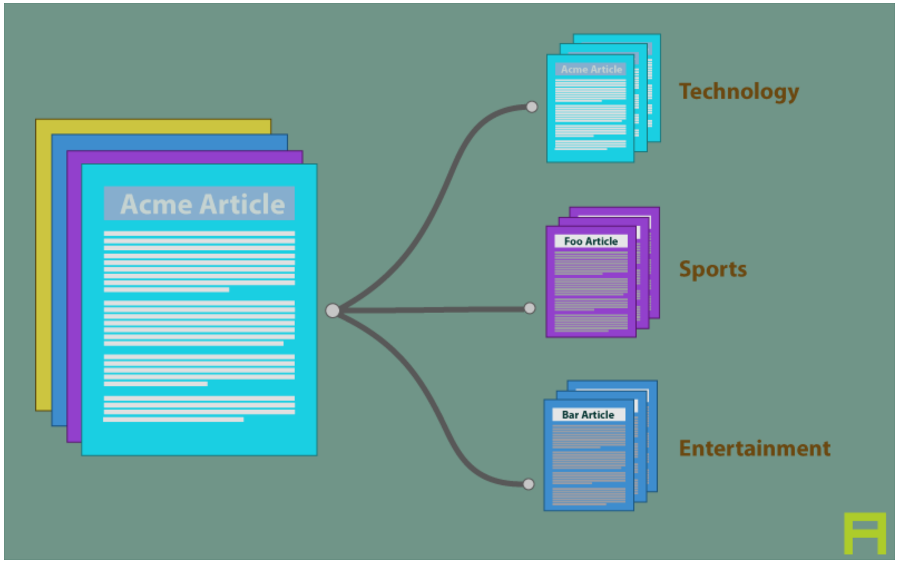


Lecture 9-1: Document Classification Part 1: Vector Space Models

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

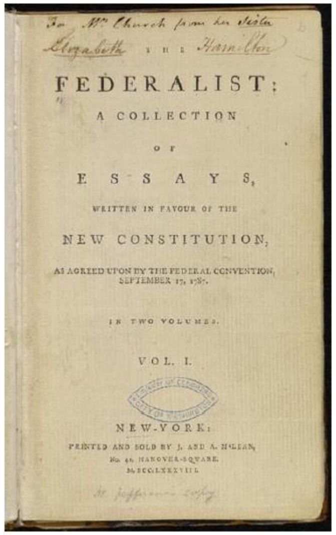
01 Document Classification: Overview

02 Naive Bayesian Classifier

03 k-Nearest Neighbor Classifier

Document Classification: Examples

- Who wrote which federalist papers?
  - ✓ 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton
  - ✓ Authorship of 12 of the letters in dispute
  - ✓ 1963: solved by Mosteller and Wallace using Bayesian methods



To Mrs Church from her Sister
Elizabeth THE Hamilton

FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

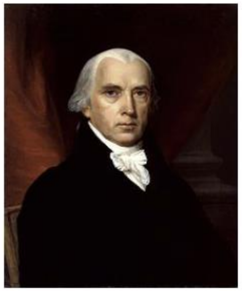
IN TWO VOLUMES.

VOL. I.

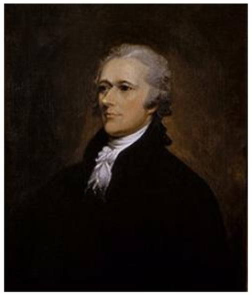
LIBRARY OF CONGRESS
CITY OF WASHINGTON

NEW-YORK:
PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE,
M,DCC,LXXXVIII.

Mr Jefferson's copy



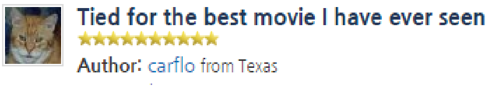
James Madison

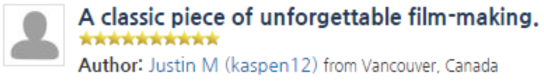


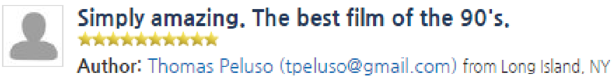
Alexander Hamilton

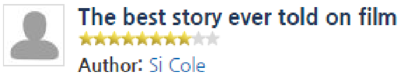
Document Classification: Examples

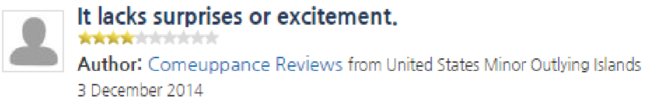
- Positive or Negative reviews?

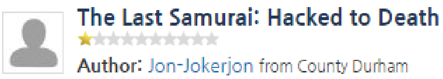

Tied for the best movie I have ever seen
★★★★★★★★★★
Author: carflo from Texas


A classic piece of unforgettable film-making.
★★★★★★★★★★
Author: Justin M (kaspen12) from Vancouver, Canada


Simply amazing. The best film of the 90's.
★★★★★★★★★★
Author: Thomas Peluso (tpeluso@gmail.com) from Long Island, NY

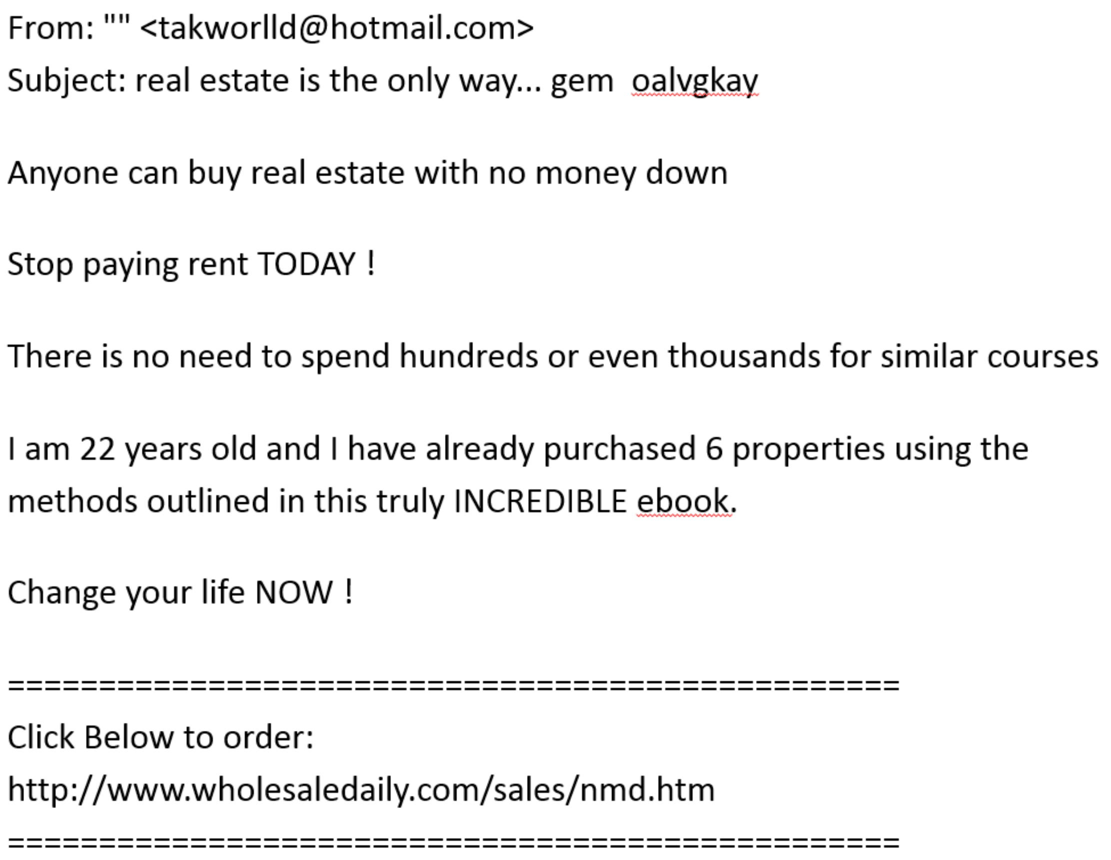

The best story ever told on film
★★★★★★★★☆☆
Author: Si Cole


It lacks surprises or excitement.
★★★★☆☆☆☆☆☆
Author: Comeuppance Reviews from United States Minor Outlying Islands
3 December 2014


