVE LEARNING

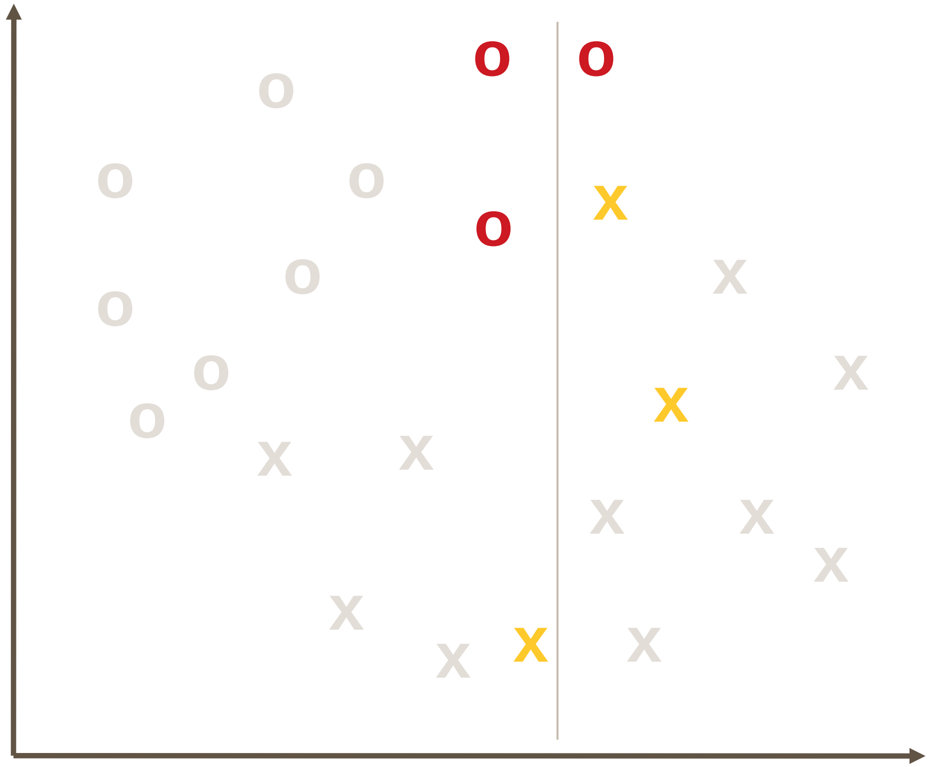
- **Key idea**: have the expert label examples most likely to be helpful in inducing a classifier
- Need fewer labels for good classification performance = less time/work/money
- Need a scoring function $f: x \rightarrow$ expected value of labeling x
- Most popular strategy: *uncertainty sampling*

UNCERTAINTY SAMPLING (W/ SVMS)



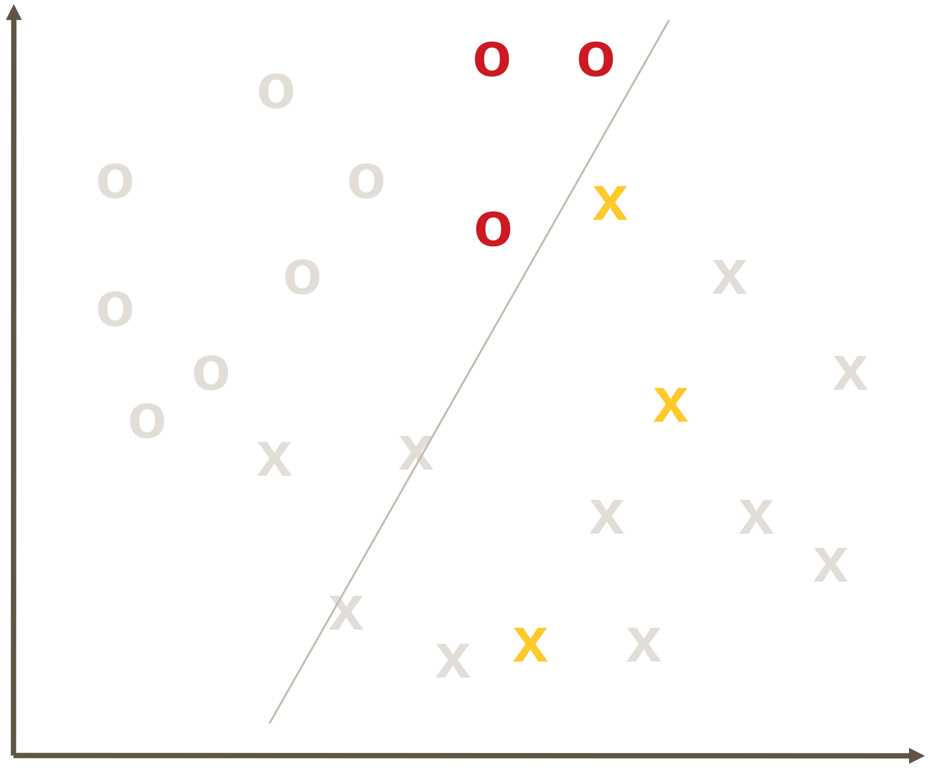
Which examples should we label next?

UNCERTAINTY SAMPLING (W/ SVMS)



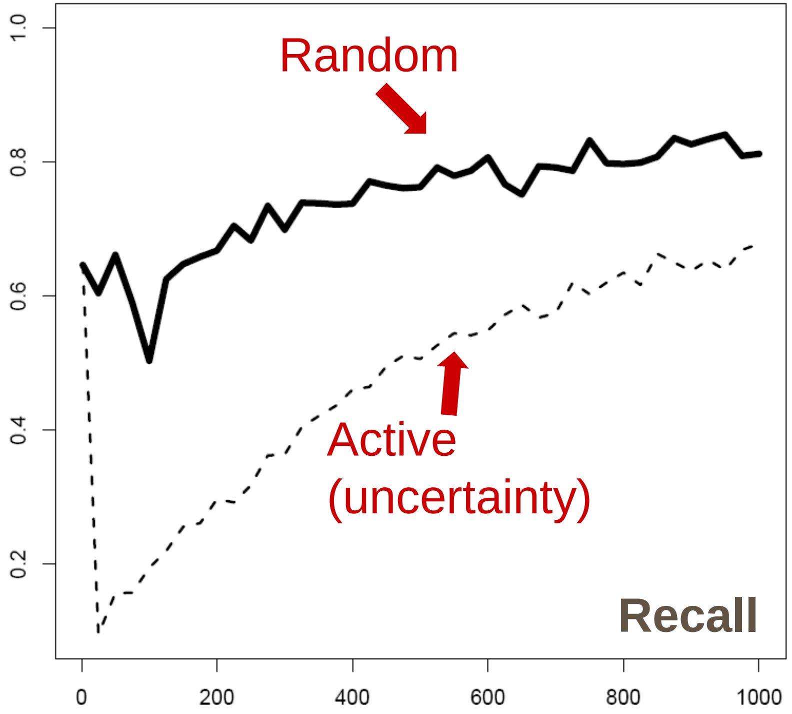
Uncertainty sampling: label the examples nearest the separating plane

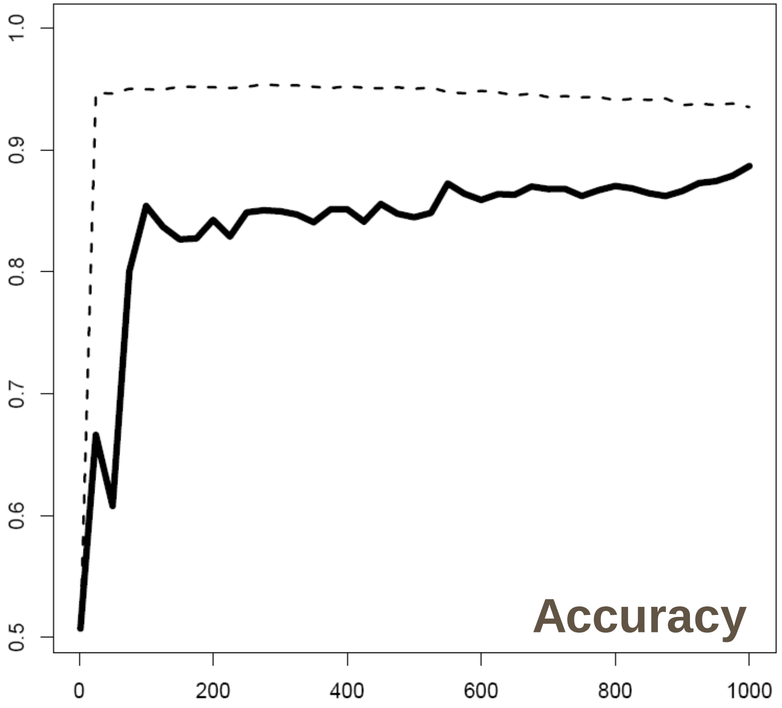
UNCERTAINTY SAMPLING (W/ SVMS)



