

Unsupervised Learning

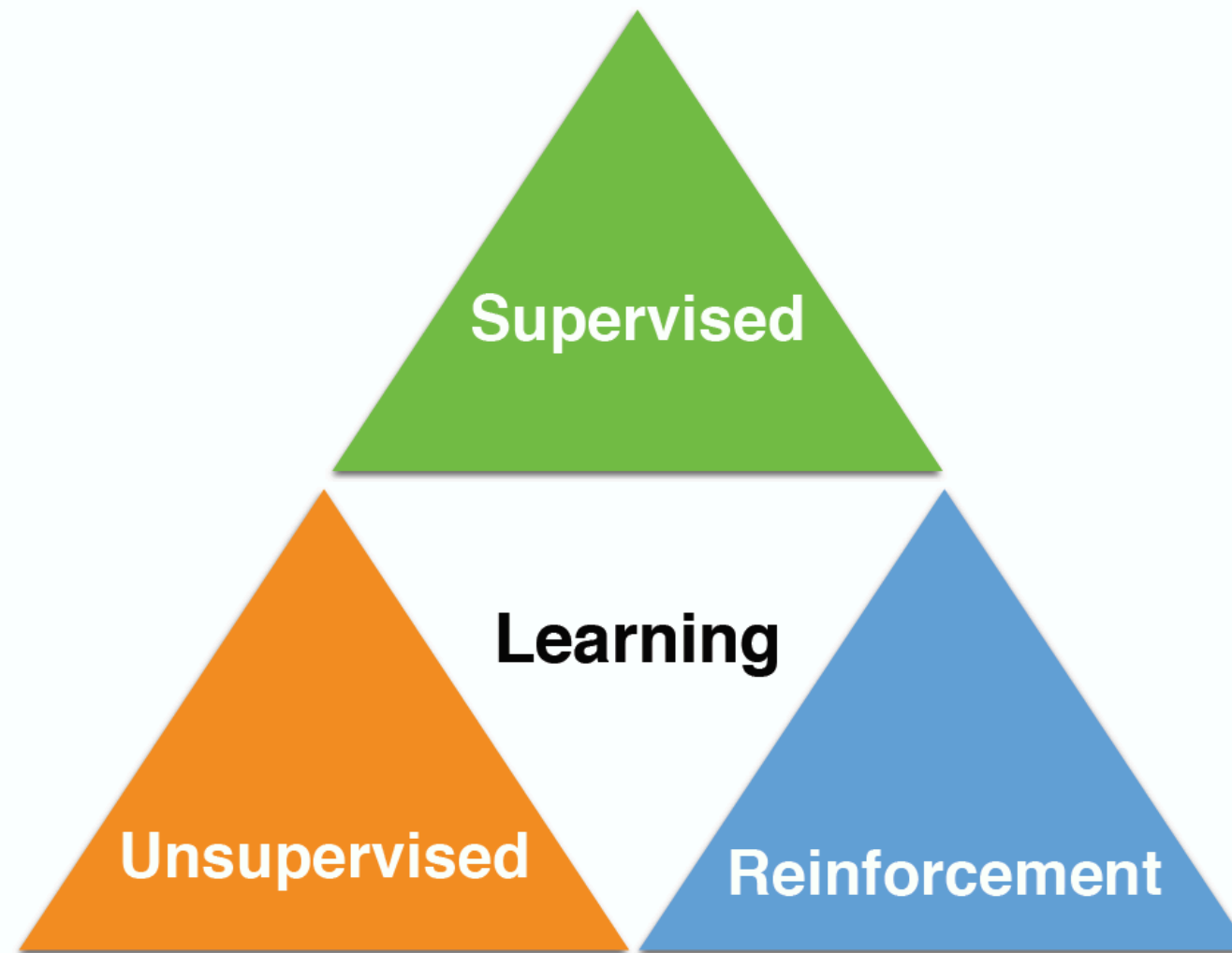
Lecture 1: Introduction and Evaluation

Tho Quan
qttho@hcmut.edu.vn

Agenda

- Introduction to supervised learning
- A case study with simple classification approach
- Evaluation the classification
- Make it practical: un upgraded version

- Labeled data
- Direct feedback
- Predict outcome/future



- No labels
- No feedback
- “Find hidden structure”

- Decision process
- Reward system
- Learn series of actions

THỂ GIỚI



Xác sinh viên gốc Việt được giấu trong một bức tường

(Dân trí) - Thi thể được tìm thấy giấu trong bức tường ở một tòa nhà của trường Đại học Yale, Mỹ, hồi cuối tuần qua đã được xác nhận là của Annie Le, sinh viên của trường bị mất tích ngay trước lễ cưới.

>> **Sinh viên gốc Việt mất tích bí Mỹ**

- Cảnh sát chống khủng bố lục soát nhiều căn nhà ở New York
- Mỹ lo ngại về hợp đồng mua vũ khí của Venezuela

THỂ THAO



Real Madrid, AC Milan trước "táp"

(Dân trí) - Xét về đẳng cấp, cả Real Madrid và AC Milan xứng đáng là UCV hàng đầu tại bán cầu Bắc. Chuyến hành quân đến sân của họ vào đêm nay sẽ là thử thách không nhỏ đối với các đội giàu truyền thống bậc nhất châu Âu.

- Nhà vô địch châu Âu thành "người thừa" ở giải Điền kinh
- Raul coi Champions League là mục tiêu sống còn

GIẢI TRÍ

Âm nhạc | Phim | Thời trang | Xem - Ăn - Chơi



"Fan cuồng" Dương Lệ Quyên lại "kết tội" Lưu Đức Hoa

(Dân trí) - "Fan cuồng" Dương Lệ Quyên bật khóc nức nở và gọi "thiên vương" Lưu Đức Hoa là kẻ giả dối khi hay tin anh kết hôn từ năm ngoái. Cô gái trẻ Dương Lệ Quyên gắn liền với sự kiện "cha nhảy sông tự vẫn để con được gặp thần tượng" năm 2007.

>> **Bố nhảy sông tự tử vì... con quá hâm mộ Lưu Đức Hoa**

[Xem tiếp](#)

- Thêm một lý do khiến Lưu Đức Hoa "bí mật" chuyện tình yêu
- Drew Barrymore với style lạ mắt

NHIP SỐNG TRÈ



Chàng trai thủ khoa từ bỏ... thủ khoa

(Dân trí) - Đỗ thủ khoa ĐH Quảng Bình, nhưng Đinh Anh Tuấn lại từ bỏ chỉ bởi đơn giản cậu học trò nghèo này muốn vào ĐH Khoa học Huế, vì theo Tuấn "học trường này phù hợp với mong muốn của mình hơn".

[Xem tiếp](#)

- Phan
- Xác sinh viên gốc Việt được giấu trong một bức tường

(Dân trí) - Drew Barrymore xuất hiện nổi bật tại LHP Toronto 2009 trong bộ váy vàng rườm rà, tóc nửa vàng nửa đen, móng tay xanh và móng chân đỏ...