The Last Samurai: Hacked to Death
★☆☆☆☆☆☆☆☆☆
Author: Jon-Jokerjon from County Durham

Document Classification: Examples

- Spam or not?



From: "" <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

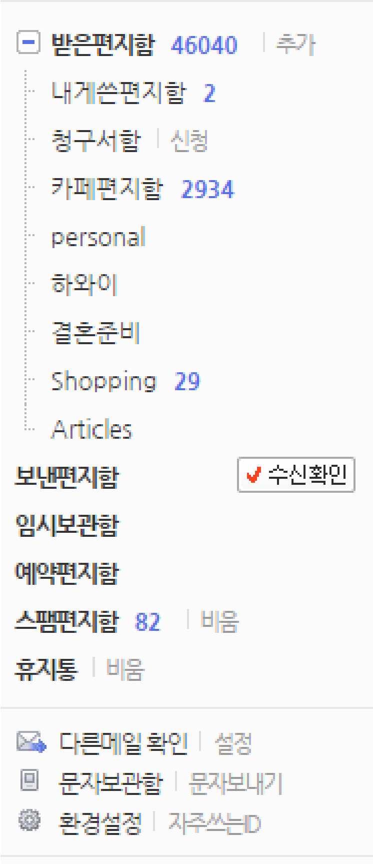
Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

===
Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
===



Document Classification: Examples

- What is the subject of this article?

MEDLINE Article

Available online at www.sciencedirect.com

Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Susanne Gahl,[a] Lise Menn,[b] Gail Ramsberger,[b] Daniel S. Jurafsky,[b] Elizabeth Elder,[b] Molly Rewega,[a] and L. Holland Audrey[a]

[illegible]

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Document Classification: Applications

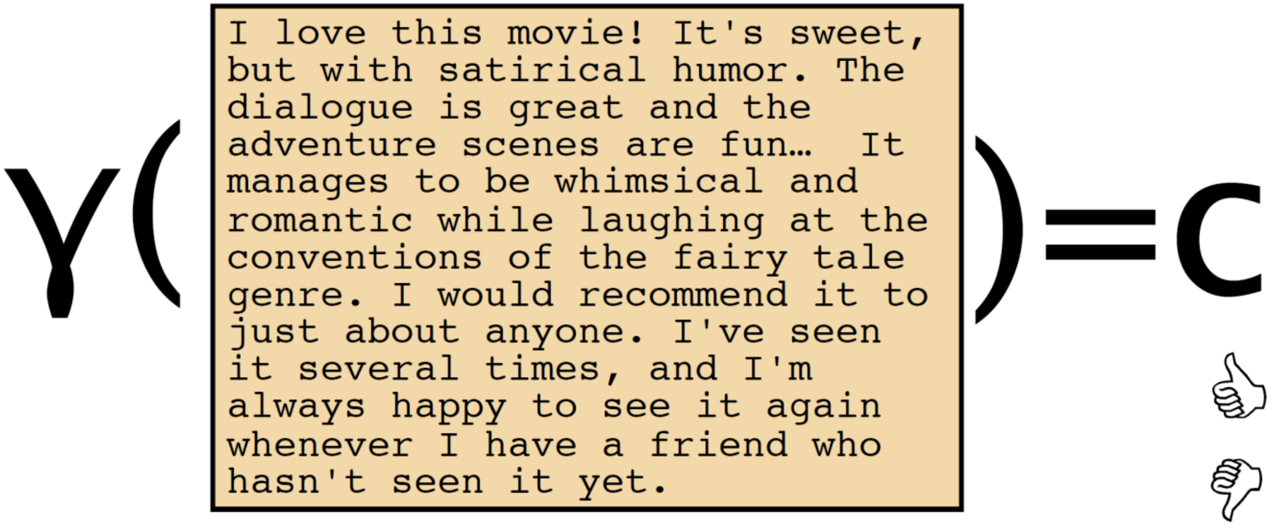
- Document Classification/Categorization
  - ✓ Assigning subject categories, topics, or genres
  - ✓ Spam detection
  - ✓ Authorship identification
  - ✓ Age/gender identification
  - ✓ Language identification
  - ✓ Sentiment analysis
  - ✓ …

Document Classification: Definition

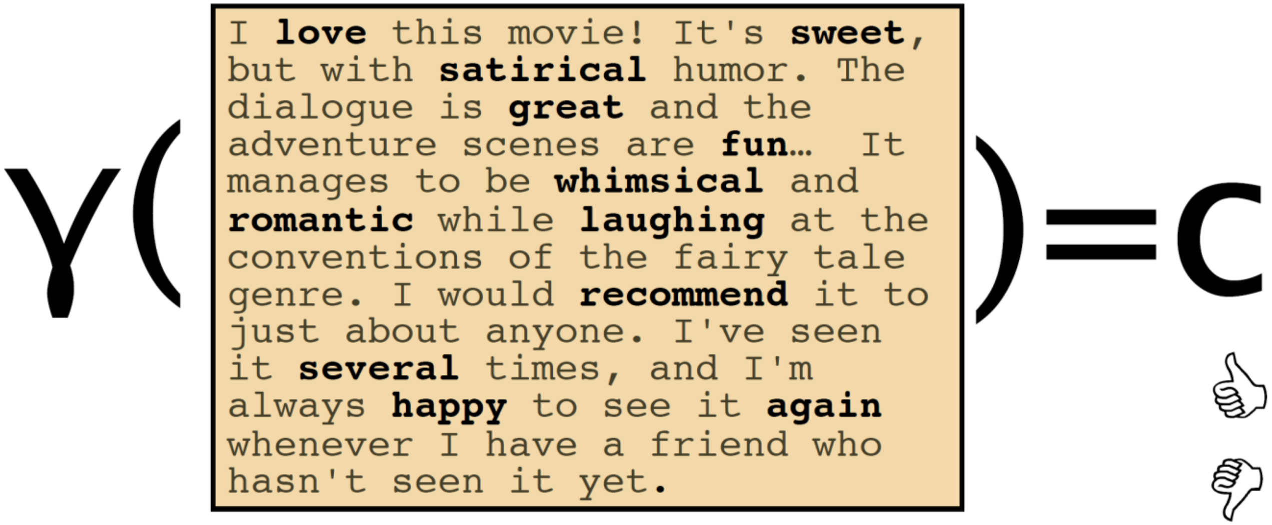
- Document Classification
  - ✓ Input (I)
    - ▪ A document d
    - ▪ A fixed set of classes C = {c_1, c_2, …, c_j}
  - ✓ Output (O)
    - ▪ A predicted class c in C
- Supervised Machine Learning
  - ✓ Input: I + a training set of m hand-labeled documents: (d_1, c_1), …, (d_m, c_m)
  - ✓ Output: a learned classifier y: f(d) = c