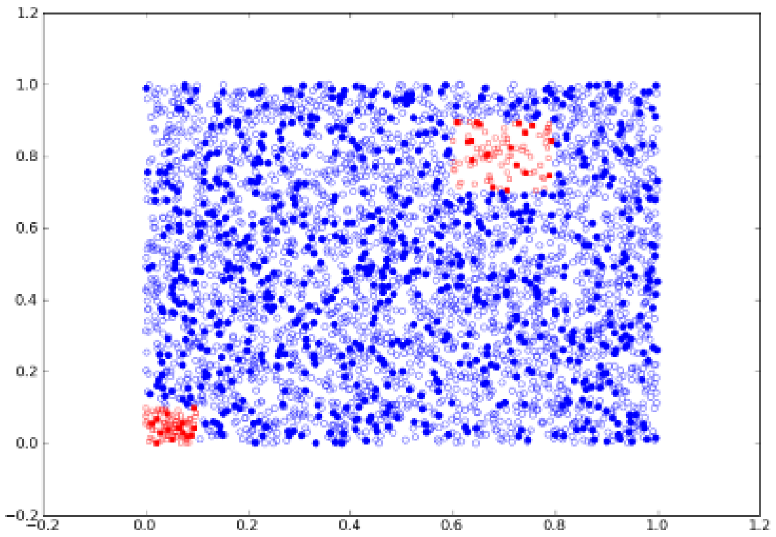
WHY 'OFF-THE-SHELF' AL DOESN'T WORK FOR CITATION SCREENING

Imbalanced data; 'relevant' class is very small (~5%), but sensitivity to this class is paramount

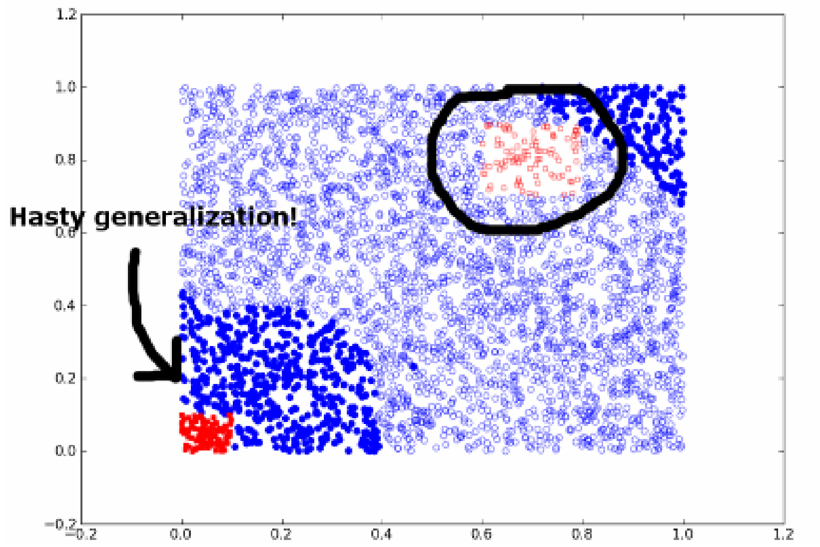




WHY MIGHT UNCERTAINTY SAMPLING FAIL?



Random sampling



Uncertainty sampling

Hasty generalization: uncertainty sampling may miss clusters

- Pre-clustering doesn't help
  - unreliable in high-dimensions
  - small clusters of interest

GUIDING AL WITH DOMAIN KNOWLEDGE

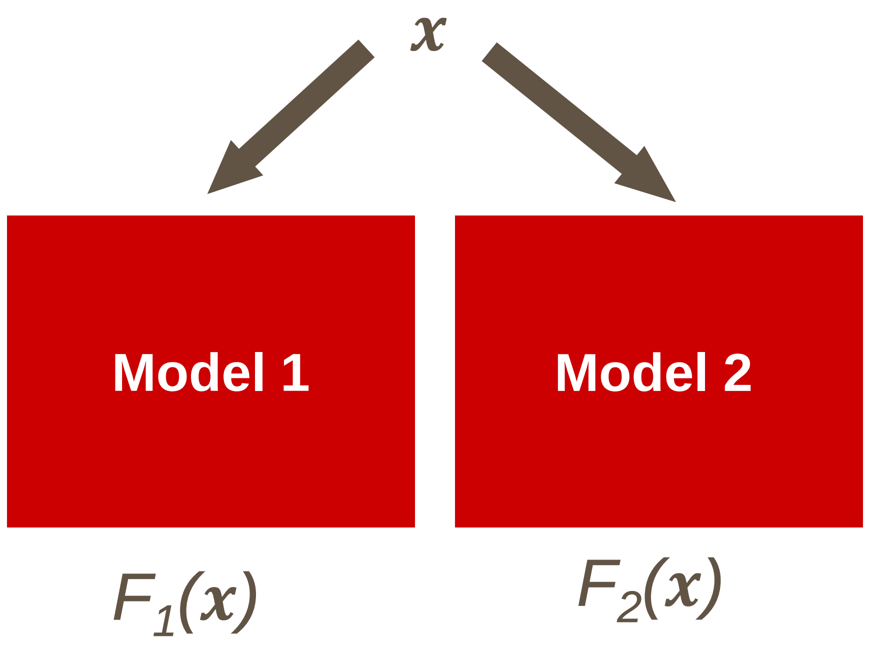
Labeled terms: terms or *n*-grams whose presence is indicative of class membership

⊕ *tension headache*, *leeches*, *aspirin*

⊖ *migraine headache*, *mice*

"Is aspirin better than leeches in inducing more than 50% relief in patients with tension headaches?"

CO-TESTING FRAMEWORK (MUSLEA ET AL., 2000)



If model 1 disagrees with model 2 about x, then x is a good point to label

LABELED TERMS + CO-TESTING

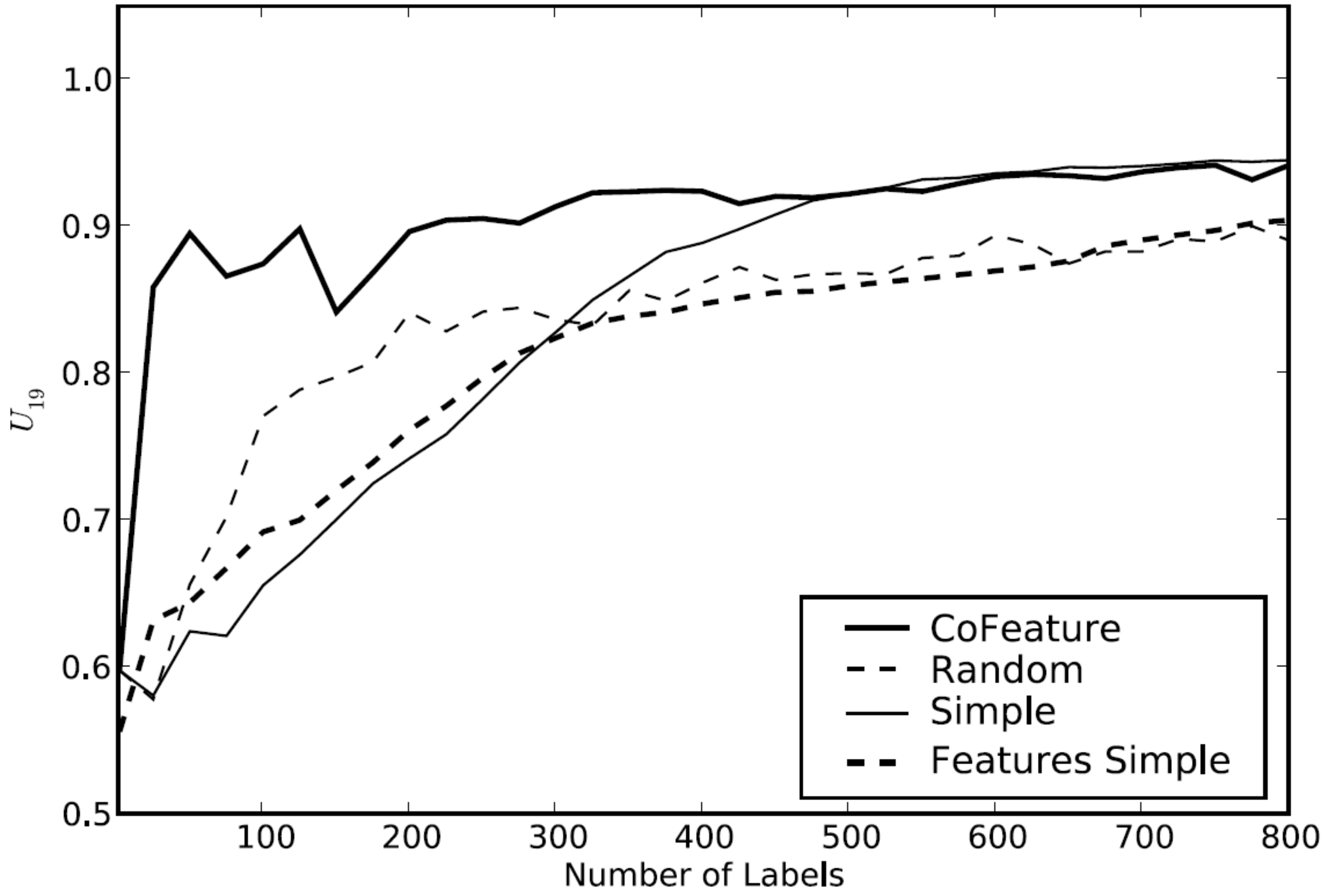
Model 1: Standard BOW (linear kernel) SVM

