AAS

Association for
Academic Surgery

Big Data and Informatics

David F. Schneider, MD, MS
Surgeon Informaticist
Assistant Professor
University of Wisconsin

Disclosures

- None

Overview

- Terminology
- Big Data
- Machine Learning
- Questions

TERMINOLOGY

Artificial Intelligence

A field of computer science that aims to make computers reason and act more like humans

Machine Learning

The primary methodology behind AI.

A collection of methods for inferring predictive models from sets of training instances.

Methods for training a computer to predict "unknowns" from a set of "knowns."

Natural Language Processing (NLP)

Sets of instructions or algorithms that allow computers to recognize and interpret human language (machine learning is one approach for accomplishing this task)

Big Data

Big Data refers to large and complex datasets prohibited from being processed with common or traditional database management tools and traditional data processing applications.

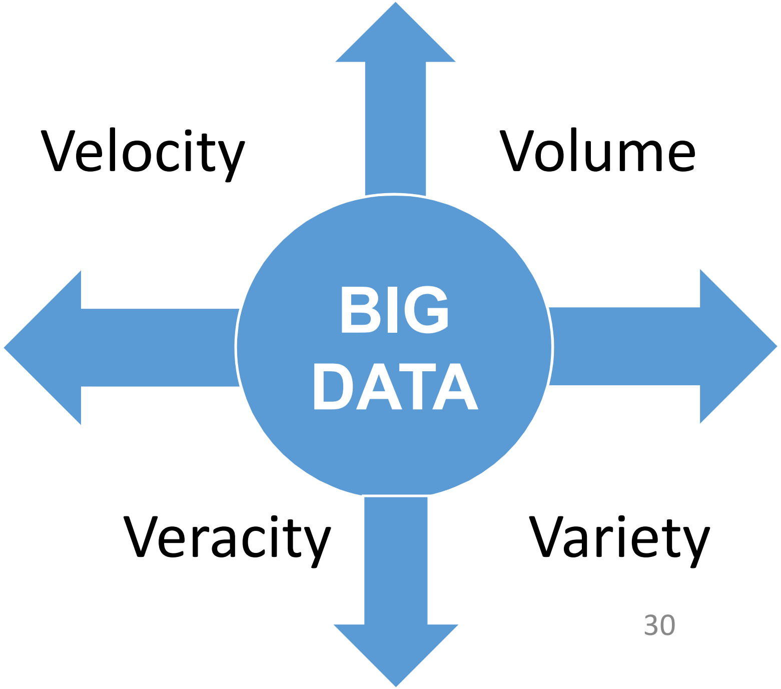
Big Data

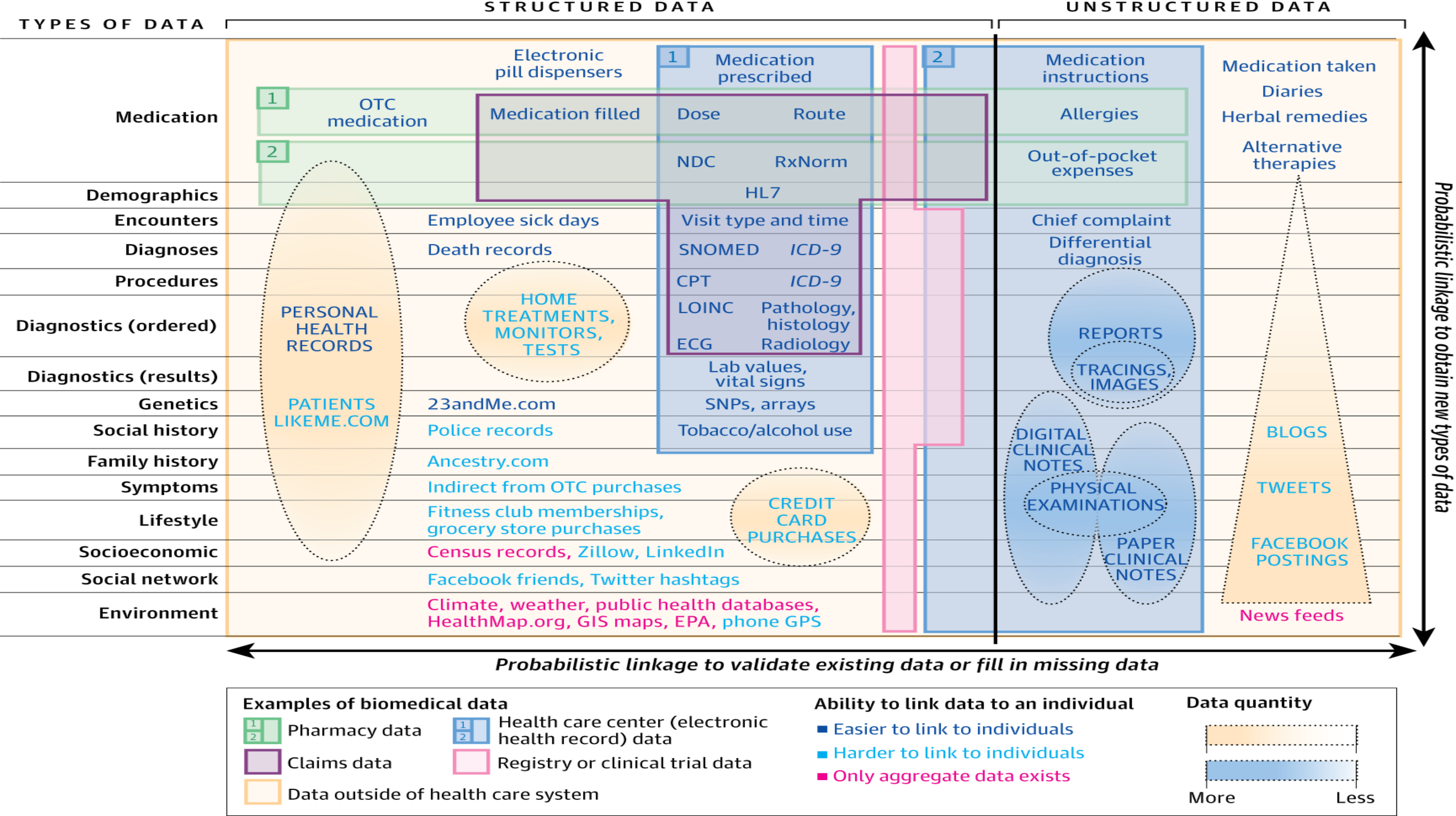
Big Data

	Traditional	Big Data
Database Management Software	Excel Access	Cloud Computing Distributed databases (Hbase)
Data Processing Tools	R STATA	Hadoop MongoDB
New Data Acquisition	slow	fast
Data Types	1-2	numerous

4 Dimensions of Big Data

1. VOLUME
2. VARIETY
3. VERACITY
4. VELOCITY





TYPES OF DATA

STRUCTURED DATA

UNSTRUCTURED DATA

Medication

Demographics

Encounters

Diagnoses

Procedures

Diagnostics (ordered)

Diagnostics (results)

Genetics

Social history

Family history

Symptoms

Lifestyle

Socioeconomic

Social network

Environment

Electronic pill dispensers

1 OTC medication

2