Document Classification: Definition

- Supervised Machine Learning with Bag-of-Words representation



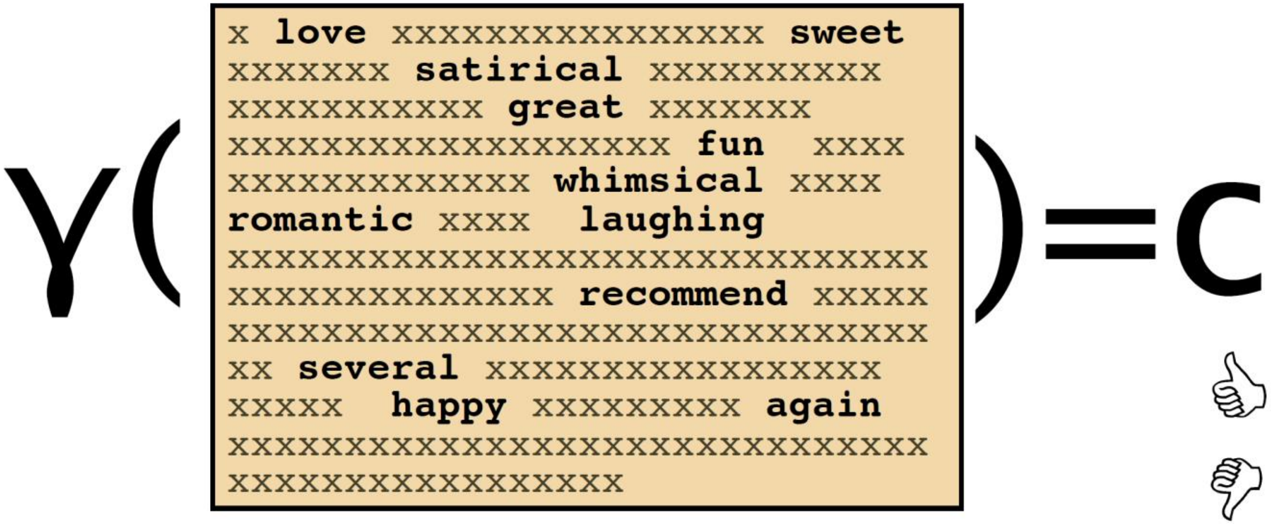
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.



I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun…** It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet.

Document Classification: Definition

- Supervised Machine Learning with Bag-of-Words representation



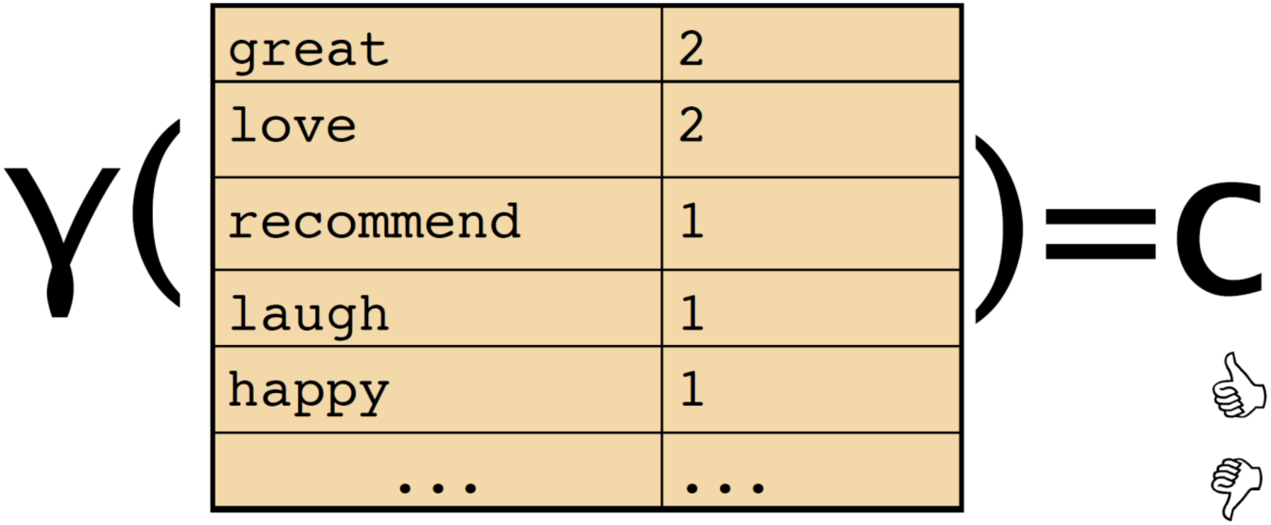
$\gamma($

```
x love xxxxxxxxxxxxxxxx sweet
xxxxxxx satirical xxxxxxxxxx
xxxxxxxxxxx great xxxxxxx
xxxxxxxxxxxxxxxxxxx fun  xxxx
xxxxxxxxxxxxx whimsical xxxx
romantic xxxx  laughing
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx recommend xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx several xxxxxxxxxxxxxxxxx
xxxxx  happy xxxxxxxxx again
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx
```