Model 2: Ratio of #pos terms to #neg terms

Query strategy:

- Find all documents about which the models disagree
- Select for labeling items of maximum disagreement

COPD:
GENETIC ASSOCIATIONS WITH COPD



MOST IMPORTANT REQUIREMENT FOR MACHINE LEARNING TO WORK: THE DATA

- Are the features predictive of the class?
- How noisy is the data? (attribute noise vs. class noise)
- Do you have enough (labeled) data?
- **Are the training samples representative?**

TRANSFER LEARNING

- A machine learning technique to improve performance leveraging on related knowledge
- A primary task on dataset T
- An auxiliary dataset T_{aux}
- T and T_{aux} are usually related and have similar distributions

Auxiliary T_{aux}

Primary T

TRANSFER LEARNING EXAMPLES

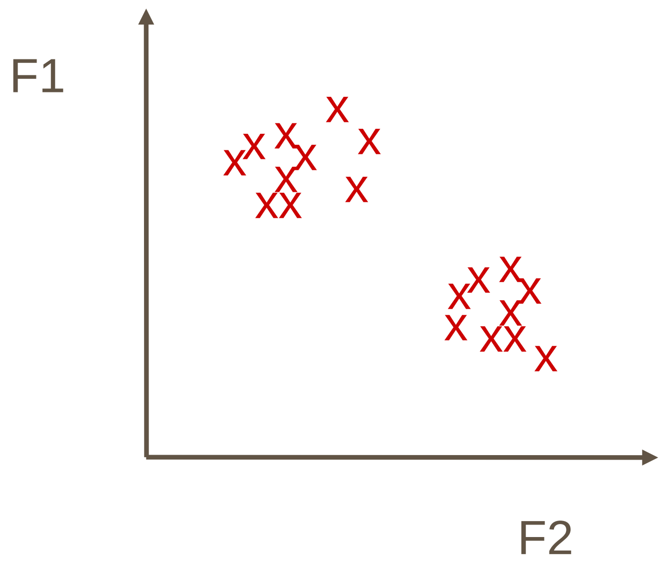
- Predicting readmission to hospitals
  - Use data from other hospitals to predict for your hospital
- Predicting MS progression
  - Combining data from multiple physicians

Fall 2017

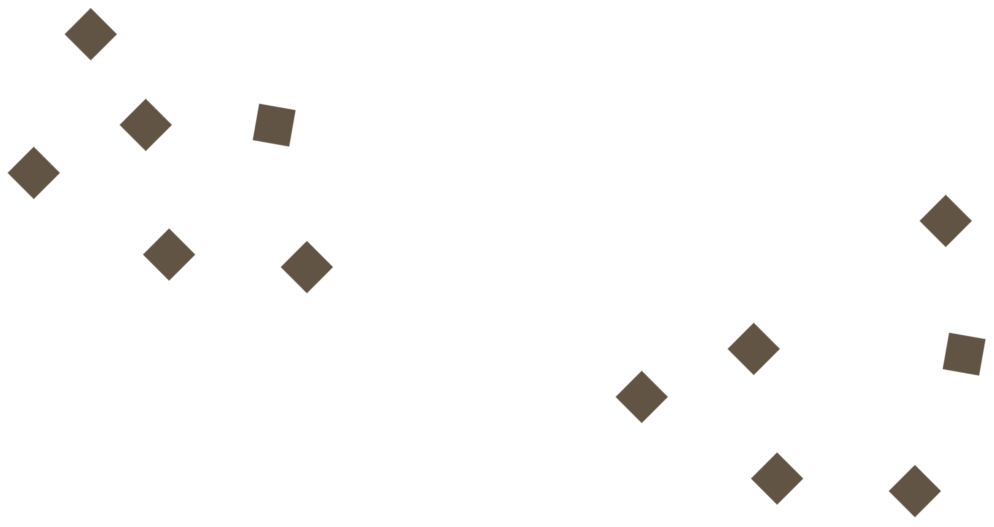
UNSUPERVISED LEARNING

CLUSTERING

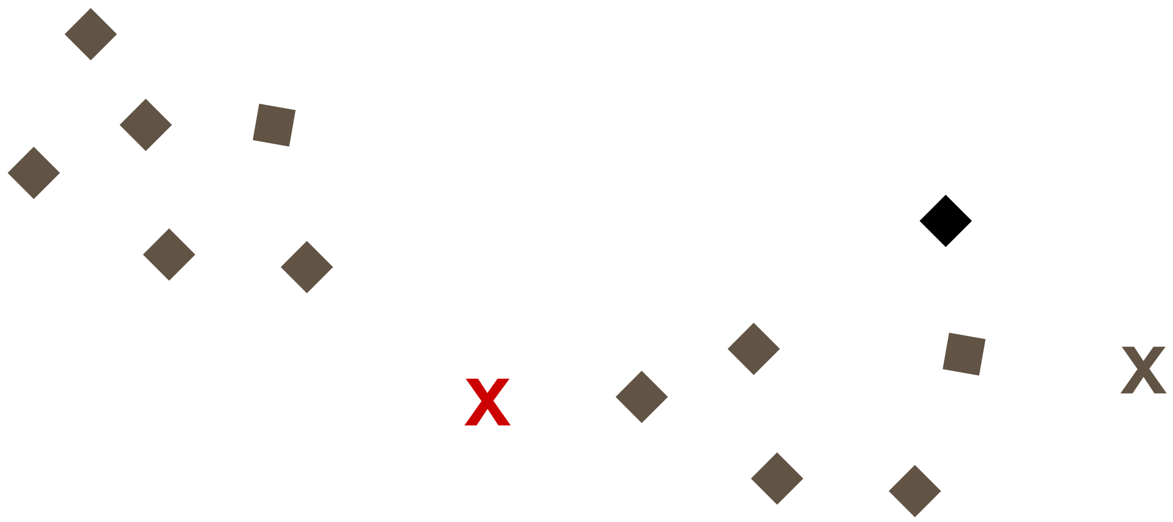
- Given a set of data points, each described by a set of attributes, find clusters such that:
  - Inter-cluster similarity is maximized
  - Intra-cluster similarity is minimized
- Requires the definition of a similarity measure