Drew Barrymore tham dự LHP Toronto tại Canada để quảng cáo cho bộ phim mới của cô mang tên *Whip It*, đây là bộ phim do Drew Barrymore làm đạo diễn kiêm diễn viên. Trong phim còn có sự tham gia của ngôi sao xinh đẹp Ellen Page.

Ngày 13/9 vừa qua, Drew Barrymore và dàn diễn viên phim *Whip It* đã tới nhà hát Ryerson ở Toronto, Canada để dự buổi công chiếu phim mới của mình. Drew Barrymore là người nổi bật nhất trên thảm đỏ, cô diện một chiếc váy vàng rườm rà, cầu kỳ, họa tiết rắc rối hiệu Alexander McQueen. Mái tóc ngắn của cô nhuộm đen ở chân tóc, móng tay tô màu xanh và móng chân tô màu đỏ. Trông Drew Barrymore rất thú vị và thậm chí hài hước.



Thêm 8 trường ĐH công bố điểm chuẩn NV2

(Dân trí) - Sáng nay, 15/9, ĐH Ngoại thương, KHTN Hà Nội, KHXH&NV Hà Nội, ĐH Giáo dục - ĐHQG Hà Nội, ĐH Hồng Đức, Tây Bắc, Quy Nhơn và An Giang công bố điểm chuẩn NV2. Các trường ĐH Tây Bắc, Hồng Đức, Quy Nhơn, An Giang tiếp tục xét tuyển NV3.

Mức điểm chuẩn công bố tính cho thí sinh ở KV3, mỗi đối tượng ưu tiên kế tiếp giảm 1 điểm, khu vực ưu tiên kế tiếp giảm 0,5 điểm.

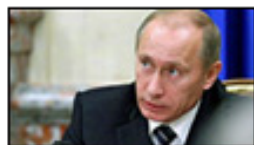
1. ĐH Tây Bắc

Ngành	Tên ngành	Khối thi	Điểm chuẩn NV2
<i>Các ngành đào tạo đại học</i>			
102	SP Tin	A	13,0
103	SP Vật lý	A	13,0

ФОРУМЫ ▶



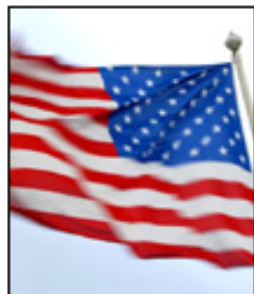
Как вы относитесь к моде на усыновление звездами сирот?



Путин подумывает о возвращении?

▶ О громких и нераскрытых убийствах и профессионализме следствия

ВАМ СЛОВО ▶



За что в России не любят Америку?

Станислав Чекалин:
"Откуда у многих россиян комплекс антиамериканизма?"

БЛОГИ ▶



Северный Кавказ глазами блоггеров

Так положено в чеченском обществе – дети после развода всегда остаются с отцом...

▶ Михаил Белый: о хамстве на дорогах, штрафах и взятках

LIVE_REPORT ▶



Неделя глазами Live_Report

Где побывали участники сообщества Live_Report? Соревнования по гребле, акция "Кто пьет, тот не водит" и многое другое...

За что в России не любят Америку?

Откуда у многих россиян комплекс антиамериканизма? Что нужно сделать, чтобы сломать этот стереотип?

Станислав Чекалин, РФ, Калининград

[Присылайте свои темы](#)

Не забудьте оставить свой номер телефона, если вы хотите принять участие в передаче.

Звоните нам в прямой эфир по телефонам:

Supervised vs. Unsupervised Learning

- *Given*
 - $\{x_i\}$, x_i is description of an object in some space, $i = 1, 2, \dots, n$.
 - y_i is some property of x_i viewed as its label, $y_i \in \{C_1, C_2, \dots, C_K\}$ or $y_i \in \mathbb{R}$
 - $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- *Find*

Function $p(y|x)$ for label data
and $p(x)$ for unlabeled data



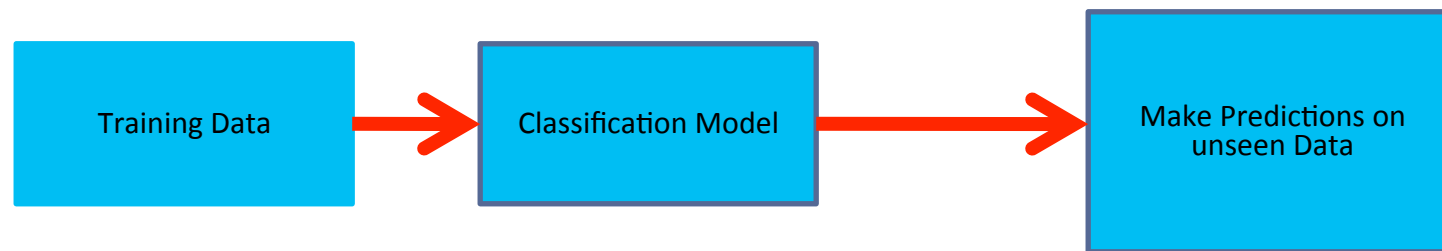
(Source: Eric Xing lecture notes)

Branch of Artificial Intelligence: Reasoning, language understanding, learning

DEDUCTION [Given $f(x)$ and x_i , deduce $f(x_i)$] vs. *INDUCTION* [Given $\{x_i\}$, infer $f(x)$]