$)=c$

👍

👎



$\gamma($

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

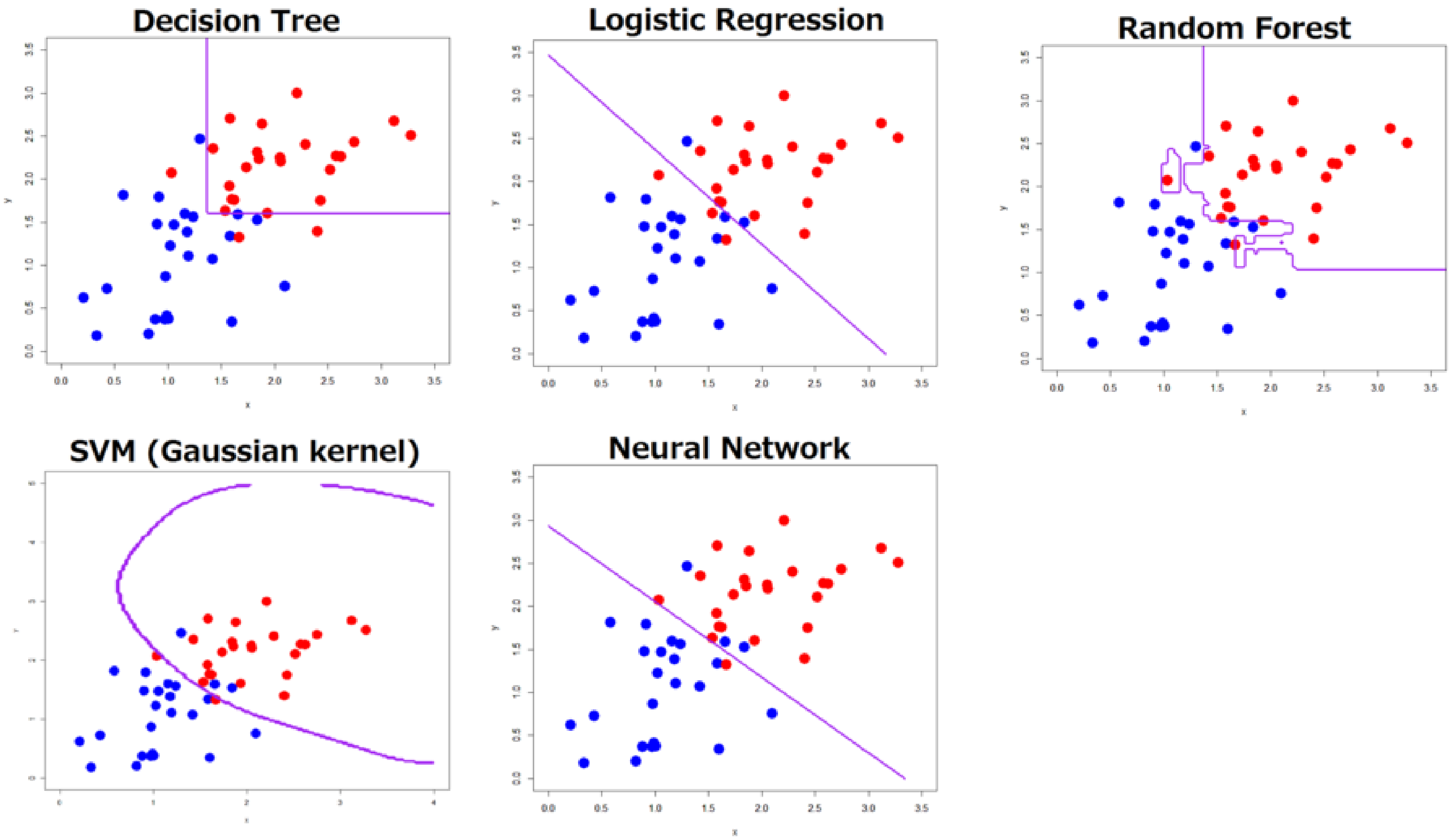
$)=c$

👍

👎

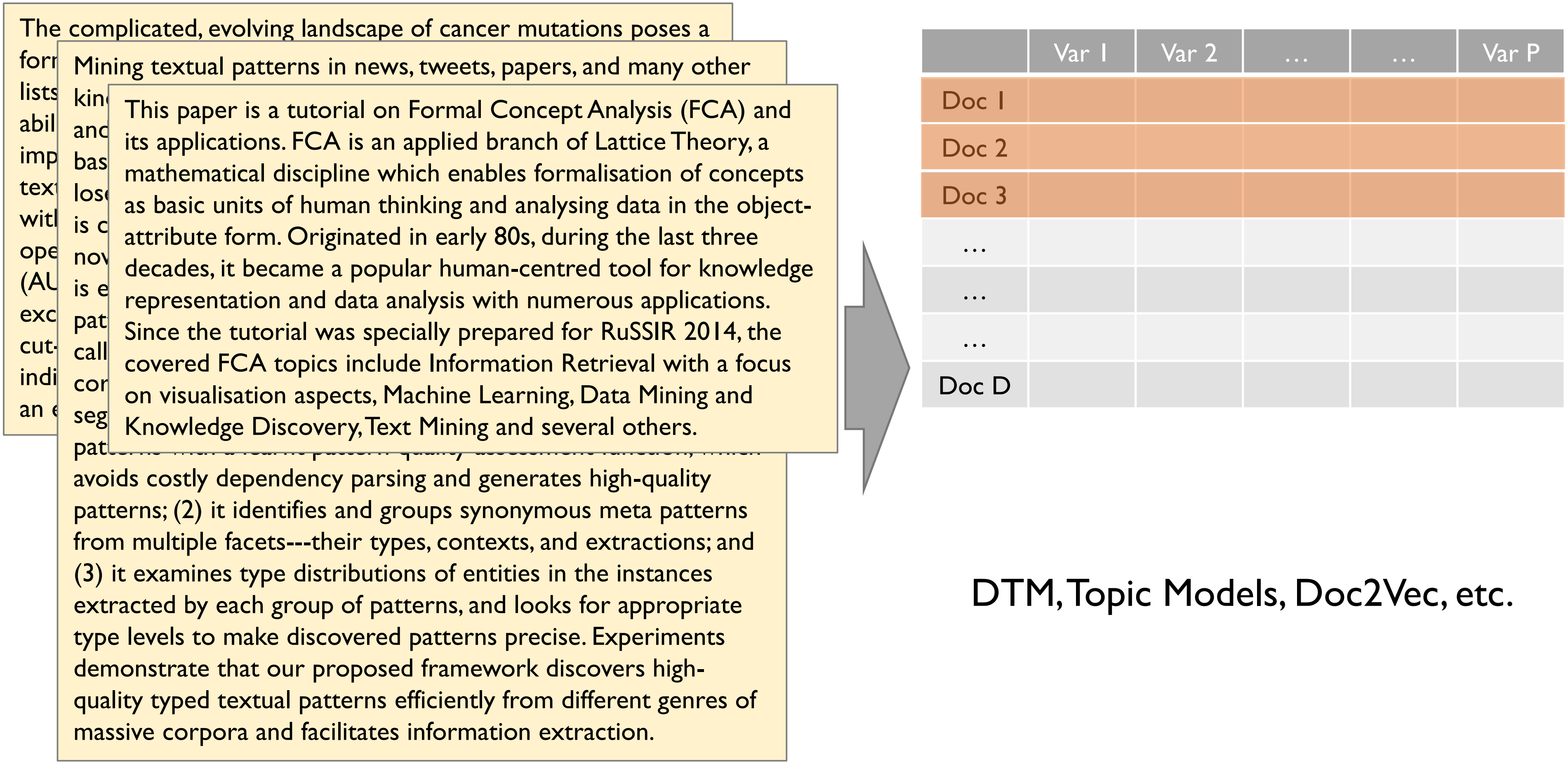
Why Are There So Many Classifiers?

- We cannot guarantee that a single classifier is always better than the others



Vector Space Model vs. Matrix-based Model

- Vector Space Model
  - ✓ A single document is transformed into a single vector



The complicated, evolving landscape of cancer mutations poses a
forr
lists
abil
imp
text
witl
ope
(AU
exc
cut-
indi
an e

Mining textual patterns in news, tweets, papers, and many other
kin
and
bas
lose
is c
nov
is e
pat
call
cor
seg
pat
avoids costly dependency parsing and generates high-quality patterns; (2) it identifies and groups synonymous meta patterns from multiple facets---their types, contexts, and extractions; and (3) it examines type distributions of entities in the instances extracted by each group of patterns, and looks for appropriate type levels to make discovered patterns precise. Experiments demonstrate that our proposed framework discovers high-quality typed textual patterns efficiently from different genres of massive corpora and facilitates information extraction.

This paper is a tutorial on Formal Concept Analysis (FCA) and its applications. FCA is an applied branch of Lattice Theory, a mathematical discipline which enables formalisation of concepts as basic units of human thinking and analysing data in the object-attribute form. Originated in early 80s, during the last three decades, it became a popular human-centred tool for knowledge representation and data analysis with numerous applications. Since the tutorial was specially prepared for RuSSIR 2014, the covered FCA topics include Information Retrieval with a focus on visualisation aspects, Machine Learning, Data Mining and Knowledge Discovery, Text Mining and several others.