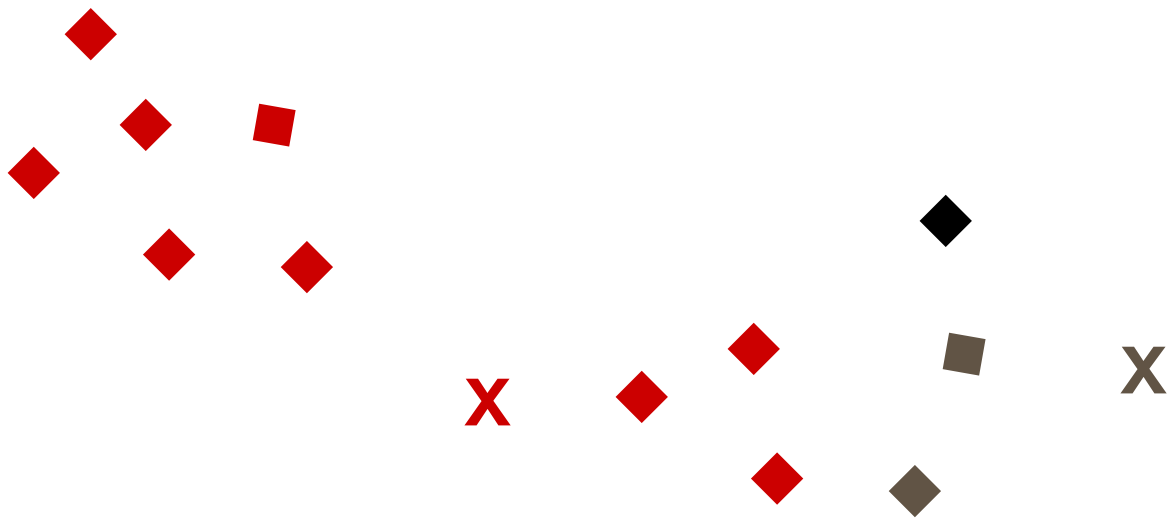
EXAMPLE: K-MEANS



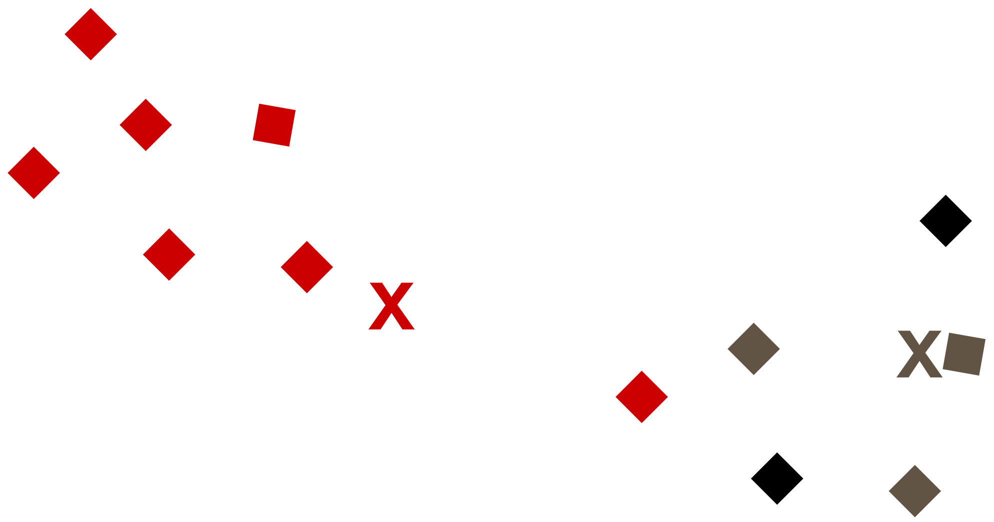
EXAMPLE: K-MEANS



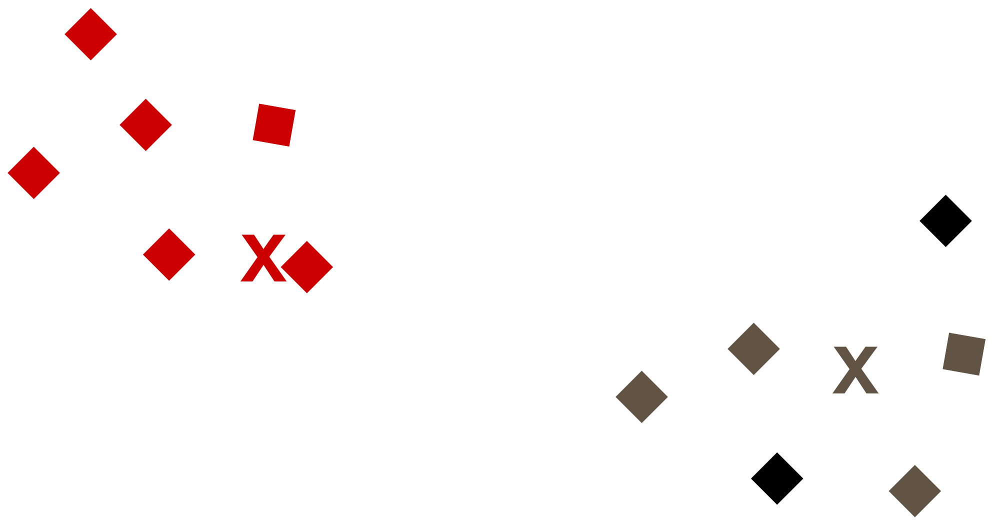
EXAMPLE: K-MEANS



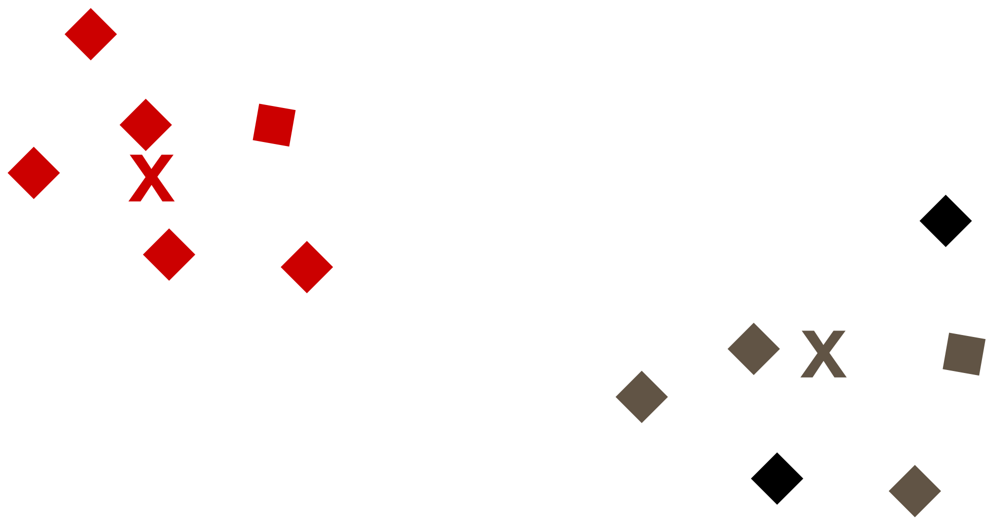
EXAMPLE: K-MEANS



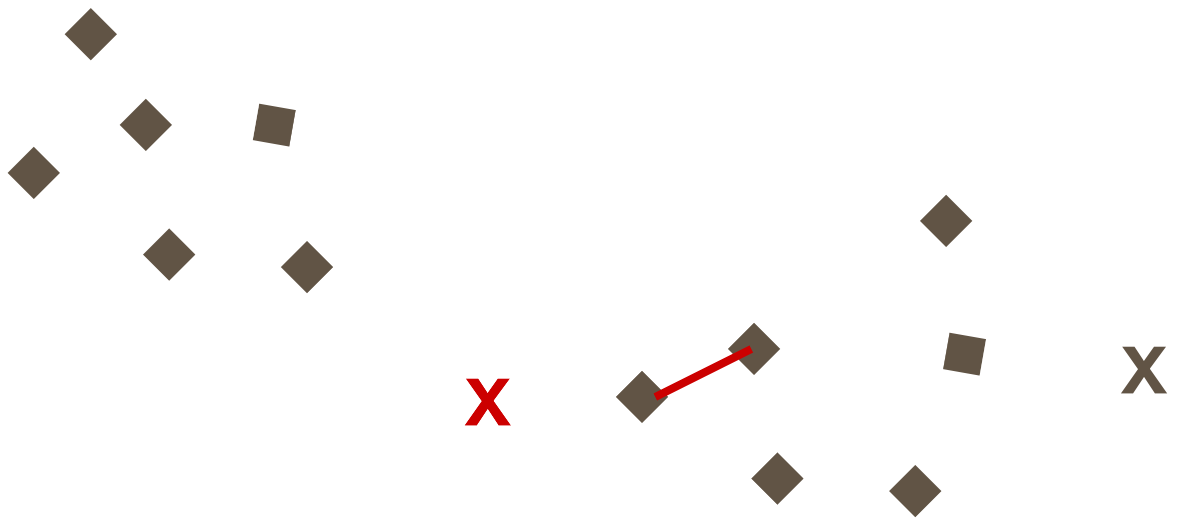
EXAMPLE: K-MEANS



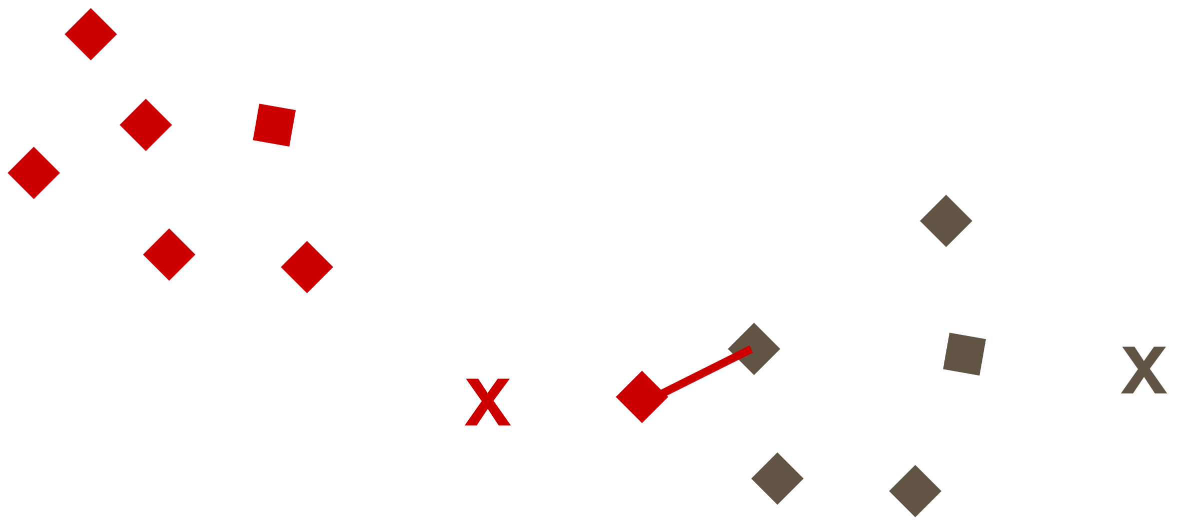
EXAMPLE: K-MEANS