PERSONAL HEALTH RECORDS

PATIENTS LIKEME.COM

Employee sick days

Death records

HOME TREATMENTS, MONITORS, TESTS

23andMe.com

Police records

Ancestry.com

Indirect from OTC purchases

Fitness club memberships, grocery store purchases

CREDIT CARD PURCHASES

Census records, Zillow, LinkedIn

Facebook friends, Twitter hashtags

Climate, weather, public health databases, HealthMap.org, GIS maps, EPA, phone GPS

1 Medication prescribed

Medication filled

Dose

Route

NDC

RxNorm

HL7

Visit type and time

SNOMED

ICD-9

CPT

ICD-9

LOINC

Pathology, histology

ECG

Radiology

Lab values, vital signs

SNPs, arrays

Tobacco/alcohol use

2 Medication instructions

Allergies

Out-of-pocket expenses

Chief complaint

Differential diagnosis

REPORTS

TRACINGS, IMAGES

DIGITAL CLINICAL NOTES

PHYSICAL EXAMINATIONS

PAPER CLINICAL NOTES

Medication taken

Diaries

Herbal remedies

Alternative therapies

BLOGS

TWEETS

FACEBOOK POSTINGS

News feeds

Probabilistic linkage to obtain new types of data

Probabilistic linkage to validate existing data or fill in missing data

Examples of biomedical data

- 🟩 Pharmacy data
- 🟦 Health care center (electronic health record) data
- 🟪 Claims data
- ▢ Registry or clinical trial data
- 🟧 Data outside of health care system

Ability to link data to an individual

- ■ Easier to link to individuals
- ■ Harder to link to individuals
- ■ Only aggregate data exists

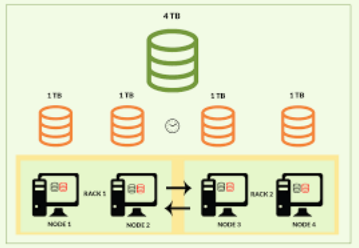
Data quantity

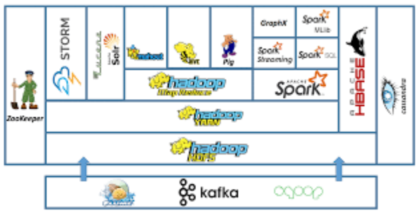
More

Less

HADOOP

- Open source project run by Apache
- Cheaply process large amounts of data, regardless of its structure
- The driving force behind big data industry