	Var 1	Var 2	…	…	Var P
Doc 1					
Doc 2					
Doc 3					
…					
…					
…					
Doc D					

DTM, Topic Models, Doc2Vec, etc.

Vector Space Model vs. Matrix-based Model

- Matrix-based Model
  - ✓ A document is represented as a N by P matrix
    - N: the maximum number of words in a single document (ex: 512 for BERT)
    - P: the word embedding dimension (ex: 128)

The complicated, evolving landscape of cancer mutations poses a formidable challenge to identify cancer genes among the large lists of mutations typically generated in NGS experiments. The ability to prioritize these variants is therefore of paramount importance. To address this issue we developed OncoScore, a text-mining tool that ranks genes according to their association with cancer, based on available biomedical literature. Receiver operating characteristic curve and the area under the curve (AUC) metrics on manually curated datasets confirmed the excellent discriminating capability of OncoScore (OncoScore cut-off threshold = 21.09; AUC = 90.3%, 95% CI: 88.1-92.5%), indicating that OncoScore provides useful results in cases where an efficient prioritization of cancer-associated genes is needed.

	Var 1	Var 2	…	…	Var P
The					
complicated					
evolving					
landscape					
…					
is					
needed					
Padding	0	0	0	0	0
Padding	0	0	0	0	0
Padding	0	0	0	0	0

AGENDA

01 Document Classification: Overview

02 Naive Bayesian Classifier

03 k-Nearest Neighbor Classifier

Naive Bayesian Classifier

- Baye’s Rule (one of the most important rules in statistics)

$$P(C_i|x_1,x_2) = \frac{P(x_1,x_2|C_i)\cdot P(C_i)}{P(x_1,x_2)}$$

- Naive: Let’s assume that all variables are statistically independent to each other

$$= \frac{P(x_1|C_i)\cdot P(x_2|C_i)\cdot P(C_i)}{P(x_1,x_2)}$$

Naive Bayesian Classifier

- Maximum likelihood estimates
  - ✓ Simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N.Doc(C = c_j)}{Total\ number\ of\ documents}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Fraction of times word w_i appears among all words in documents of class c_j

- Limitation
  - ✓ What if we have seen no training document with the word "fantastic" and classified in in the class "positive"?

$$\hat{P}(\text{fantastic}|\text{positive}) = \frac{count(\text{fantastic}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

  - ✓ Zero probabilities cannot be conditioned away, no matter the other evidence!

Naive Bayesian Classifier

- Example: Classifying movie reviews into Positive/Negative class
  - ✓ Review 1: This movie was awesome! I really enjoyed it.
  - ✓ Review 2: This movie was boring and waste of time.
- Step 1: Estimate the conditional probabilities for each class

Words	P(Word/Positive)	P(Word/Negative)
This	0.1	0.1
Movie	0.1	0.1
Was	0.1	0.1
Awesome	0.4	0.01
I	0.2	0.2
Really	0.3	0.05
enjoyed	0.5	0.05
It	0.1	0.1
Boring	0.02	0.3
And	0.1	0.1
Waste	0.02	0.35
Of	0.02	0.02
Time	0.15	0.15

Naive Bayesian Classifier

- For the Review 1
  - ✓ Review 1: This movie was awesome! I really enjoyed it.

$$\prod_i P(Word_i|Pos) = 120 \times 10^{-8} \; > \; \prod_i P(Word_i|Neg) = 0.5 \times 10^{-8}$$

Words	P(Word/Positive)	P(Word/Negative)
This	0.1	0.1
Movie	0.1	0.1
Was	0.1	0.1
Awesome	0.4	0.01
I	0.2	0.2
Really	0.3	0.05
enjoyed	0.5	0.05
It	0.1	0.1
Boring	0.02	0.3
And	0.1	0.1
Waste	0.02	0.35
Of	0.02	0.02
Time	0.15	0.15

Naive Bayesian Classifier

- For the Review 1
  - ✓ Review 2: This movie was boring and waste of time.

$$\prod_i P(Word_i|Pos) = 0.012 \times 10^{-8} \; < \; \prod_i P(Word_i|Neg) = 3.15 \times 10^{-8}$$

Words	P(Word/Positive)	P(Word/Negative)
This	0.1	0.1
Movie	0.1	0.1
Was	0.1	0.1
Awesome	0.4	0.01
I	0.2	0.2
Really	0.3	0.05
enjoyed	0.5	0.05
It	0.1	0.1
Boring	0.02	0.3
And	0.1	0.1
Waste	0.02	0.35
Of	0.02	0.02
Time	0.15	0.15

Naive Bayesian Classifier

- Smoothing Techniques
  - ✓ Laplace (add-1) smoothing

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)} = \frac{count(w_i, c_j) + 1}{\sum_{w \in V} count(w, c_j) + |V|}$$

- Example

Model pos

0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model neg

0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

P(s|pos) > P(s|neg)