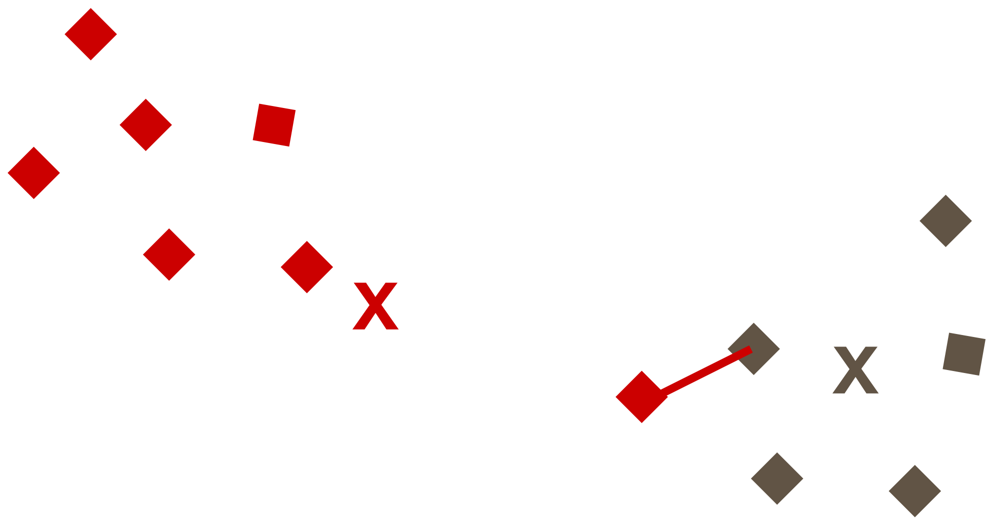
EXAMPLE: CONSTRAINED K-MEANS



EXAMPLE: CONSTRAINED K-MEANS



EXAMPLE: CONSTRAINED K-MEANS

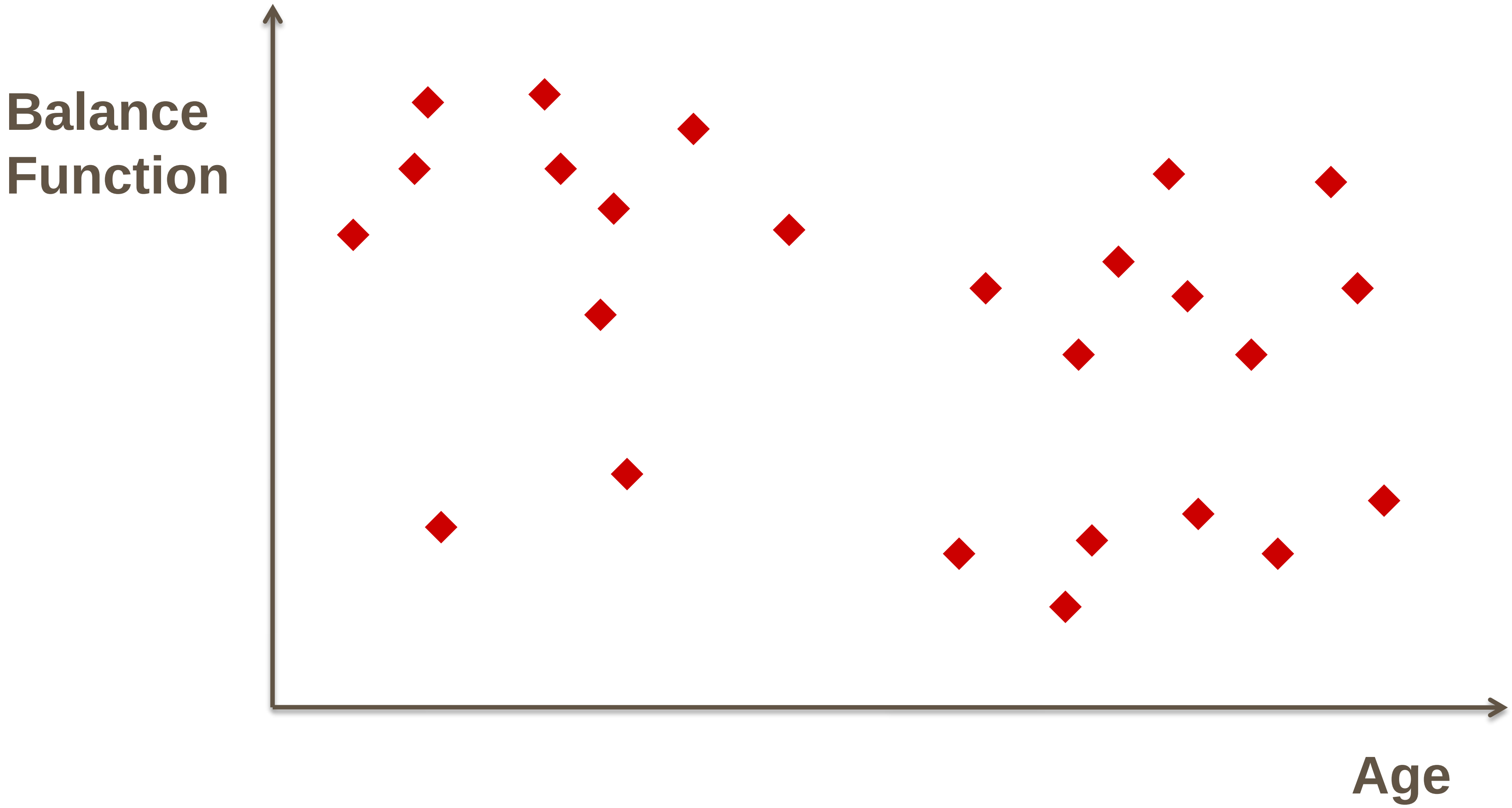


CHALLENGES IN CLUSTERING MEDICAL DATA

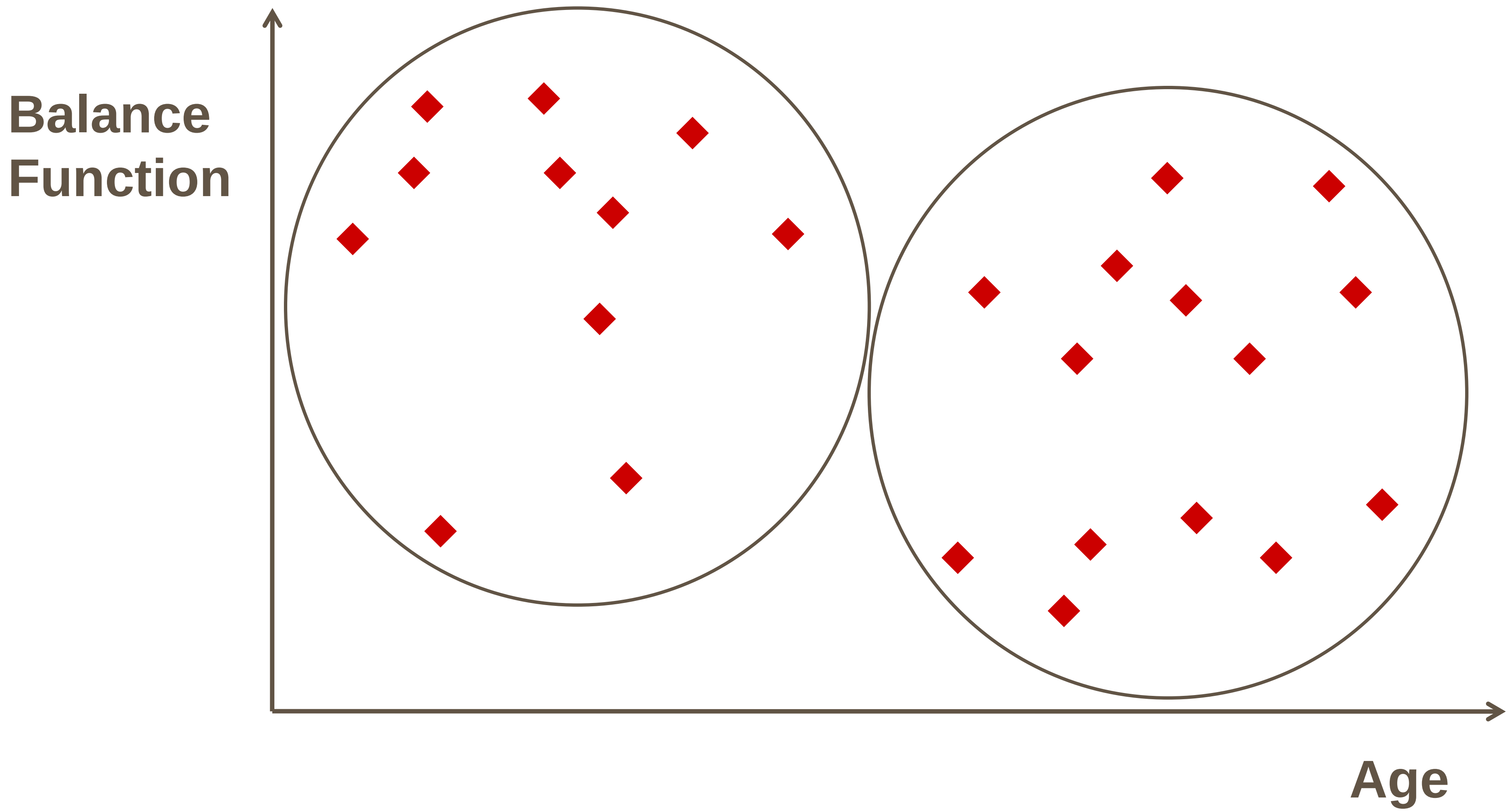
- Confounding factor:
  - One or a set of features whose effect will lead to undesirable clustering solution if not removed