Classification: Overview

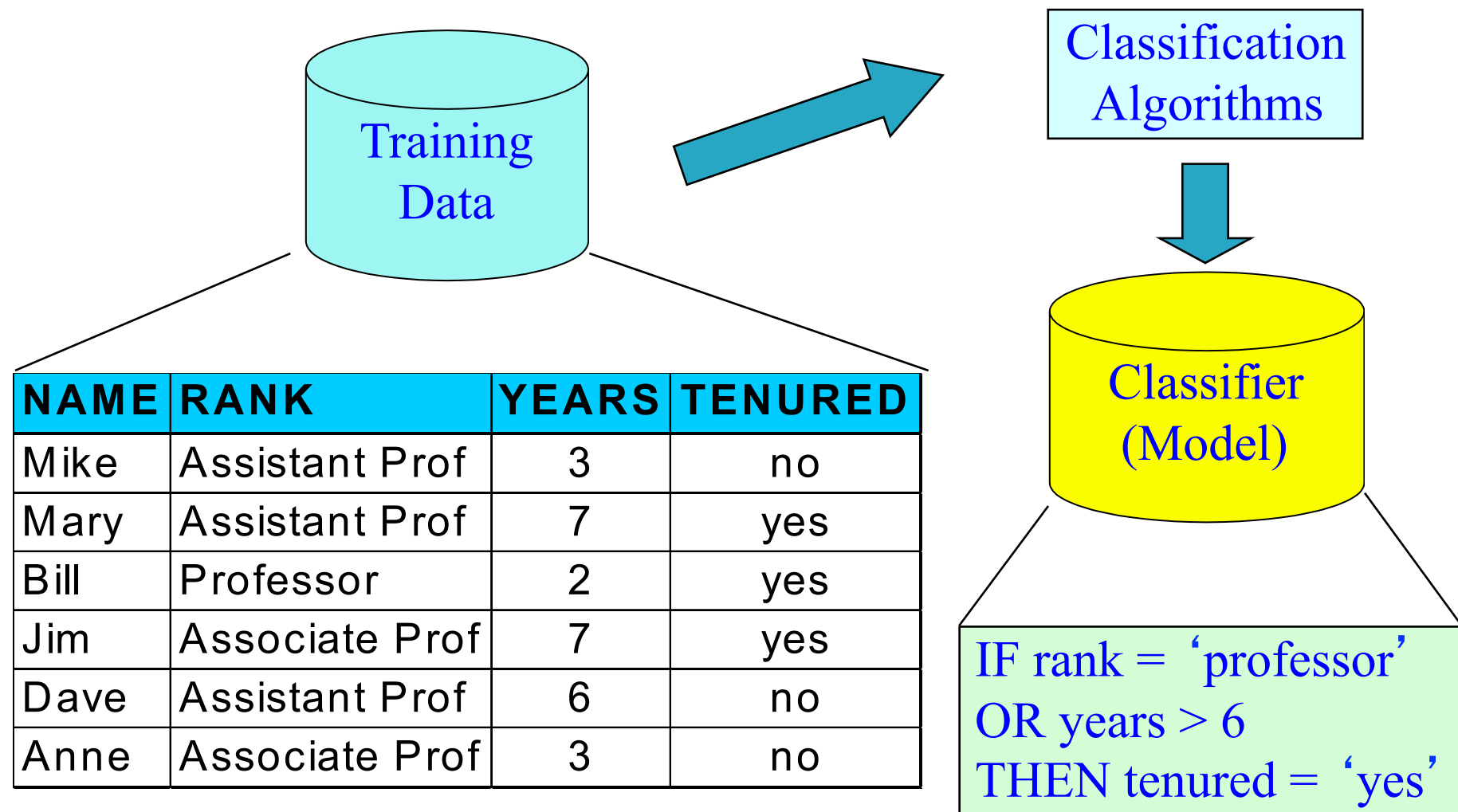
- Classification



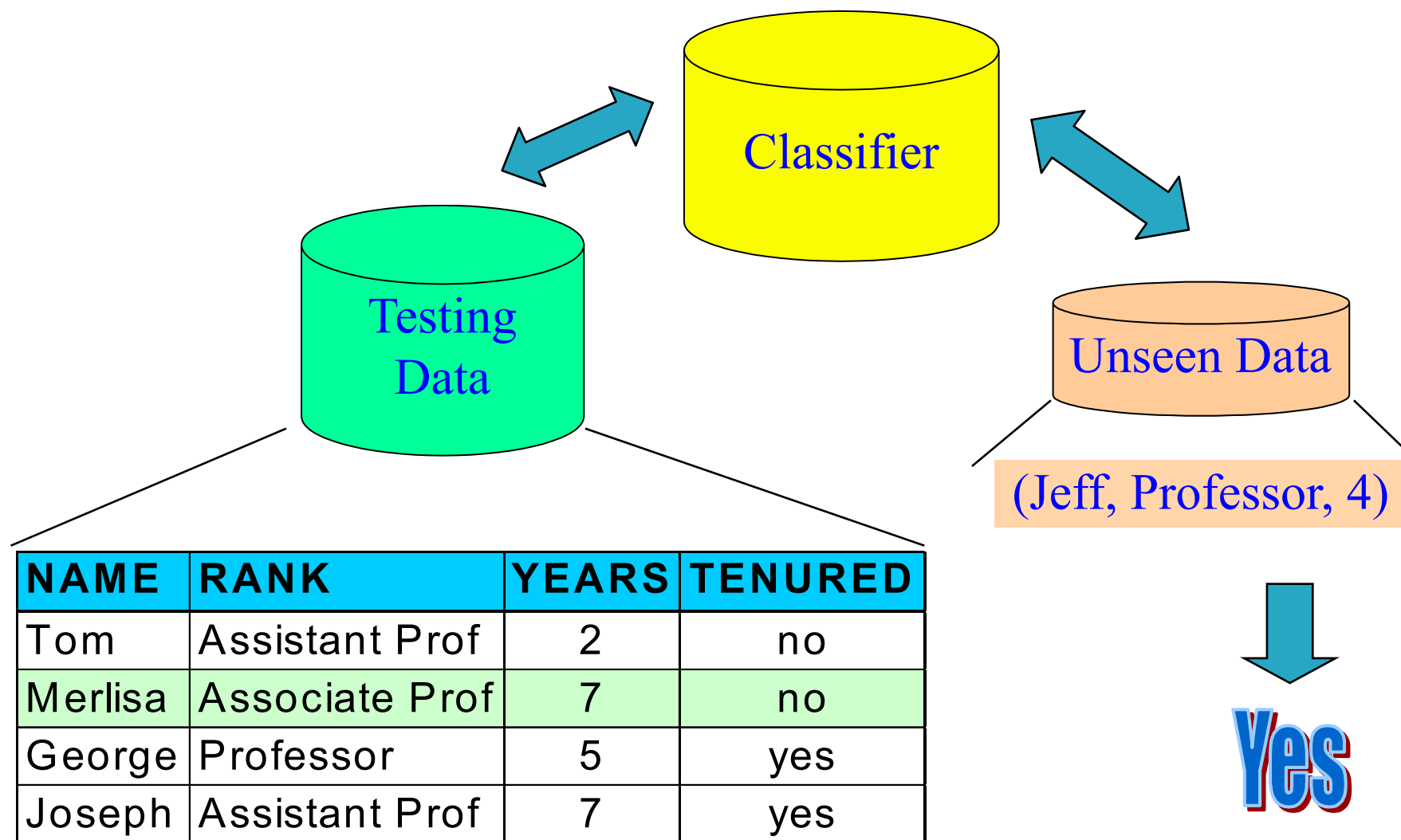
- Example :

- Play goft
- Neuron Network

Process 1: Model Construction



Process 2: Using the Model in Prediction



Classification Algorithms

- Classification Algorithms:
 - Support Vector Machines
 - Neural Network (multi-layer perceptron)
 - Decision Tree
 - K-Nearest Neighbor
 - Naive Bayes Classifier...