HADOOP

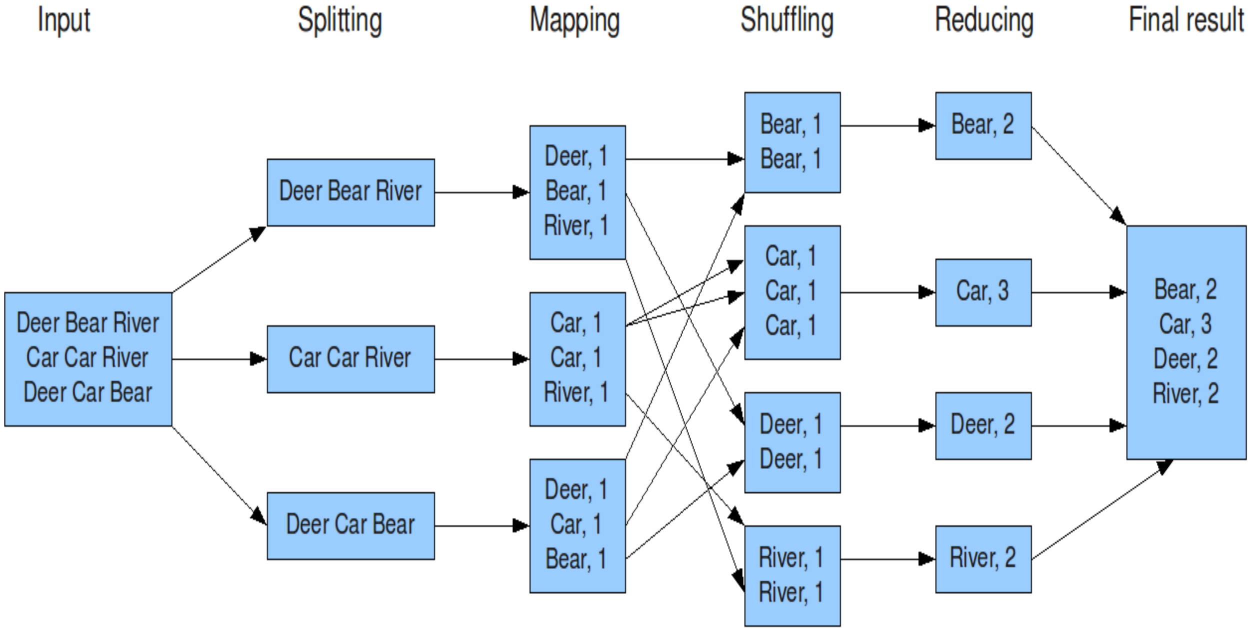
- Distributed processing of large datasets on clusters of computers
- Parallel computation, workflow
- Massive scalability and speed

- HDFS (**H**adoop **D**istributed **F**ile **S**ystem) - runs in a clustered environment
- MapReduce – programming paradigm for running processes over clustered environments

Pseudocode for MapReduce

AAS
Association for Academic Surgery

- Iterate over a large number of records
- Extract something of interest from each record
- Shuffle and sort intermediate results
- Aggregate intermediate results
- Generate a final output



Machine Learning

What is Machine Learning?

A collection of methods for inferring predictive models from sets of training instances

OR

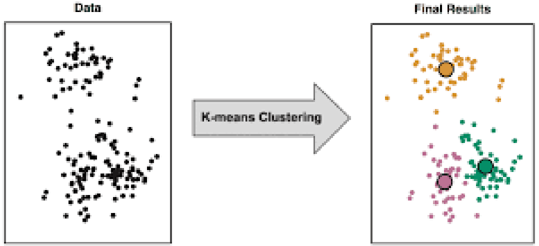
The methods behind artificial intelligence

OR

The science of getting computers to act without specific programming

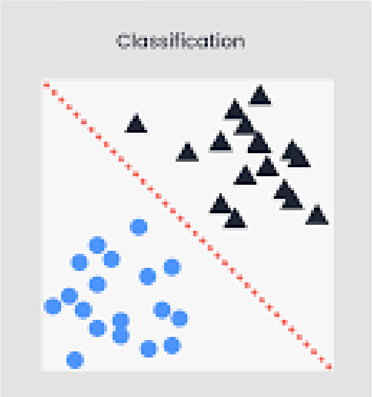
Machine Learning: Unsupervised

Methods to discover patterns from a dataset that has not been classified, labeled, or categorized. Commonly, unsupervised machine learning methods cluster the cases in a dataset by their similarity or differences of their features.