- Clustering clinical data:
  - Physician subjectivity
  - Age for neurological test scoring in MS

EXAMPLE: VESTIBULAR DISORDERS



CLUSTERING WITH K = 2



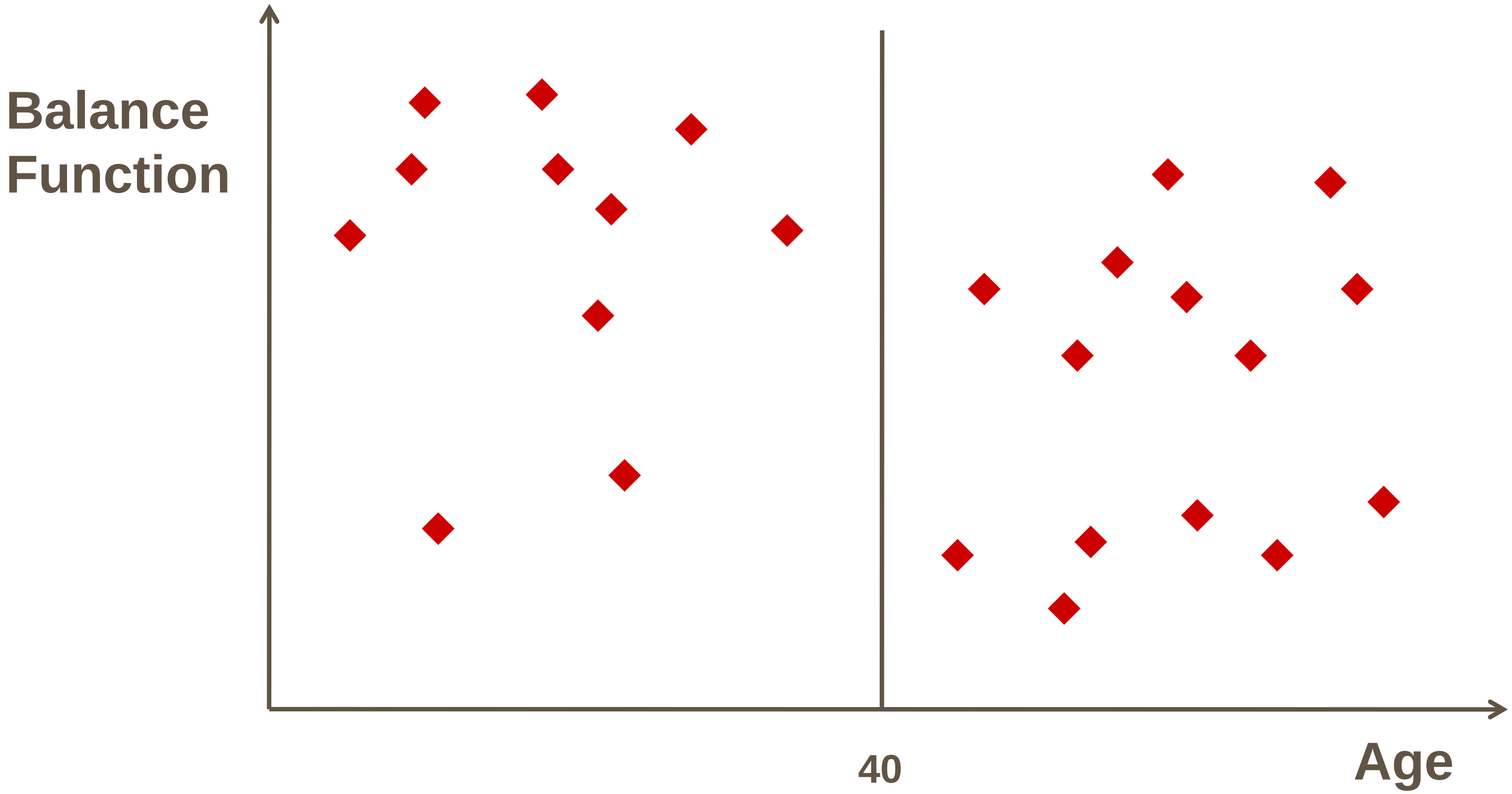
PROPOSED SOLUTION

- Remove the impact of confounding factor F via constraint-based clustering:
  1. Bin the data into homogeneous groups w.r.t. F
  2. Apply clustering to each group and generate pair-wise instance constraints
  3. Apply constraint-based clustering to entire data

STEP 1: BINNING (STRATIFICATION)

- **Categorical F:**
  - Create one bin per category
  - Example: one bin per physician for MS data
- **Numeric F:**
  - Create bins of:
    - *Uniform ranges or uniform bin sizes*
    - *Domain knowledge*
    - *More sophisticated binning methods, such as nonparametric density estimation, etc*