K-Nearest Neighbor

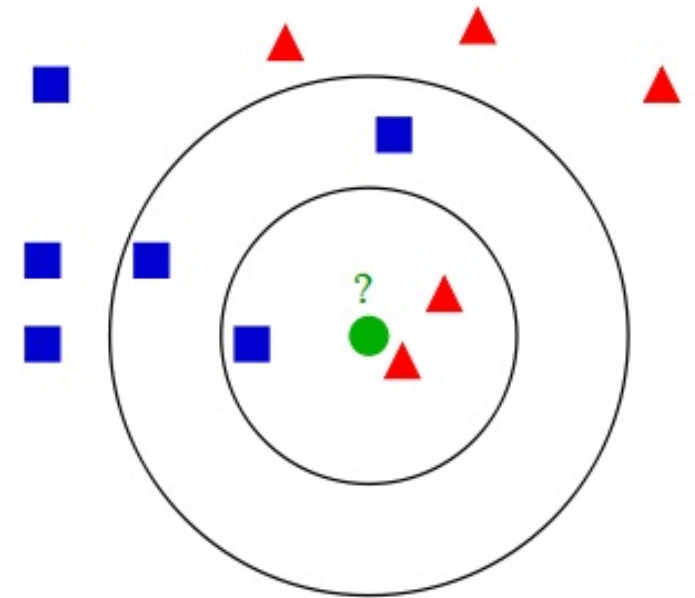
- Classifying objects based on closest training examples in the **feature space**
- Approximated locally
- All computation is deferred until classification
- Classified by a majority vote of its neighbors

Data

- The training examples are vectors in a multidimensional feature space, each with a class label
- The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples

Algorithm

- k is a user-defined constant
- Choose k nearest neighbors (**NNS**)
- Label is label which is most frequent among the k training samples nearest.



K-Nearest Neighbor

- Classifying objects based on closest training examples in the **feature space**
- Approximated locally
- All computation is deferred until classification
- Classified by a majority vote of its neighbors

Similarity and Representation

- How do we define similarity?
- How do we represent the objects whose similarities we wish to measure?

Motivation for the VSM

- We get back to the case study of document classification
 - VSM is an algebraic model for representing text documents as vectors of **index term**
 - A document is represented as a vector. Each dimension corresponds to a separate term. Its values are **tf-idf weighting**

Document Collection

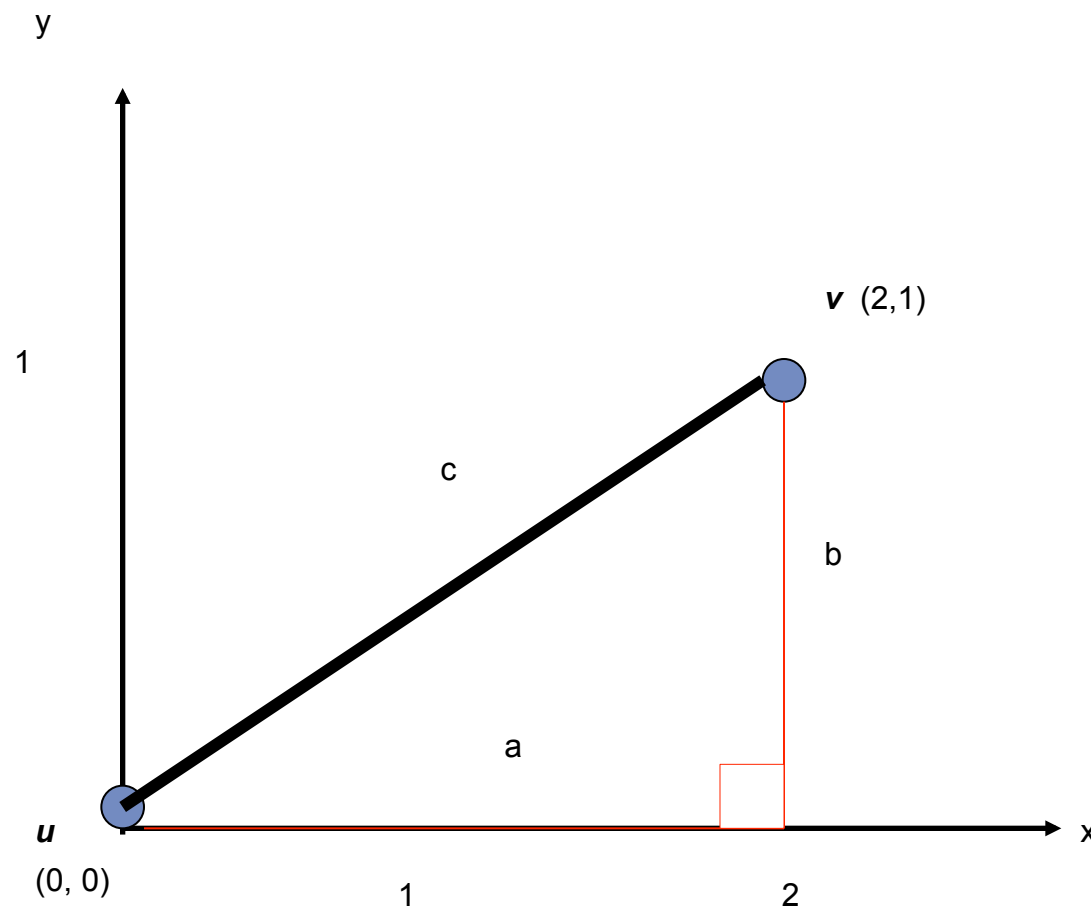
- A collection of n documents can be represented in the vector space model by a term-document matrix.
- An entry in the matrix corresponds to the “weight” of a term in the document; zero means the term has no significance in the document or it simply doesn’t exist in the document.

	T_1	T_2	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
($:$	$:$	$:$		$:$
$:$	$:$	$:$		$:$
D_n	w_{1n}	w_{2n}	...	w_{tn}

Measuring Distance

- Euclidean Distance
- Manhattan Distance
- Cosine Similarity
- Vector Length & Normalization

Euclidean Distance



$$a^2 + b^2 = c^2$$

$$2^2 + 1^2 = c^2$$

$$c = \sqrt{5}$$

Euclidean Distance

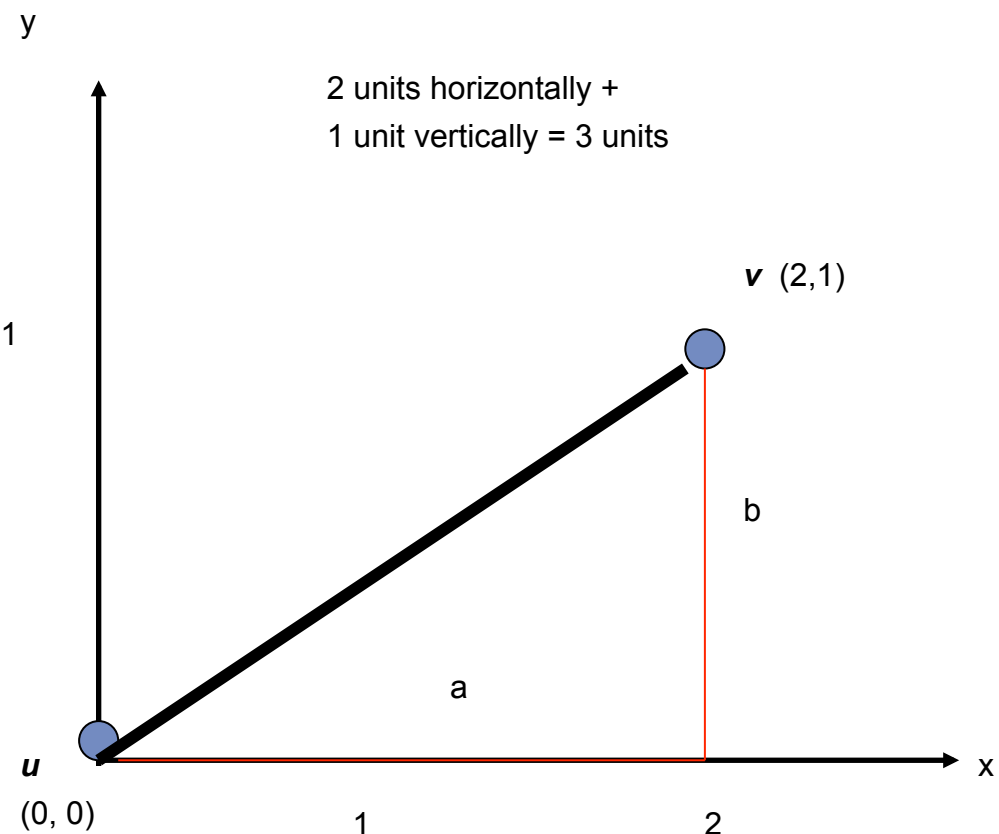
In p dimensions, let Euclidean distance between two points u and v be:

$$\text{dist}(u, v) = \sqrt{\sum_{i=1}^p (u_i - v_i)^2}$$

Manhattan Distance

- The Manhattan distance (also known as a city block distance) is the **number of units** on **a rectangular grid** it takes to travel from point u to v .

$$D_M(u, v) = \sum_{i=1}^p |u_i - v_i|$$



Cosine Similarity

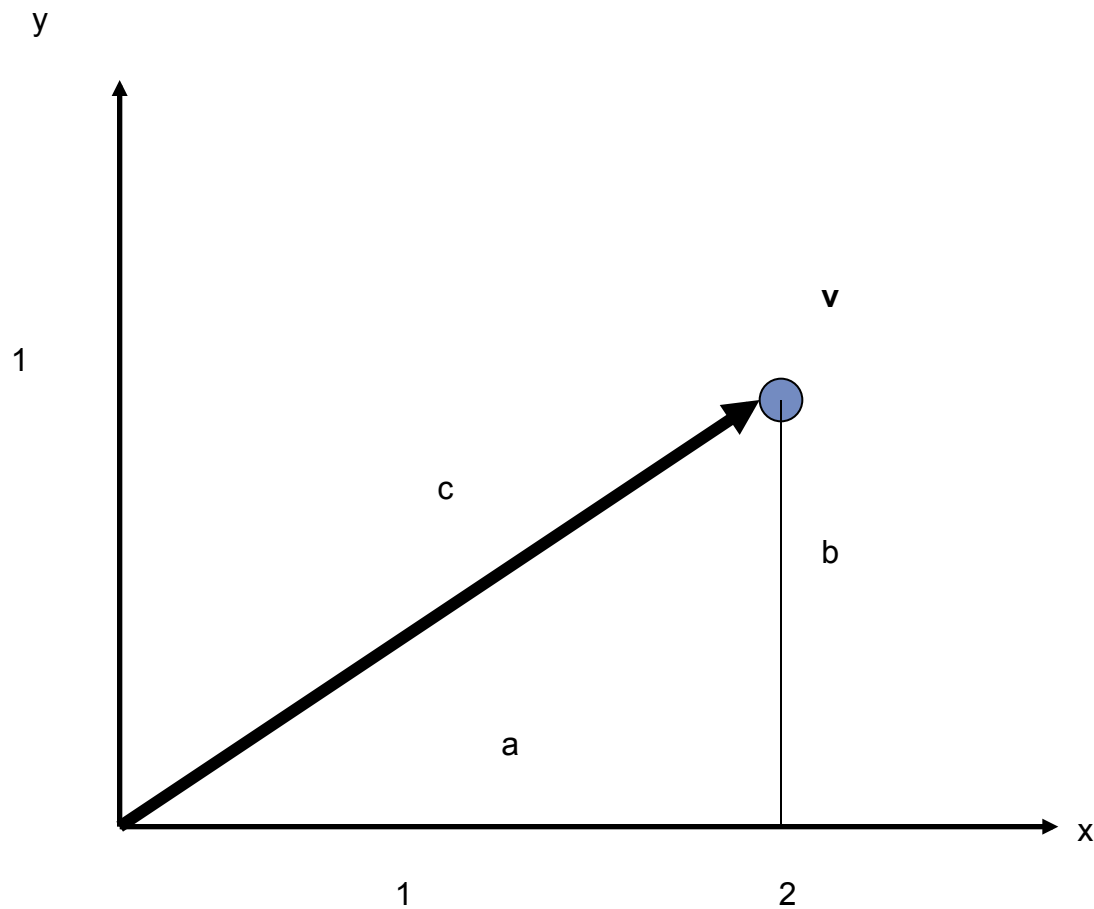
- The traditional vector space model is based on a different notion of similarity: **the cosine of the angle between two vectors**.
- To start our consideration of cosine similarity, consider our document vectors not as points in term space, but as arrows that travel from the origin of the space to a particular address.

Calculating Cosine(x, y)

$x(x_1, \dots, x_p)$ & $y(y_1, \dots, y_p)$ are two
vectors in p-dimensions:

$$\begin{aligned}\cos(x, y) &= \frac{x \cdot y}{\sqrt{\left(\sum_{i=1}^p x_i \cdot x_i\right)} \sqrt{\left(\sum_{i=1}^p y_i \cdot y_i\right)}} \\ &= \frac{\sum_{i=1}^p x_i \cdot y_i}{\sqrt{\left(\sum_{i=1}^p x_i \cdot x_i\right)} \sqrt{\left(\sum_{i=1}^p y_i \cdot y_i\right)}} \quad (1)\end{aligned}$$

Vector Length and Normalization



$$a^2 + b^2 = c^2$$

$$2^2 + 1^2 = c^2$$

$$c = \sqrt{5}$$

Vector Length and Normalization

- Vector length: $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^p v_i^2} = \sqrt{\mathbf{v} \cdot \mathbf{v}}$

- Normalization: $\mathbf{u} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$

Thus

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Using the Cosine for IR

- From now on, unless otherwise specified, we can assume that **all document vectors** in our term-document matrix **A** have been **normalized to unit length**.
- Likewise we always normalize our query to unit length.
- Given these assumptions, we have the classic vector space model for IR.

Accuracy Measures

- A natural criterion for judging the performance of a classifier is the probability for making a misclassification error.
- Misclassification
 - The observation belongs to one class, but the model classifies it as a member of a different class.
- A classifier that makes no errors would be perfect
 - Do not expect to be able to construct such classifiers in the real world
 - Due to “noise”
 - Not having all the information needed to precisely classify cases.