AGENDA

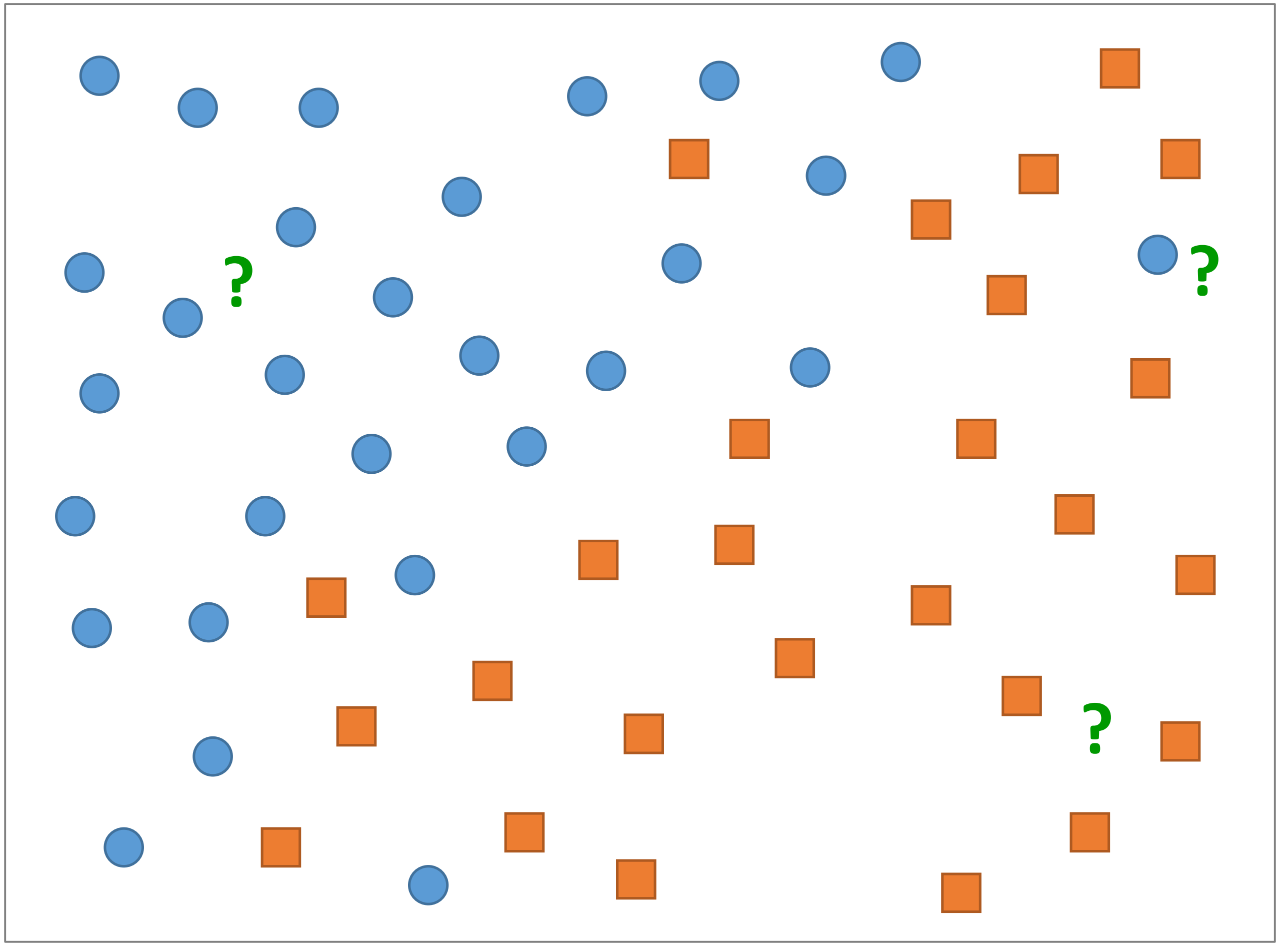
01 Document Classification: Overview

02 Naive Bayesian Classifier

03 k-Nearest Neighbor Classifier

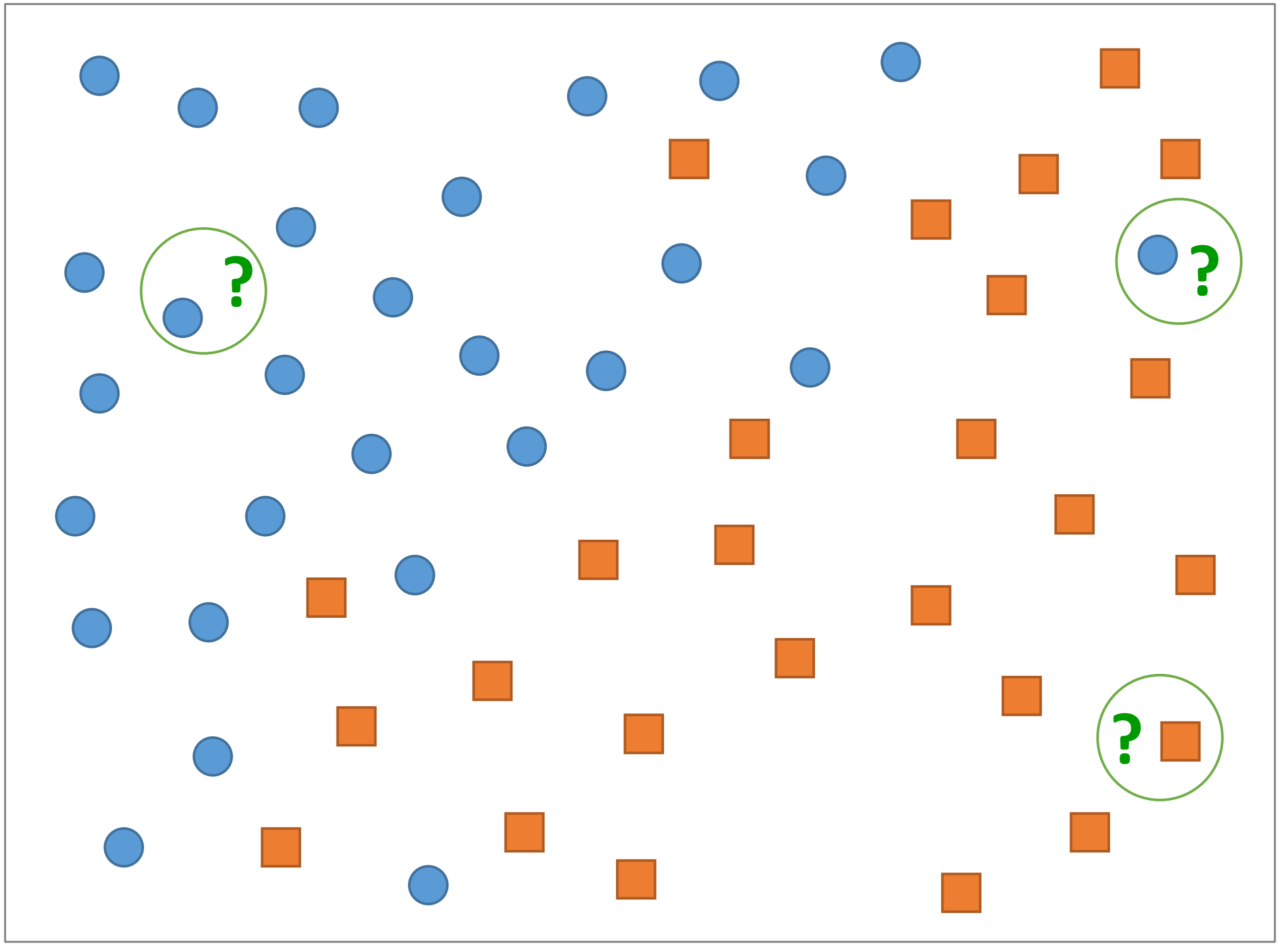
k-Nearest Neighbor Classification

- Which class does the question mark belong to?