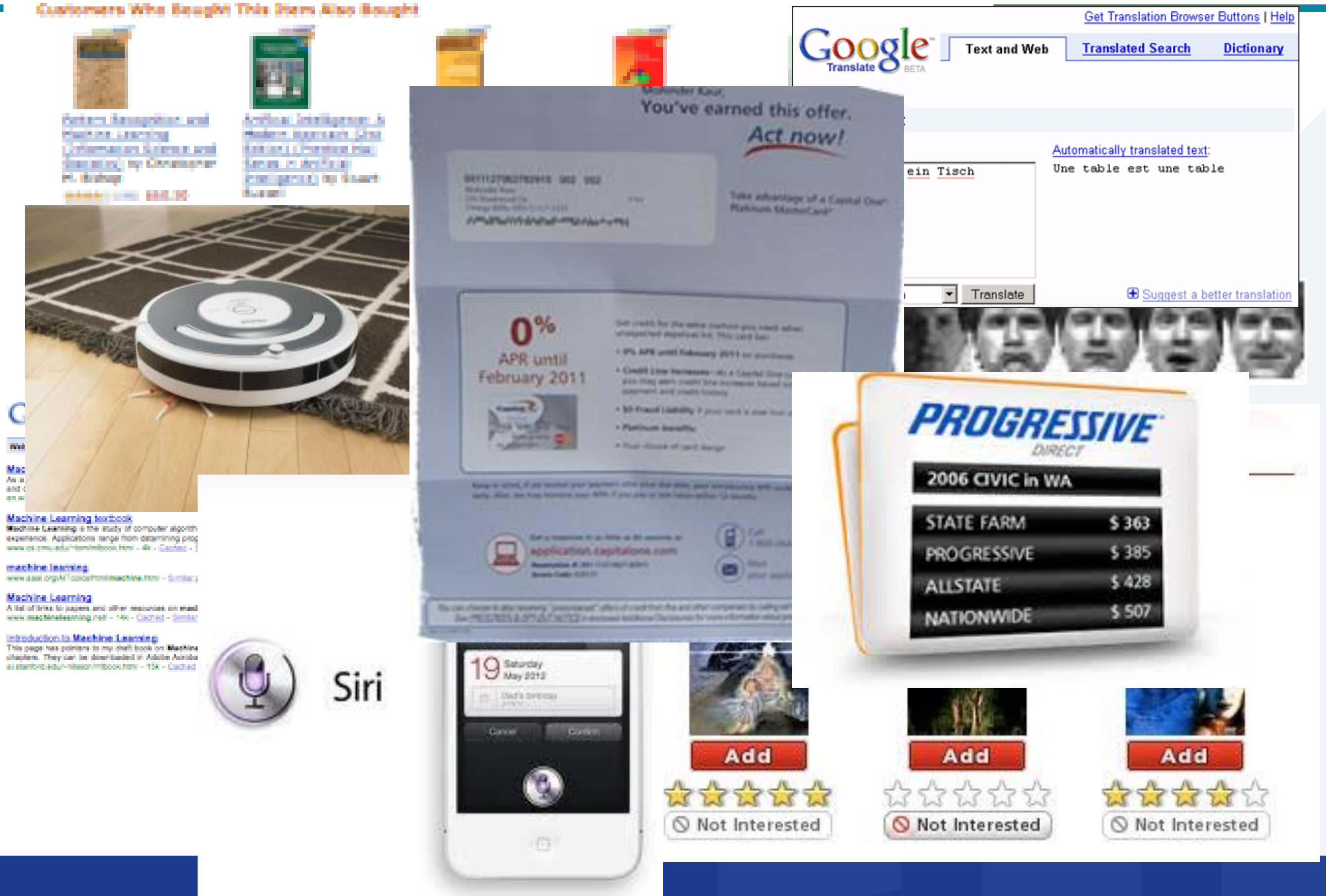


Machine Learning: Supervised

Methods to make predictions using a dataset that is labeled or classified by the outcome of interest with the purpose of then applying the algorithm to a test (unlabeled) set.

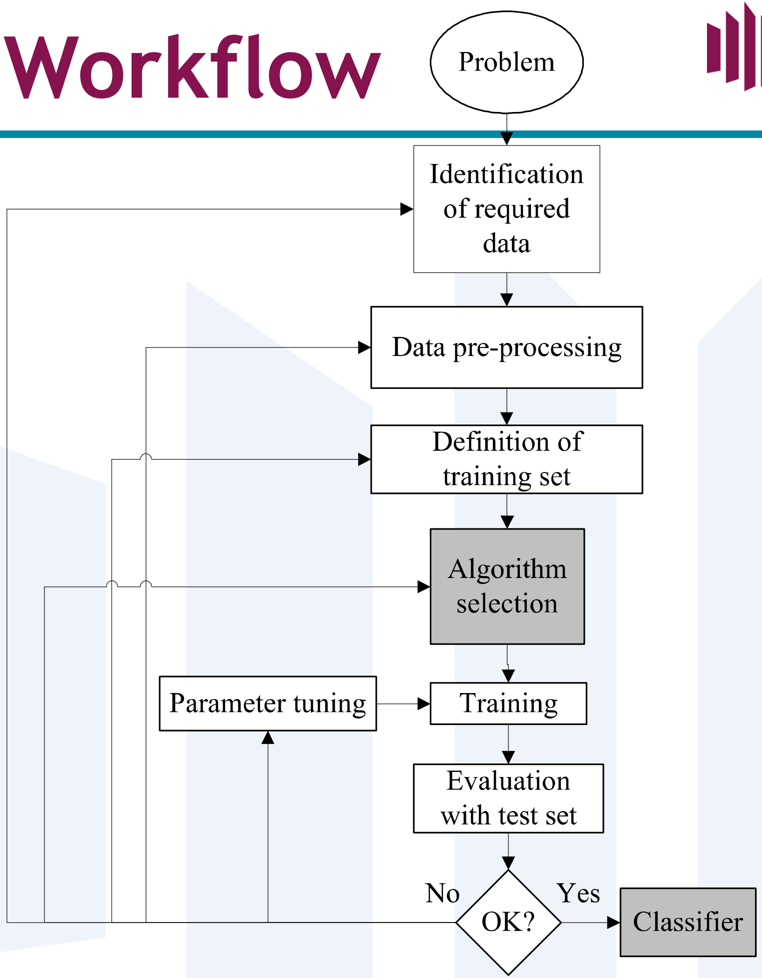


Machine Learning is All Around Us



Generalized Workflow

- Iterative
- Open
  - Consider presentation
  - audience
- Data sets
  - Requires planning
  - Consider n