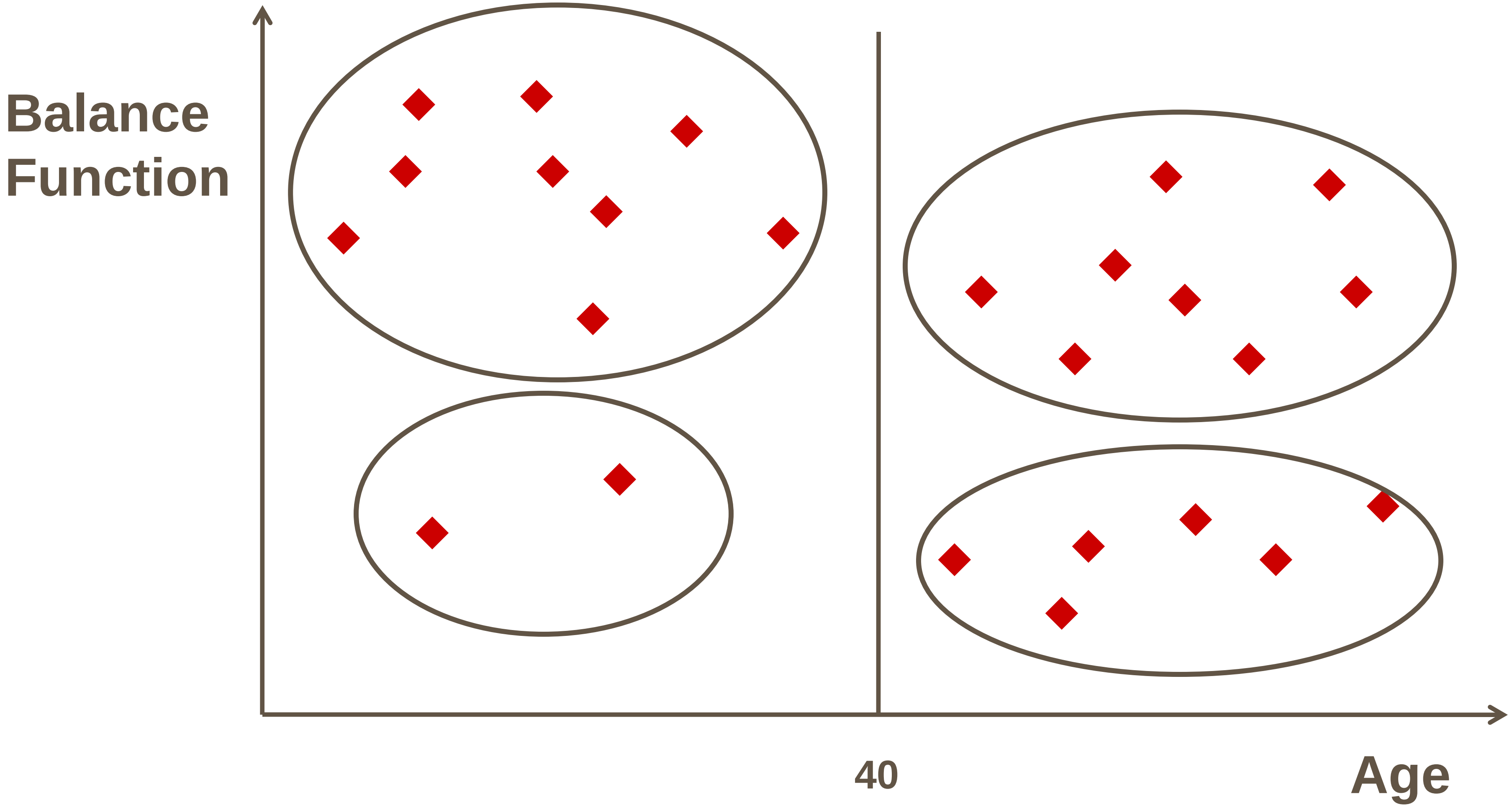
STEP 1: BINNING



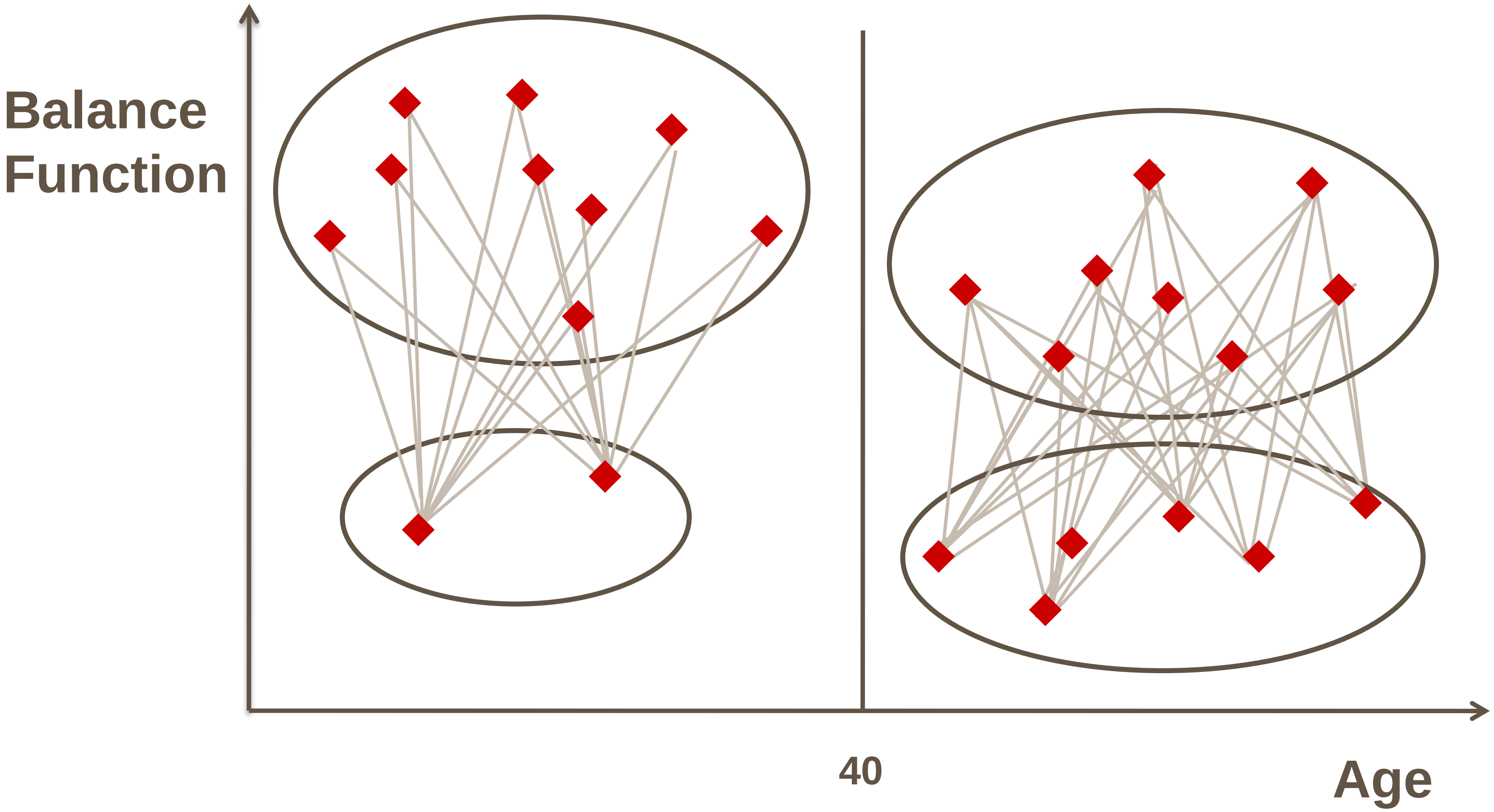
STEP 2: CLUSTER IN EACH BIN AND GENERATE CONSTRAINTS

- In each bin:
  - Apply clustering (e.g., EM over a mixture of Gaussians)
    - *Number of clusters can be specified by domain knowledge or inferred using criteria such as BIC*
  - Generate “must-not-link” constraints for pairs of instances in different clusters

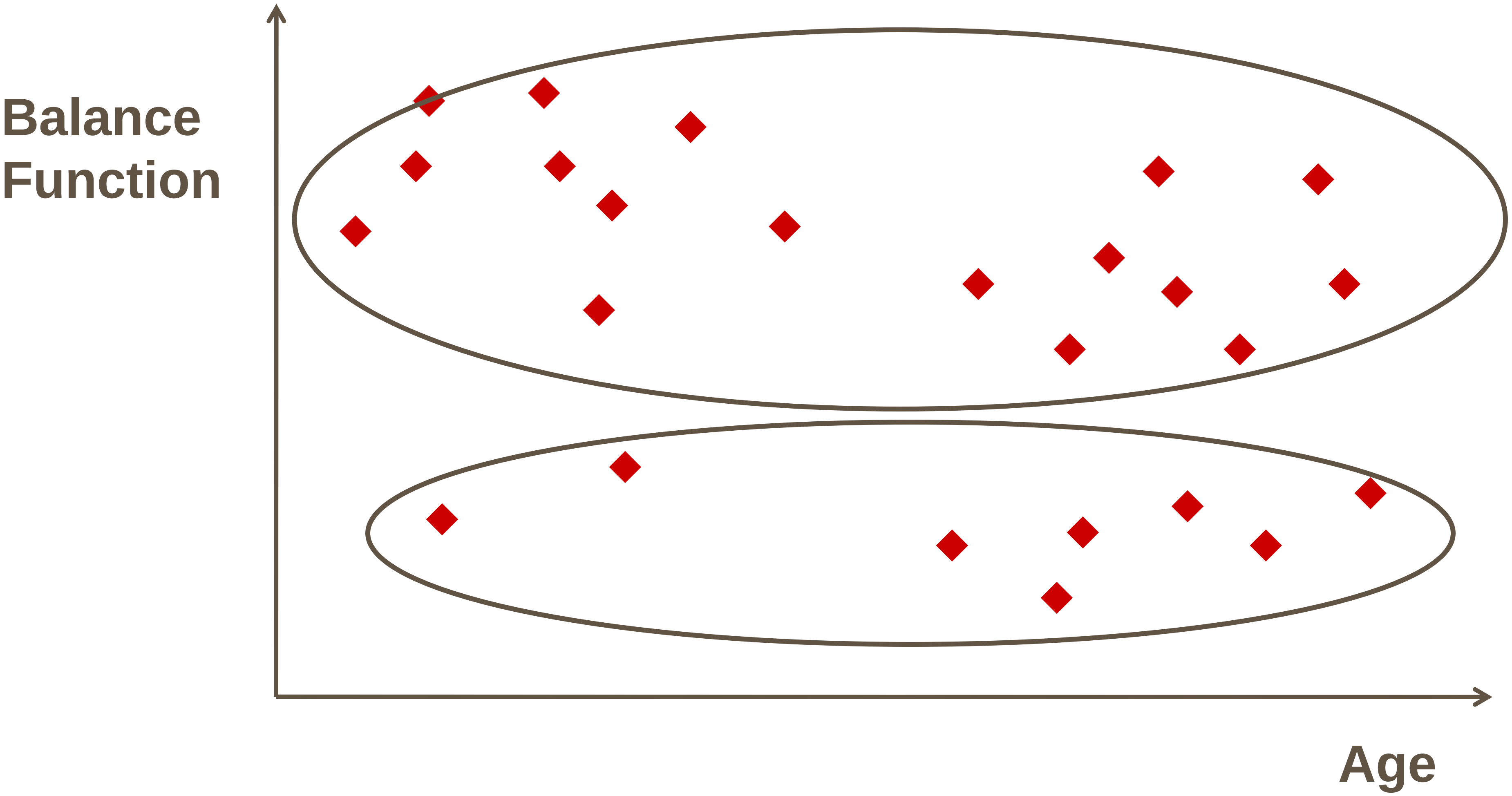
STEP 2: CLUSTER EACH BIN



STEP 2: GENERATE CONSTRAINTS



STEP 3: APPLY CONSTRAINT BASED CLUSTERING TO THE ENTIRE DATA

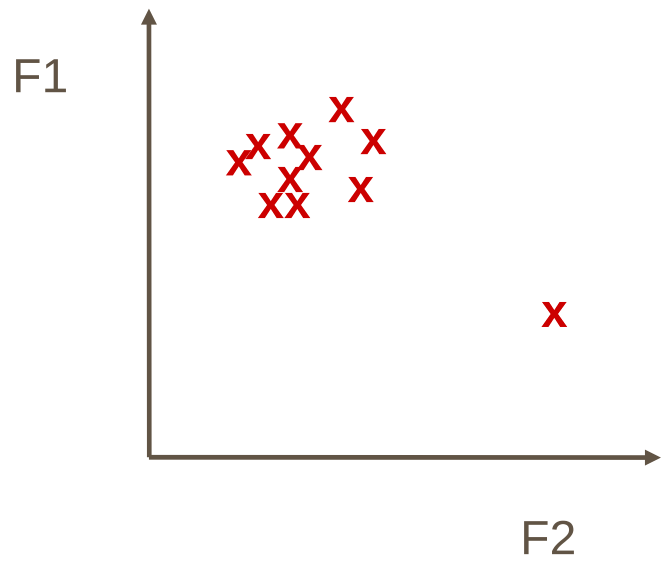


Fall 2017

ANOMALY DETECTION

ANOMALY DETECTION

- Given a set of data points, each described by a set of attributes, points that are far away from most of the other points – also called outliers
- Requires the definition of a similarity measure