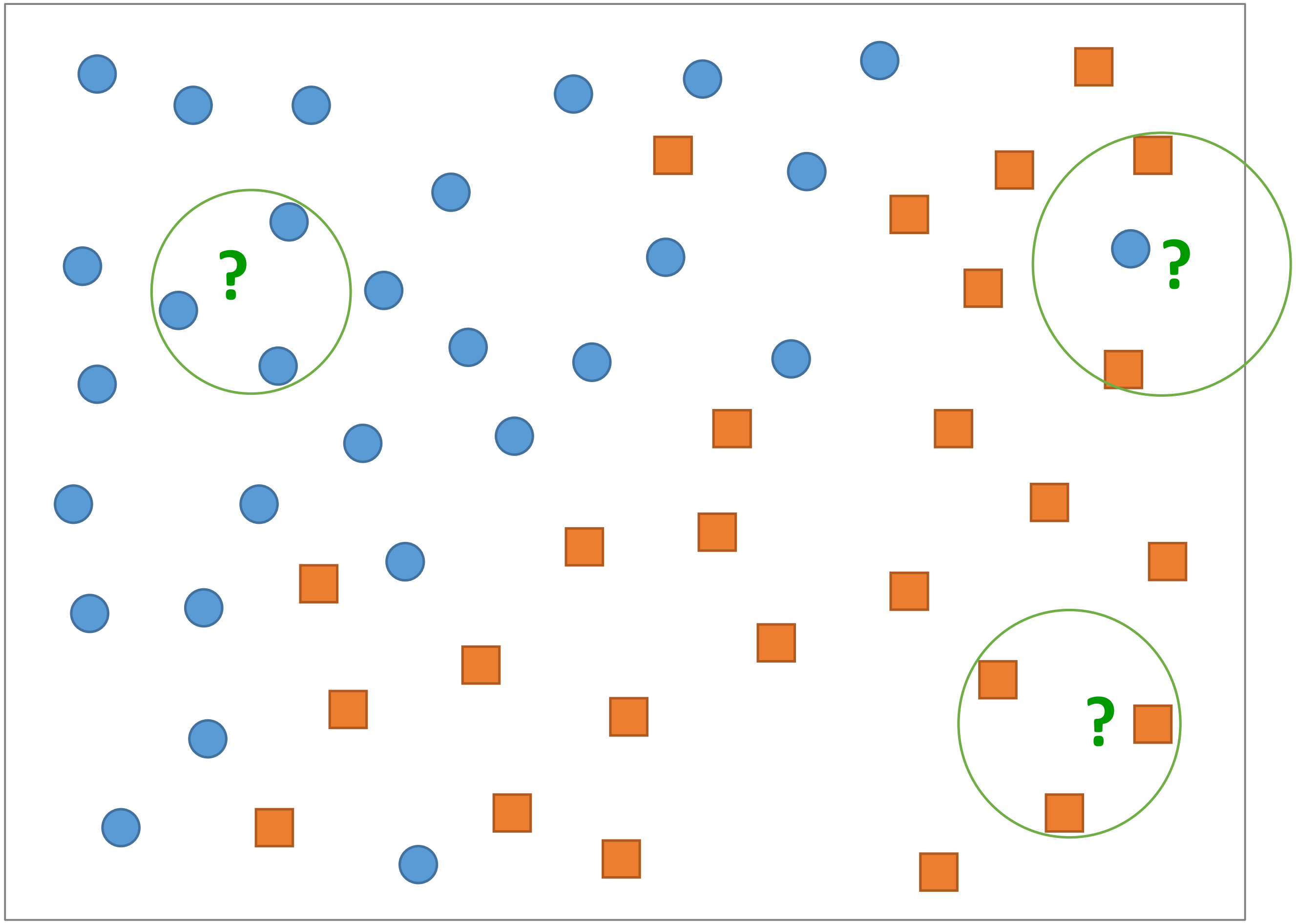
k-Nearest Neighbor Classification

- Which class does the question mark belong to?



k-Nearest Neighbor Classification

- Which class does the question mark belong to?



k-Nearest Neighbor Classification

- Motivation

類類相從　近墨者黑

“Birds of a feather flock together”

k-Nearest Neighbor Classification

k-NN Classification Process

- Step 1: Prepare the reference data
  - ✓ Define attributes
    - BoW representation or distributed representation
  - ✓ Collect sufficient number of records from each class

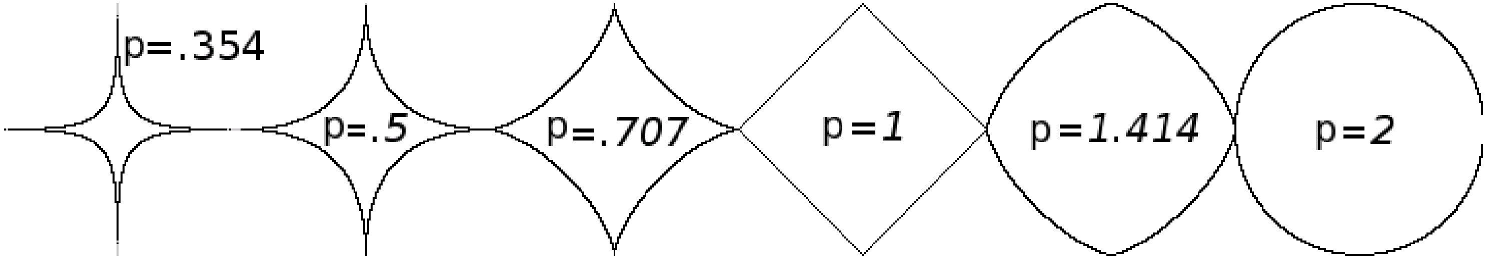
Doc.	Vision	Finance	…	Class
1	2.54	0.15	…	TPAMI
2	1.78	0	…	TPAMI
3	0.12	2.65	…	JoF
4	0.25	3.52	…	JoF
…	…	…	…	…
N	3.12	0.14	…	TPAMI

k-Nearest Neighbor Classification

k-NN Classification Process

- Step 2: Define the similarity measure
  - ✓ Similarity ∝ 1/distance
  - ✓ Minkovski distance with order p

$$\text{distance}\Big(P=(x_1,x_2,\ldots,x_n)\ ,\ Q(y_1,y_2,\ldots,y_n)\Big)=\Big(\sum_{i=1}^{n}|\,x_i-y_i\,|^p\Big)^{\frac{1}{p}}$$