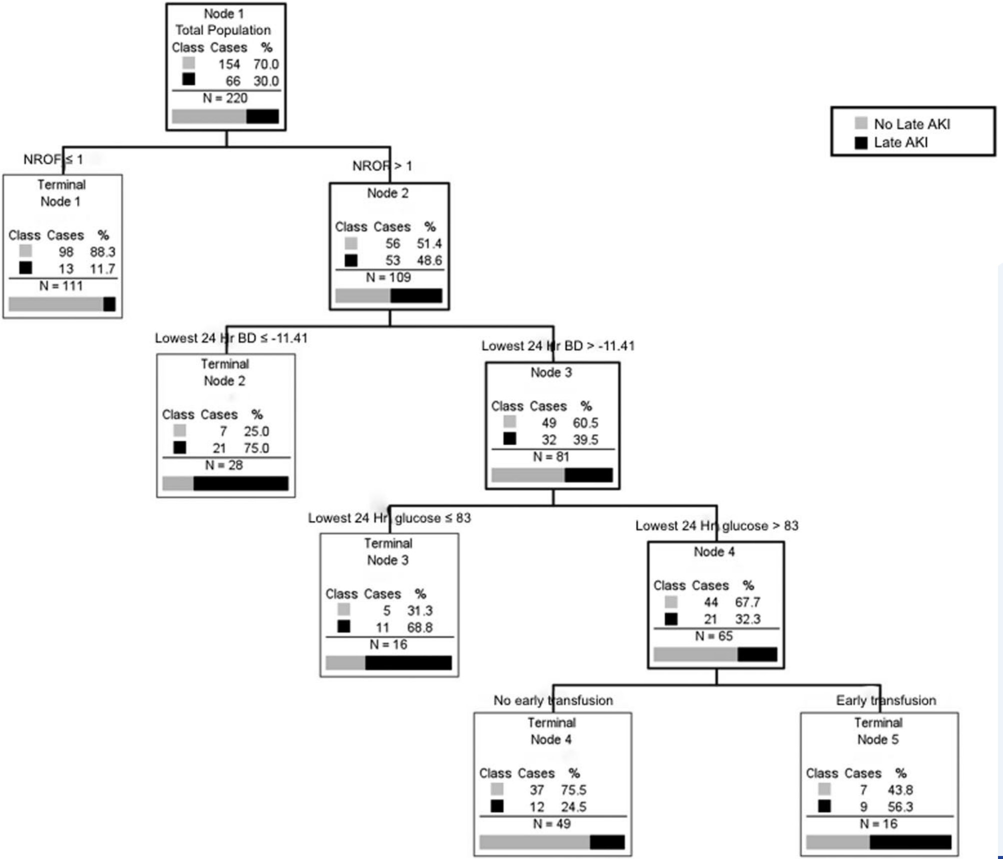
Types of Supervised ML

Continuous	Categorical
Regression linear polynomial	Logistic Regression
Decision Trees	KNN
Random Forests	Trees
Neural Networks	Bayesian
	SVM
	Rule-based

CART

- **C**lassification **A**nd **R**egression **T**ree
- predicts late acute kidney injury in burn patients.
- Uses segmentation by outcome label
- Non-parametric – rules
- Stopping, pruning

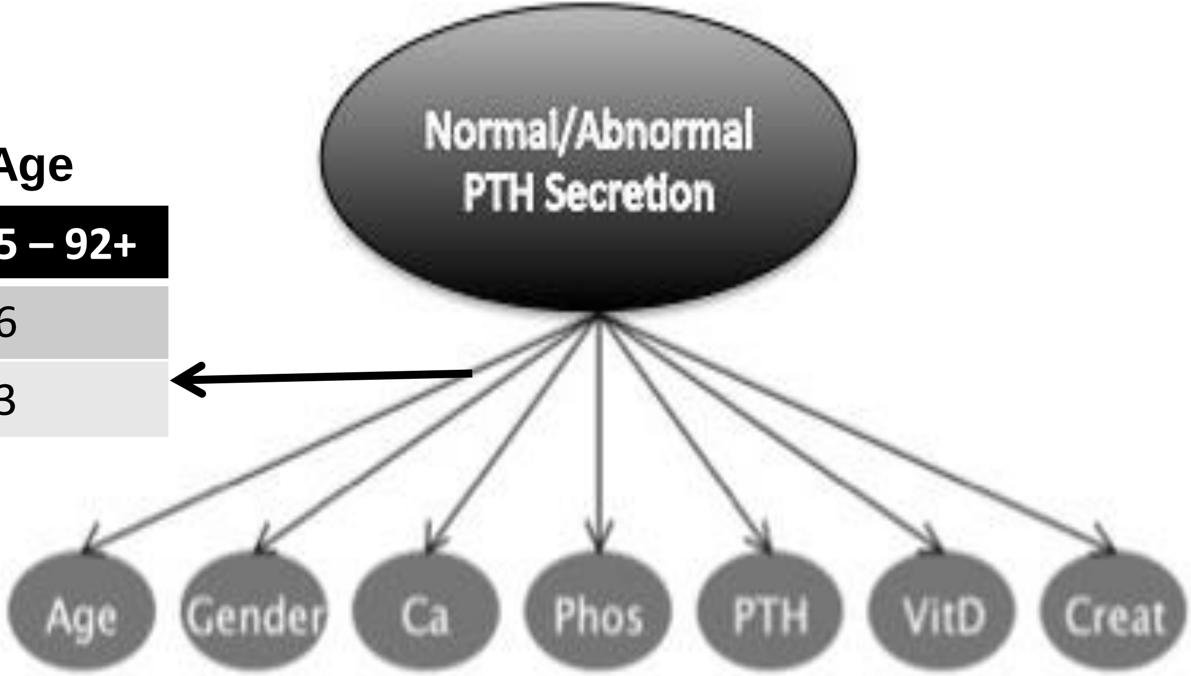
Schneider DF, et al. Predicting acute kidney injury in burn patients: A CART analysis. *J Burn Care Res*, 2012; 33:242-251