Accuracy Measures

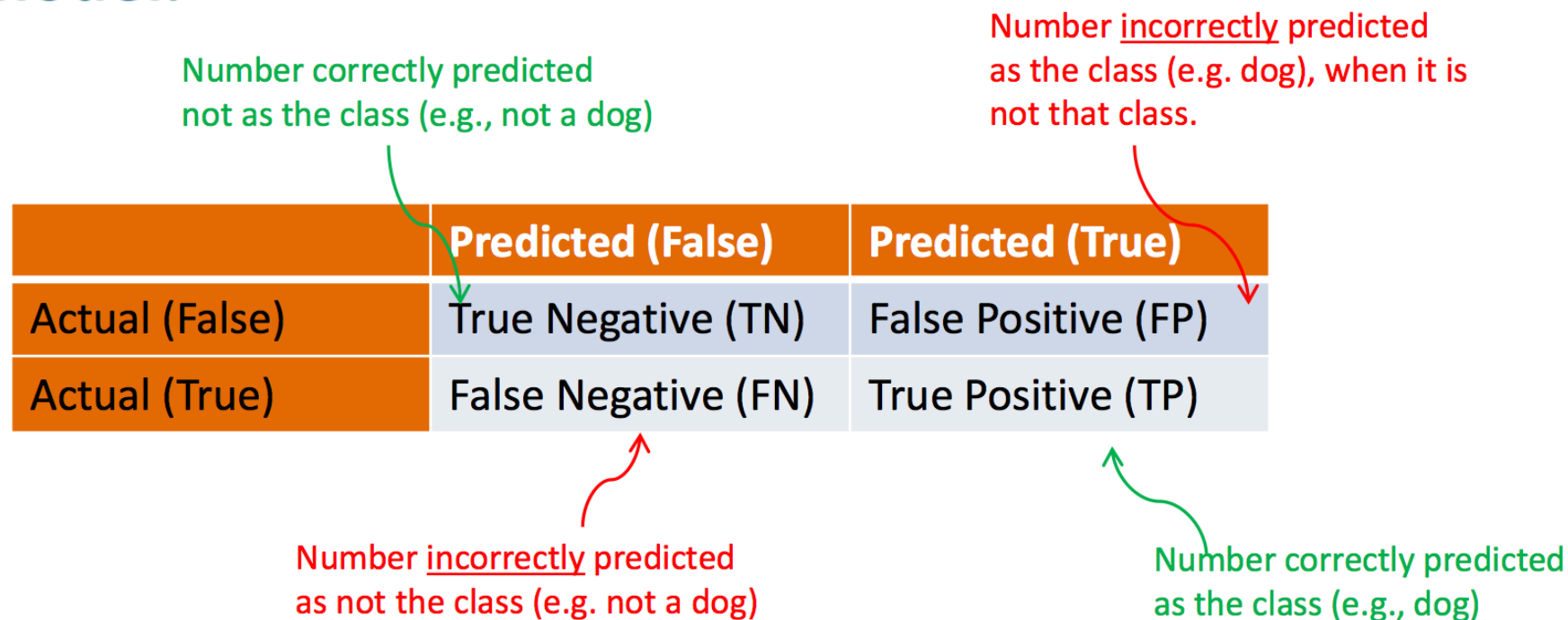
- To obtain an honest estimate of classification error, we use the classification matrix that is computed from the validation data.
 - We first partition the data into training and validation sets by random selection of cases.
 - We then construct a classifier using the training data,
 - Apply it to the validation data,
 - Yields predicted classifications for the observations in the validation set.
 - We then summarize these classifications in a classification matrix.

Problem with accuracy

- If the training and test data are skewed towards one classification, then the model will predict everything as being that class.
 - In Titanic training data, 68% of people died. If one trained a model to predict everybody died, it would be 68% accurate.

Confusion Matrix

Four Quadrant Measurement on “Performance” of a model.



The diagram shows a 2x2 confusion matrix with orange headers and light blue data cells. Annotations with arrows explain each quadrant: True Negative (TN) is 'Number correctly predicted not as the class (e.g., not a dog)'; False Positive (FP) is 'Number incorrectly predicted as the class (e.g. dog), when it is not that class.'; False Negative (FN) is 'Number incorrectly predicted as not the class (e.g. not a dog)'; and True Positive (TP) is 'Number correctly predicted as the class (e.g., dog)'.

	Predicted (False)	Predicted (True)
Actual (False)	True Negative (TN)	False Positive (FP)
Actual (True)	False Negative (FN)	True Positive (TP)

- **Accuracy** = $(TP + TN) / N$
- **Misclassification** = $(FP + FN) / N$
- **Precision** = $TP / (TP + FP)$

Example
Measurements

		Predicted Class	
		C_0	C_1
Actual Class	C_0	$n_{0,0}$ = Number of correctly classified C_0 cases	$n_{0,1}$ = Number of C_0 cases incorrectly classified as C_1
	C_1	$n_{1,0}$ = Number of C_1 cases incorrectly classified as C_0	$n_{1,1}$ = Number of correctly classified C_1 cases

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

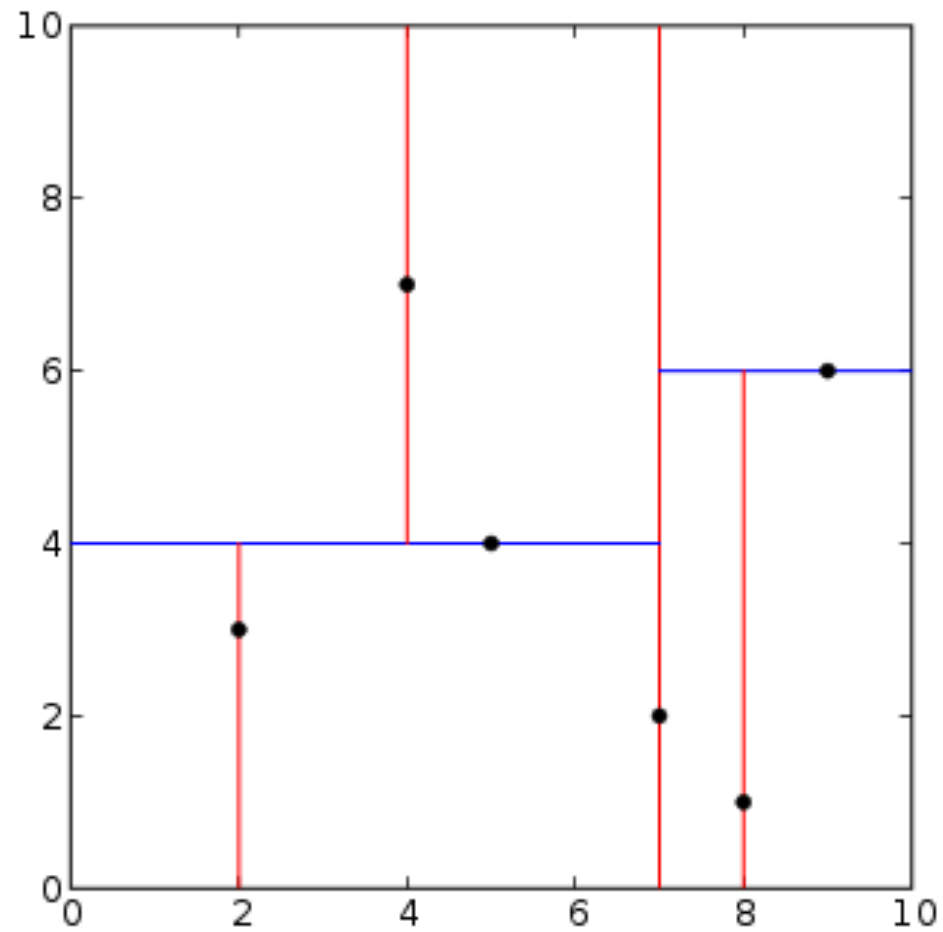
Practical Metrics for Data Science Projects