  - ✓ p=2: Euclidean distance
  - ✓ p=1: Mahattan distance

k-Nearest Neighbor Classification

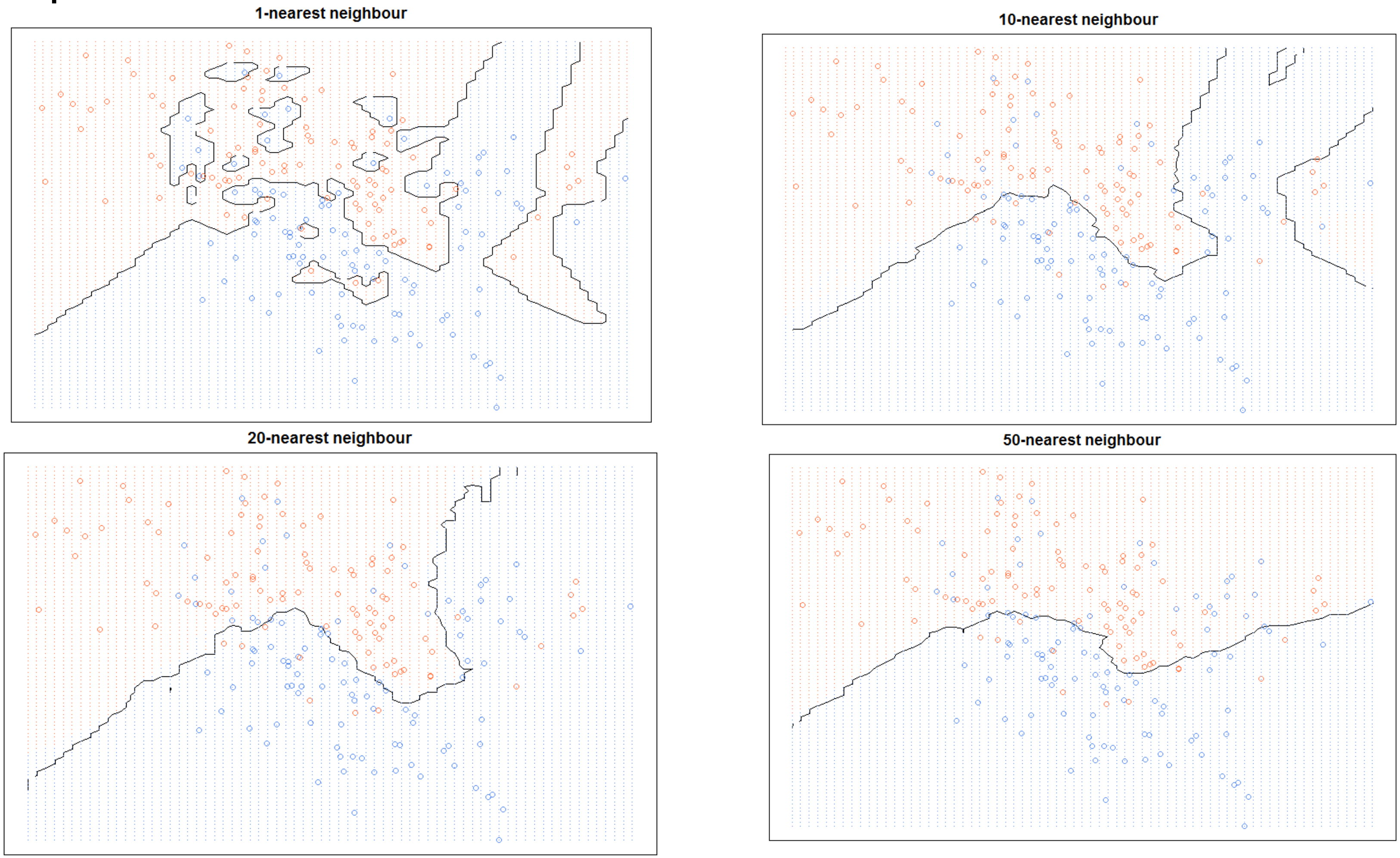
k-NN Classification Process

- Step 3: Initialize the set of candidate values for k
  - ✓ If k is too small, then the classification will be highly locally sensitive (over-fitting).
  - ✓ If k is too large, then it will lose the ability to capture the local structure (under-fitting).
  - ✓ A proper k should be chosen among a set of candidates.
  - ✓ Use the validation data.

k-Nearest Neighbor Classification

k-NN Classification Process

- Step 3: Initialize the set of candidate values for k



1-nearest neighbour

10-nearest neighbour

20-nearest neighbour

50-nearest neighbour

k-Nearest Neighbor Classification

k-NN Classification Process

- Step 4: Determine the combining rule

✓ Majority voting vs. Weighted voting

For a
new
data
X

Neighbor	Class	Distance	1/distance	Weight
N1	TPAMI	1	1.00	0.44
N2	JoF	2	0.50	0.22
N3	TPAMI	3	0.33	0.15
N4	JoF	4	0.25	0.11
N5	JoF	5	0.20	0.08

✓ Majority voting: P(X in TPAMI) = 2/5 = 0.4

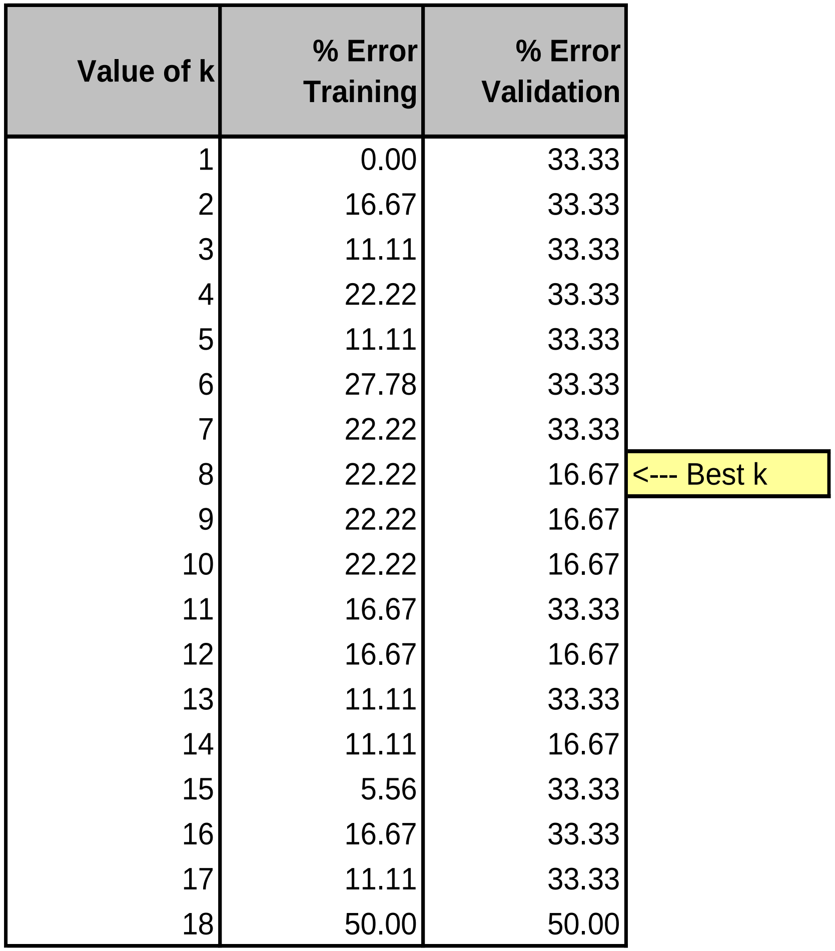
✓ Weighted voting: P(X in TPAMI) = 0.59

✓ If the cut-off is set to 0.5 X is classified as JoF by the majority voting while classified as TPAMI by the weighted voting

k-Nearest Neighbor Classification

k-NN Classification Process

- Step 5: Find the best k using the validation dataset



Value of k	% Error Training	% Error Validation	
1	0.00	33.33	
2	16.67	33.33	
3	11.11	33.33	
4	22.22	33.33	
5	11.11	33.33	
6	27.78	33.33	
7	22.22	33.33	
8	22.22	16.67	<--- Best k
9	22.22	16.67	
10	22.22	16.67	
11	16.67	33.33	
12	16.67	16.67	
13	11.11	33.33	
14	11.11	16.67	
15	5.56	33.33	
16	16.67	33.33	
17	11.11	33.33	
18	50.00	50.00	

k-Nearest Neighbor Classification

- k-NN Issue 1: Normalization
  - ✓ Normalization or scaling must be done before finding k-nearest neighbors
  - ✓ If not, variables with large measuring units are over-emphasized while variables with small measuring units are under-evaluated

[Before Normalization]

No.	Height	Weight	BFS	Gender
1	187	93	15	M
2	165	51	25	F
3	174	68	14	M
4	156	48	29	F
…	…	…	…	…
N	168	59	12	M
Avg.	165	65	20	-
Stdev.	15	10	5	-

[After Normalization]

No.	Height	Weight	BFS	Gender
1	1.47	2.80	-1.00	M
2	0.00	-1.40	1.00	F
3	0.60	0.30	-1.20	M
4	-0.60	-1.70	1.80	F
…	…	…	…	…
N	0.20	-0.60	-1.60	M

k-Nearest Neighbor Classification

- k-NN Issue 2: Cut-off
  - ✓ Consider the prior probability of each class
  - ✓ Assume that N(M) = 100, N(F) = 400

For a new data

X

Neighbor	Class
N1	M
N2	F
N3	M
N4	F
N5	F

Majority voting

P(X=M)=0.4

  - ✓ If the cut-off is set to 0.5 (assuming equal class distribution), then X is classified as F.
  - ✓ If the cut-off is set to 0.2 (proportion of M among the people), then X is classified as M.

ANY
questions?