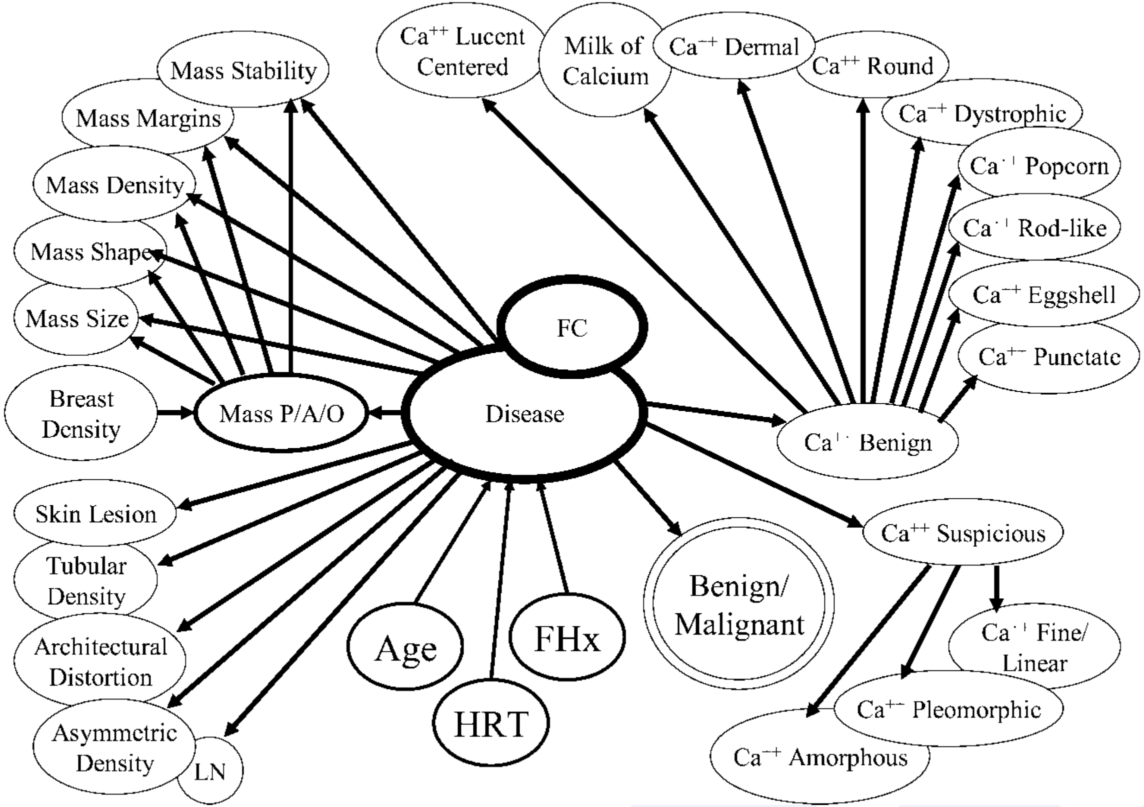
Bayesian Networks

Probability Table for Age

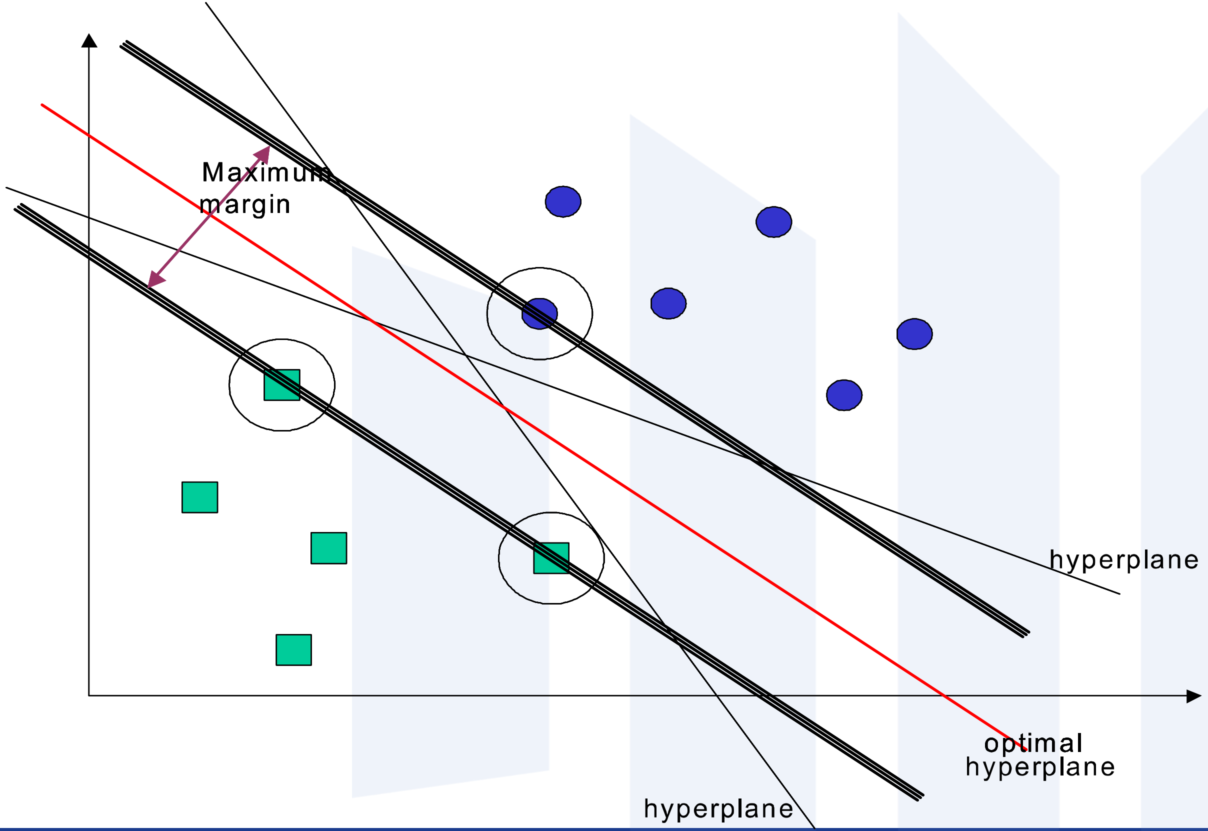
Group	0 – 53.5	53.5 – 92+
Normal	0.54	0.46
Abnormal	0.27	0.73