- Case study: in the Flight Delay project with NTU, we applied the following metrics
 - Accuracy
 - Precision
 - Recall
 - F-measure

Practical k-NN

- Need a data structure for **calculate** near neighbors fast.
- Data structure
 - **KD-tree**

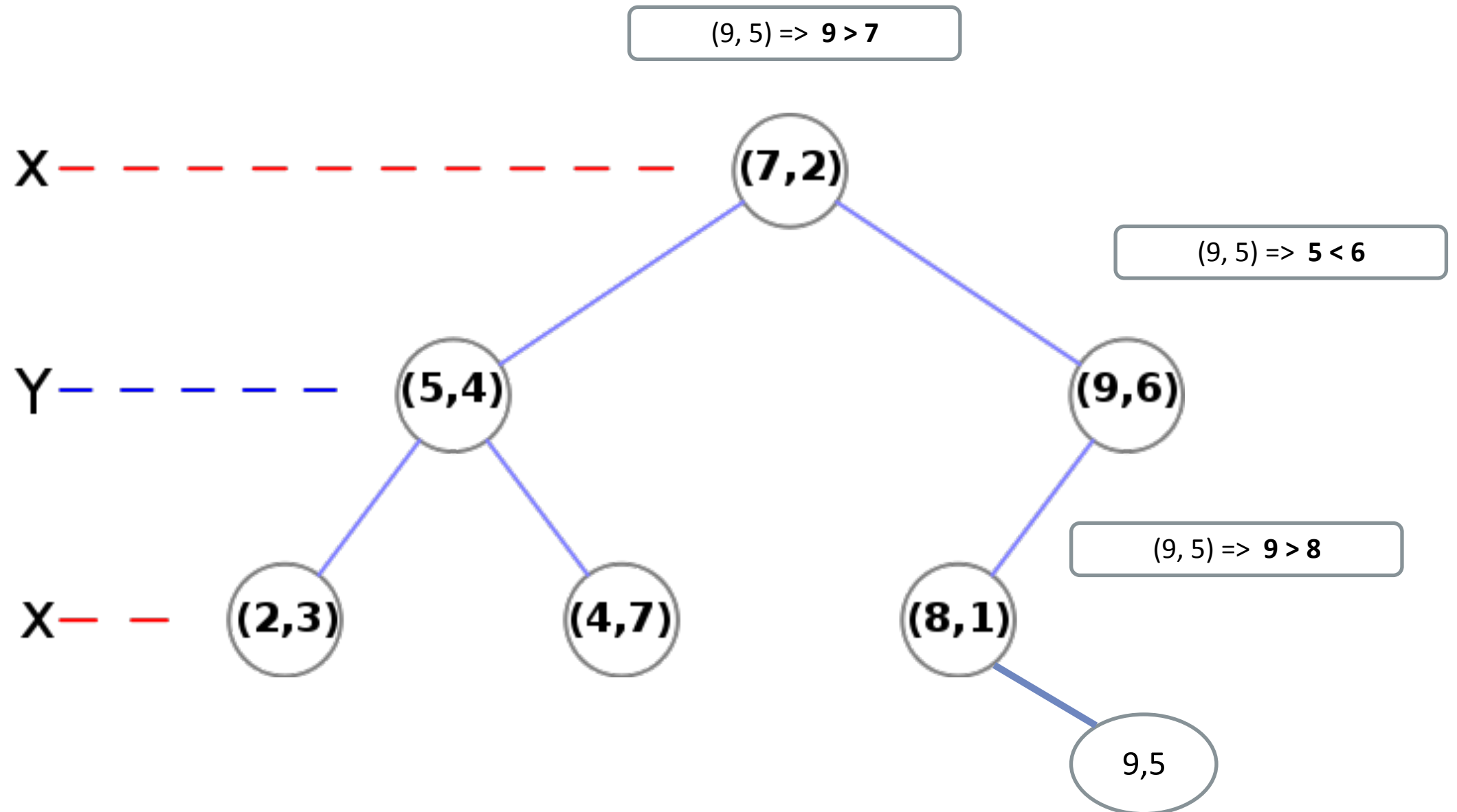
Example: 2-D Tree



2-d Tree for these points :

$(2, 3)$, $(5, 4)$, $(9, 6)$, $(4, 7)$, $(8, 1)$, $(7, 2)$

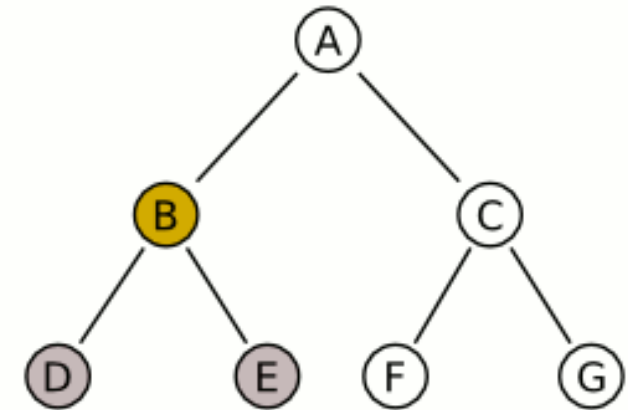
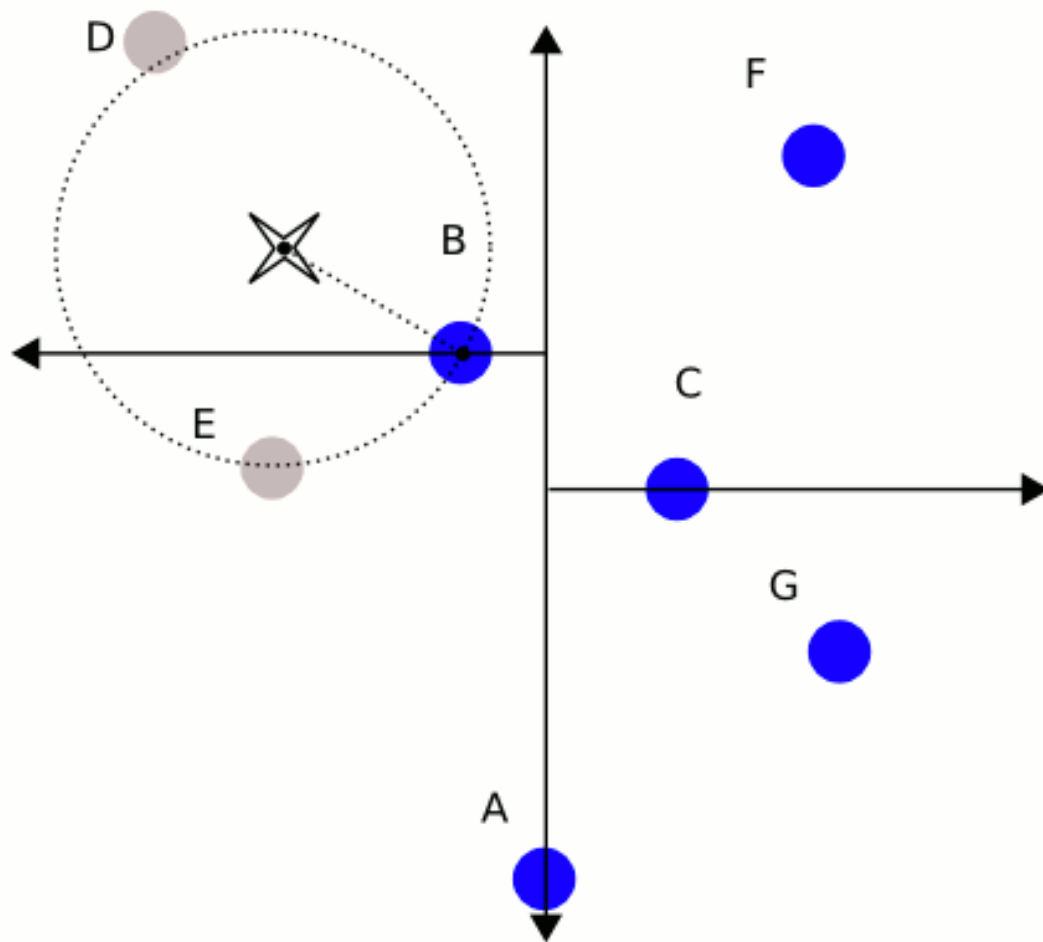
Tree Operation – Insert