TYPES OF ANOMALY DETECTION

- Supervised
  - Labelled normal and anomalous data
  - Similar to rare (minority) class mining
- Semi-supervised
  - Labels available only for normal data
- Unsupervised
  - No labelled data
  - Assumption: anomalies are rare compared to normal data

COMPLEXITIES OF ANOMALY DETECTION

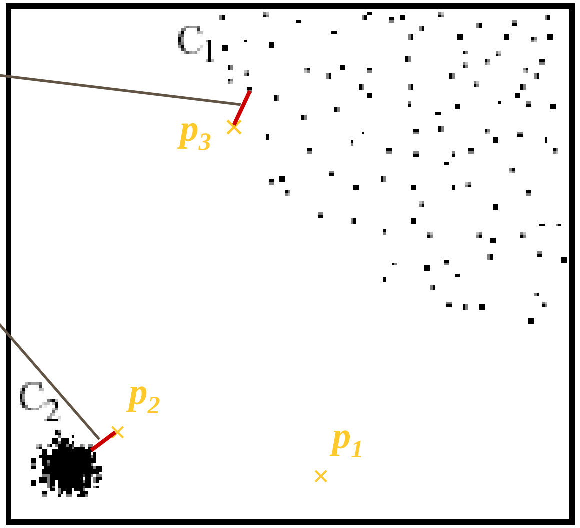
- Where does the “normal” data come from?
- Feature selection
- Metric
- Different parts of the space may have different densities

COMPLEXITIES OF ANOMALY DETECTION

- Which of P1, P2 and P3 are anomalies?

Distance from p_3 to nearest neighbor

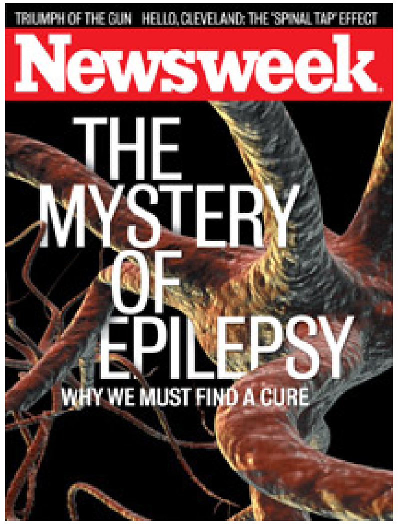
Distance from p_2 to nearest neighbor



ANOMALY DETECTION EXAMPLE: DETECTING CORTICAL LESIONS

- **50 million affected by epilepsy worldwide**
  - One-third remain refractory to treatment
  - One of the most common causes of TRE: Focal Cortical Dysplasia (FCD)
- **Treatment:**
  - Surgical resection of the abnormal cortical tissue (aka lesion)

Visual MRI Exam → Lesion Identification & Tracing → Inter-Cranial EEG Analysis → Resective Surgery



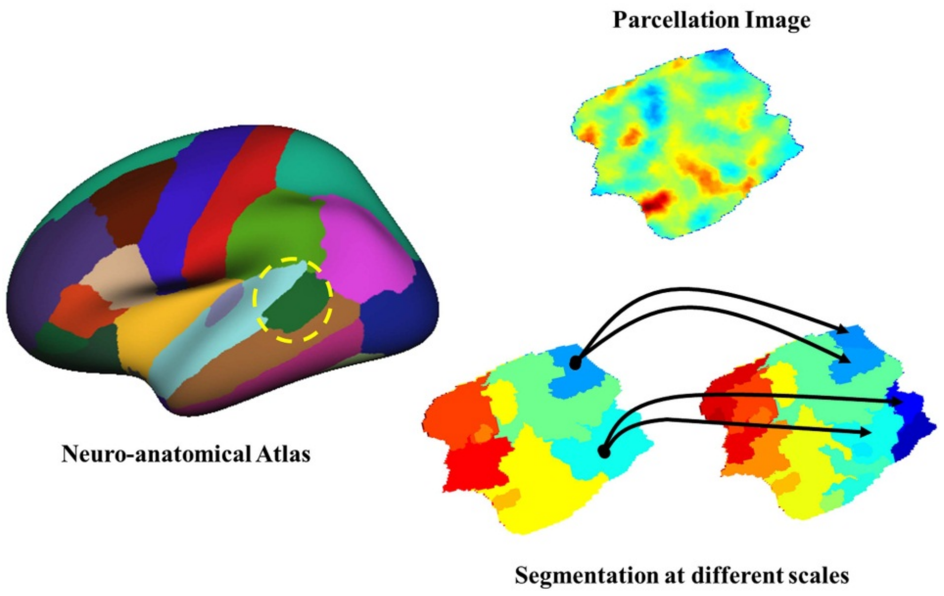
- **70-80%** of histologically verified FCD cases have **normal MRI**
- Chances of being seizure free after surgery:
  - **MRI-Positive: 66%** **MRI-Negative: 29%**

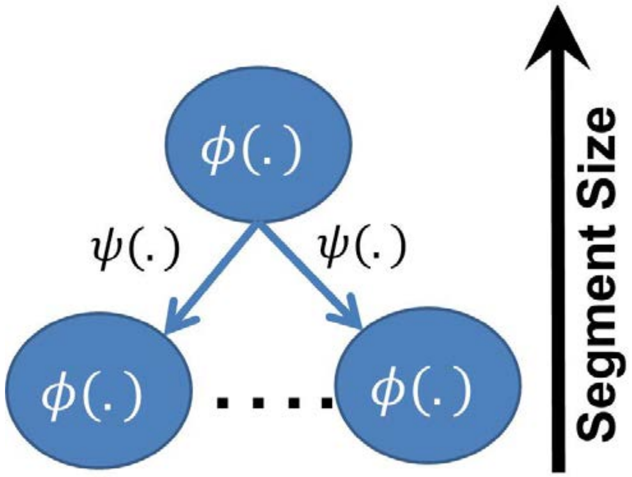
MACHINE LEARNING CHALLENGES

- Input data
  - Surfaces of FCD patients (MRI)
  - Resected tissue (MRI-Negatives): histopathologically verified
    - *Generous margins to ensure complete lesion removal*
    - *Exact location of the lesion is unknown*
- Labels
  - Resection zones for MRI-negatives
  - Lesion tracings by neuroradiologists for MRI-positives
    - *False positives in training data*
    - *False negatives in training data from long untreated epilepsy, trauma, etc.*

PROPOSED SOLUTION

- Hierarchical Conditional Random Fields for Outlier Detection
  - Discard pixel-level labels and use only image-level labels
  - Redefine FCD lesion as: *a cortical region which is an outlier when compared to the same region across a population of normal controls*

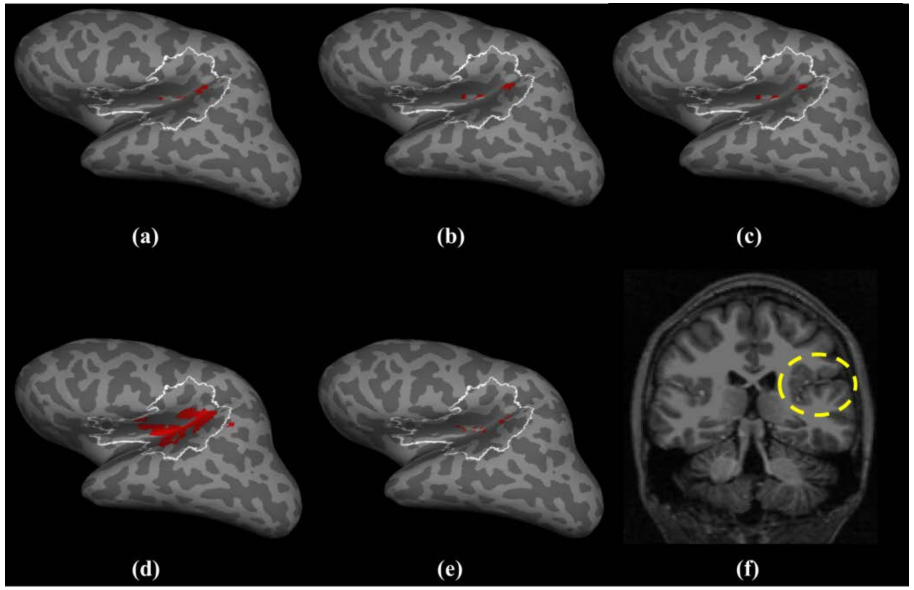




Hierarchical Conditional Random Field

RESULTS

- Tested on fifteen MRI-negative patients with successful surgery
- High detection rate **(80%)** for MRI-negative patients with higher average recall and precision



MY LAST WORDS…

There are many, many different learning algorithms, but the key to success is in having the right training data.

MLHC is a great conference….