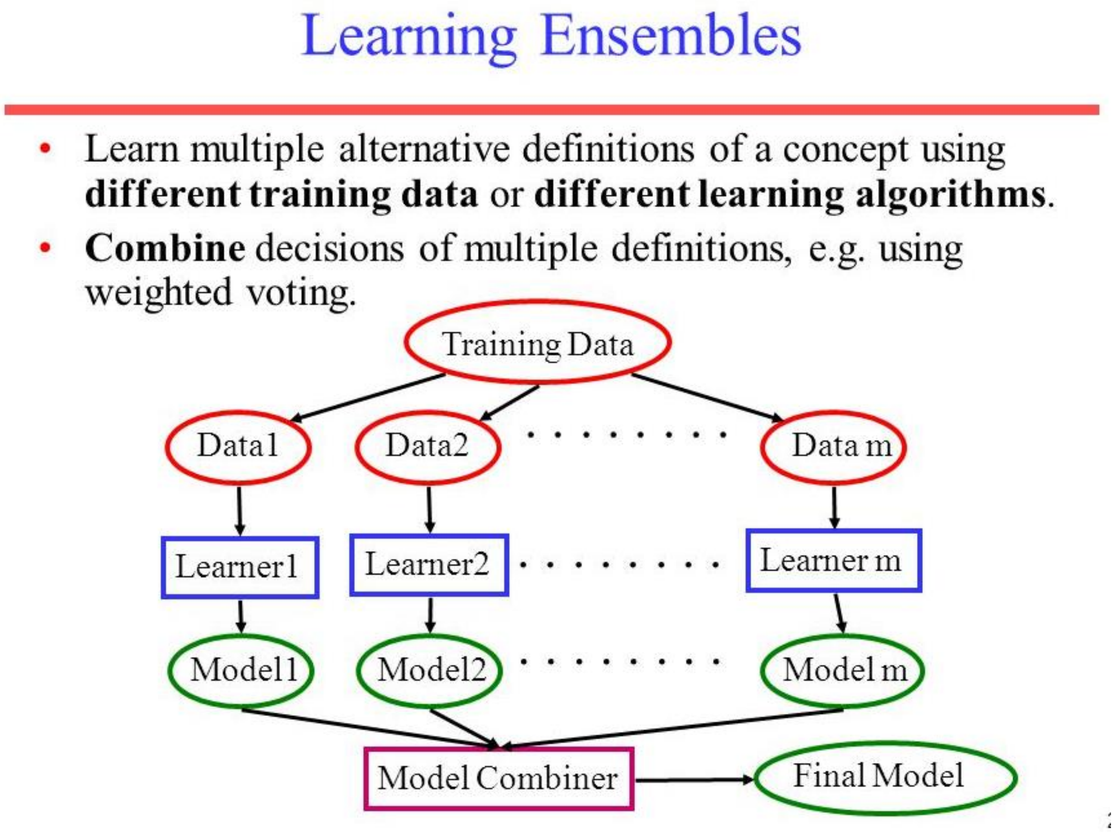
Bayesian Networks



Support Vector Machine



Ensembles



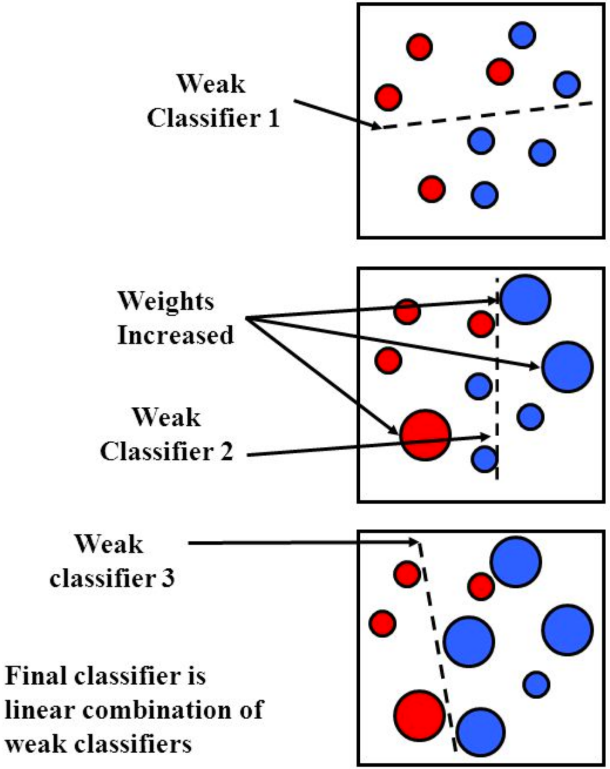
Learning Ensembles

- Learn multiple alternative definitions of a concept using **different training data** or **different learning algorithms**.
- **Combine** decisions of multiple definitions, e.g. using weighted voting.

Adaptive Boosting (AdaBoost)

Adaptive Boosting

- Output of multiple, “weaker” classifiers are combined into a weighted sum in the final, “boosted” classifier
- ***Adaptive* -** ”weak” learners are adjusted to account for misclassified instances by previous learners