Complexity

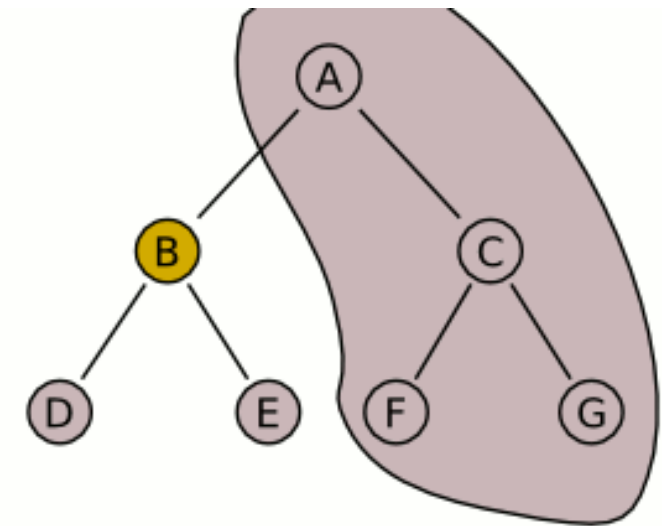
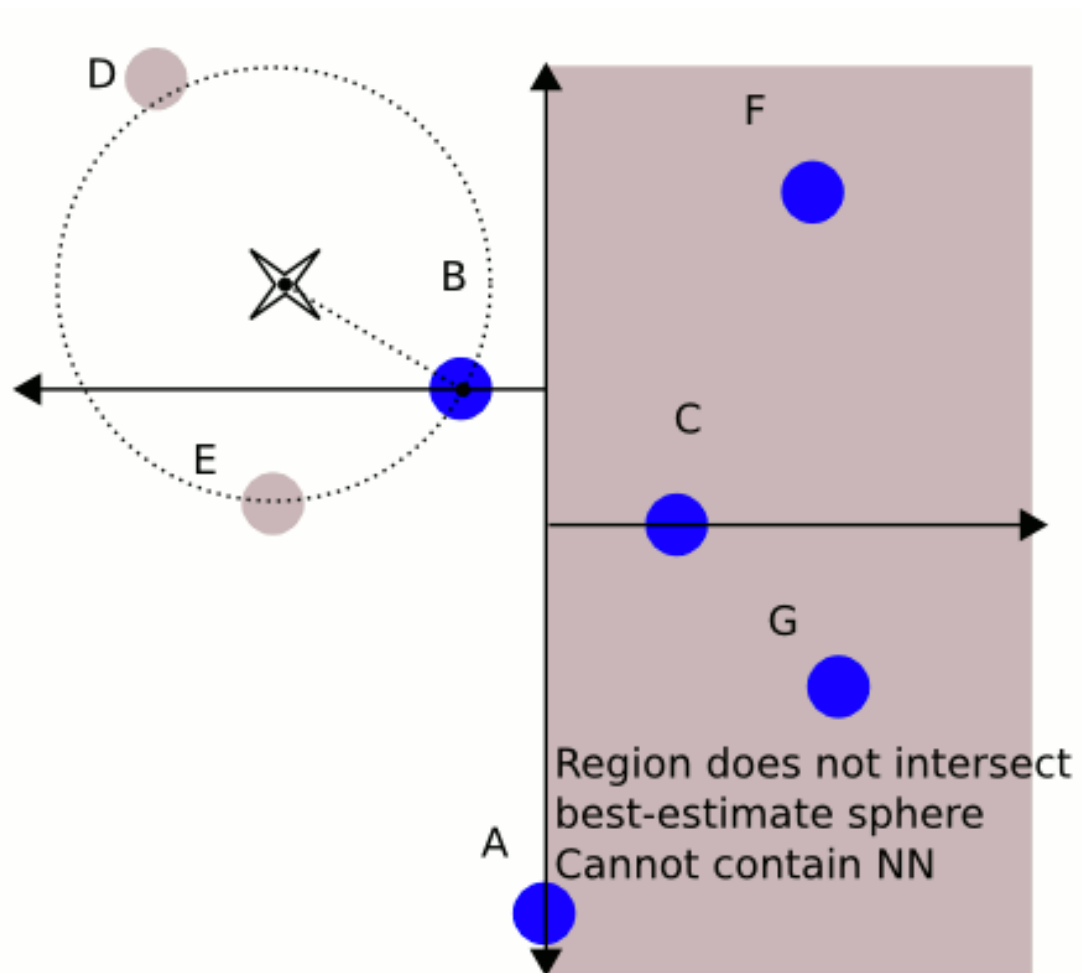
- Building a static k-d tree from n points $O(n \log n)$
- Inserting a new point into a balanced k-d tree takes $O(\log n)$ time.
- Removing a point from a balanced k-d tree takes $O(\log n)$ time.
- Finding 1 nearest neighbour in a balanced k-d tree with randomly distributed points takes $O(\log n)$ time on average.

Search Nearest Neighbor



D & E Discarded as B
(already visited) is closer.
B is the best estimate for B's sub-branch
Proceed back to parent node

Search Nearest Neighbor



A's children have all been searched,
B is the best estimate for entire tree

Search K-Nearest Neighbor

- Apply the Search Nearest Neighbor with maintain the list of points.
- **Kd**-trees are not suitable for efficiently finding the nearest neighbour in high-dimensional spaces and approximate nearest-neighbour methods should be used instead.