- ***Adaptive Boosting with BayesNet as the “weak learners”***

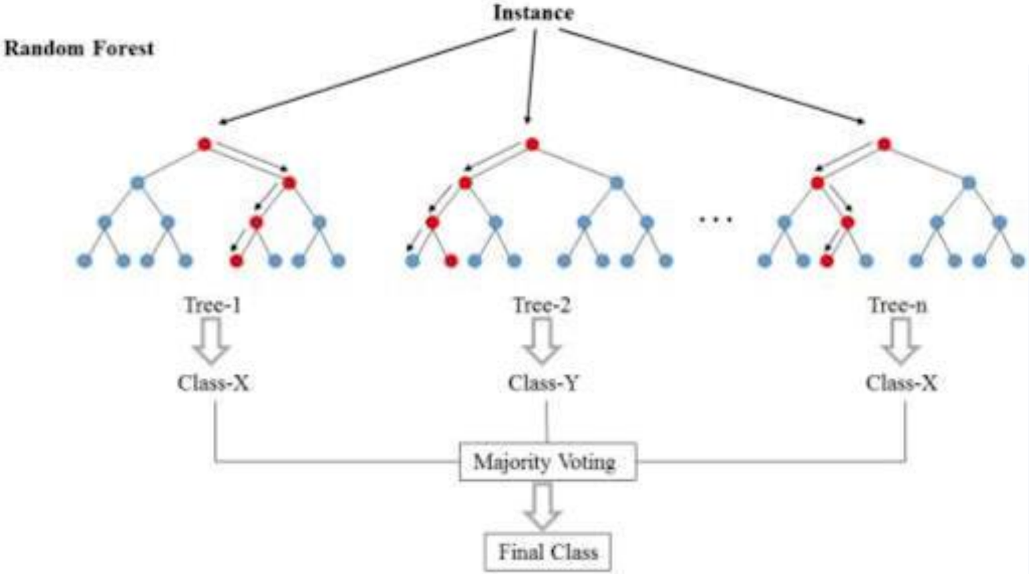


Final classifier is linear combination of weak classifiers

Random Forests

Association for Academic Surgery

- **Some ML methods really are ensembles**
- Example: Random Forests
- Ensemble of Trees
- Uses ***bagging:*** each classifier gets a vote (unweighted)
- Re-samples the training set
- Randomness to tree induction



Questions?

Wisconsin Surgical Outcomes Research Program
Department of Surgery
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Endocrine Surgery
Department of Surgery
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

schneiderd@surgery.wisc.edu