Informatics 225
Computer Science 221

Information Retrieval

Lecture 27

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

Introduction to Text classification

Information Retrieval

Categorization/Classification

- Given:
  - A representation of a document *d*
    - Issue: how to represent text documents.
    - Usually some type of high-dimensional space – bag of words
  - A fixed set of classes:

    $C = \{c_1, c_2, \ldots, c_J\}$

- Determine:
  - The category of *d*: $\gamma(d) \in C$, where $\gamma(d)$ is a classification function

  - We want to build classification functions (“classifiers”).

Supervised Learning

- Given:
  - A document *d*
  - A fixed set of classes:

    $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set *D* of documents each with a label in *C*

Supervised Learning

- Given:
  - A document *d*
  - A fixed set of classes:

    $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set *D* of documents each with a label in *C*

- Determine:
  - A learning method or algorithm which will enable us to learn a classifier *γ*
  - For a new test document *dt*, we assign it the class

$$\gamma(dt) \in C$$

Supervised Learning classification methods

- A few methods that are used for text
  - Naive Bayes (simple & common)
  - k-Nearest Neighbors (simple & powerful)
  - Support-vector machines (generally more powerful)
  - Decision trees → random forests → gradient-boosted decision trees (e.g., xgboost)
  - Neural networks
  - … plus many other methods
  - *Supervised learning: need hand-classified training data*
- Many research and commercial systems use a mix of methods
  - Ensemble methods

Features

- Classifiers can use any sort of feature
  - URL, email address, punctuation, capitalization, dictionaries, network features
- In the simplest possible view: bag of words of documents
  - We use **only** word features
  - We use **all** of the words in the text (not a subset)

The bag of words representation

$\gamma($

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.

$)=c$

The bag of words representation

$\gamma($

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

$)=c$

Feature Selection: Why?

- Text collections have a large number of features
  - 10,000 – 1,000,000 unique words … and more
- Selection may make a particular classifier feasible
  - Most classifiers can't deal with 1,000,000 features *(the curse of dimensionality)*
- Reduces training time
  - Training time for some methods is quadratic or worse in number of features
- Makes runtime models smaller and faster
- Can improve generalization (performance)
  - Eliminates noisy features
  - Avoids overfitting

Feature Selection: Frequency of words?

- The simplest feature selection method:
  - Just use the commonest terms for a certain frequency threshold
  - No particular foundation
  - But it make sense why this works
    - They' re the words that can be well-estimated and are most often available as evidence
  - In practice, this is ~90% as good as better methods *(but mileage may vary as this ~90% is averaged over all possible classes)*

Remember: Vector Space Representation

- Each document is a vector,
  one component for each term (= word).
- Normally normalize vectors to unit length.
- High-dimensional vector space:
  - **Terms are axes**
  - 10,000+ dimensions, or even 100,000+
  - Docs are vectors in this space
- How can we do classification in this space?

Classification Using Vector Spaces

- In vector space classification, training set corresponds to a labeled set of points (equivalently, vectors)

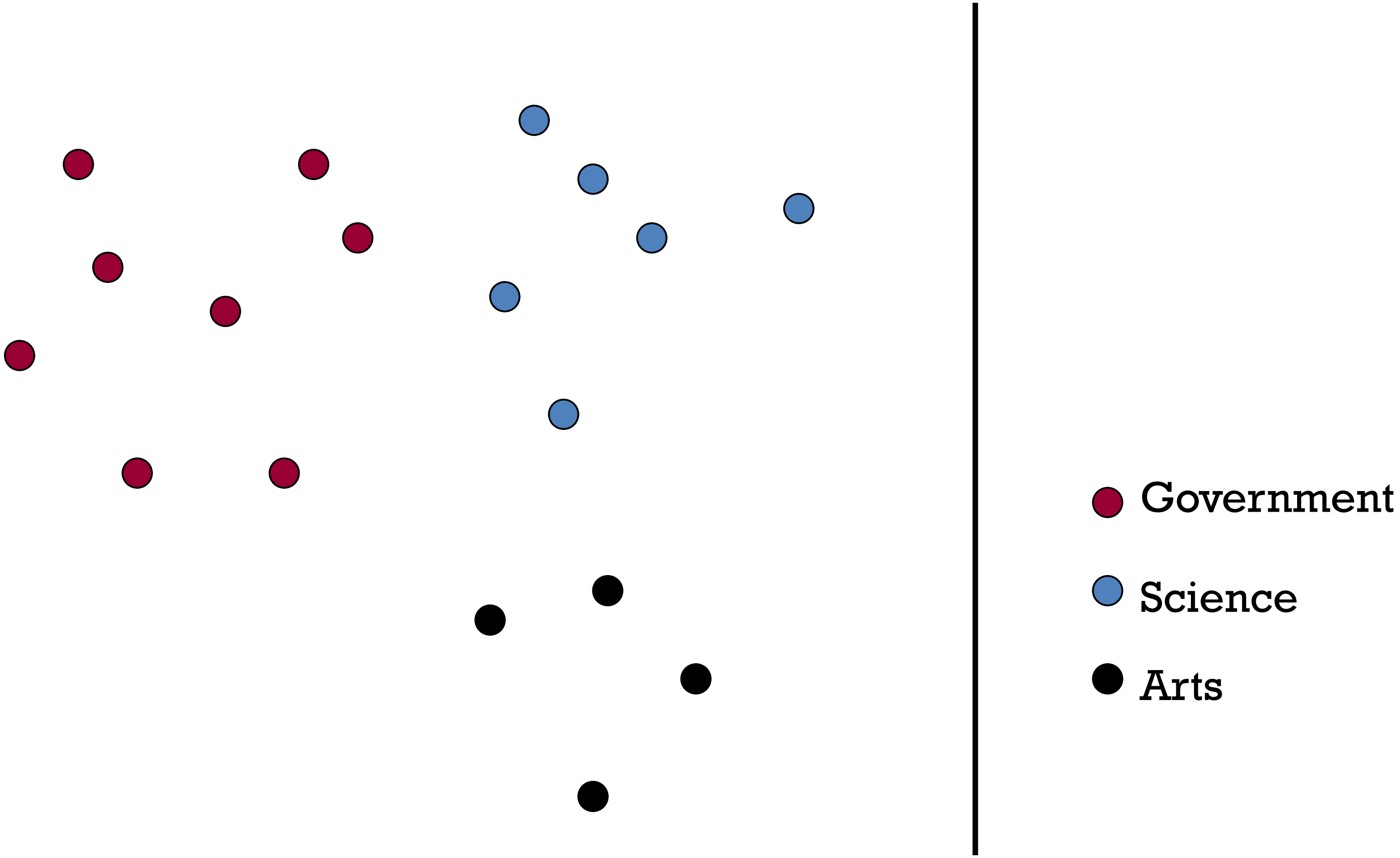
- Premise 1: Documents in the same class form a contiguous region of space

- Premise 2: Documents from different classes don't overlap (much)

- Learning a classifier:

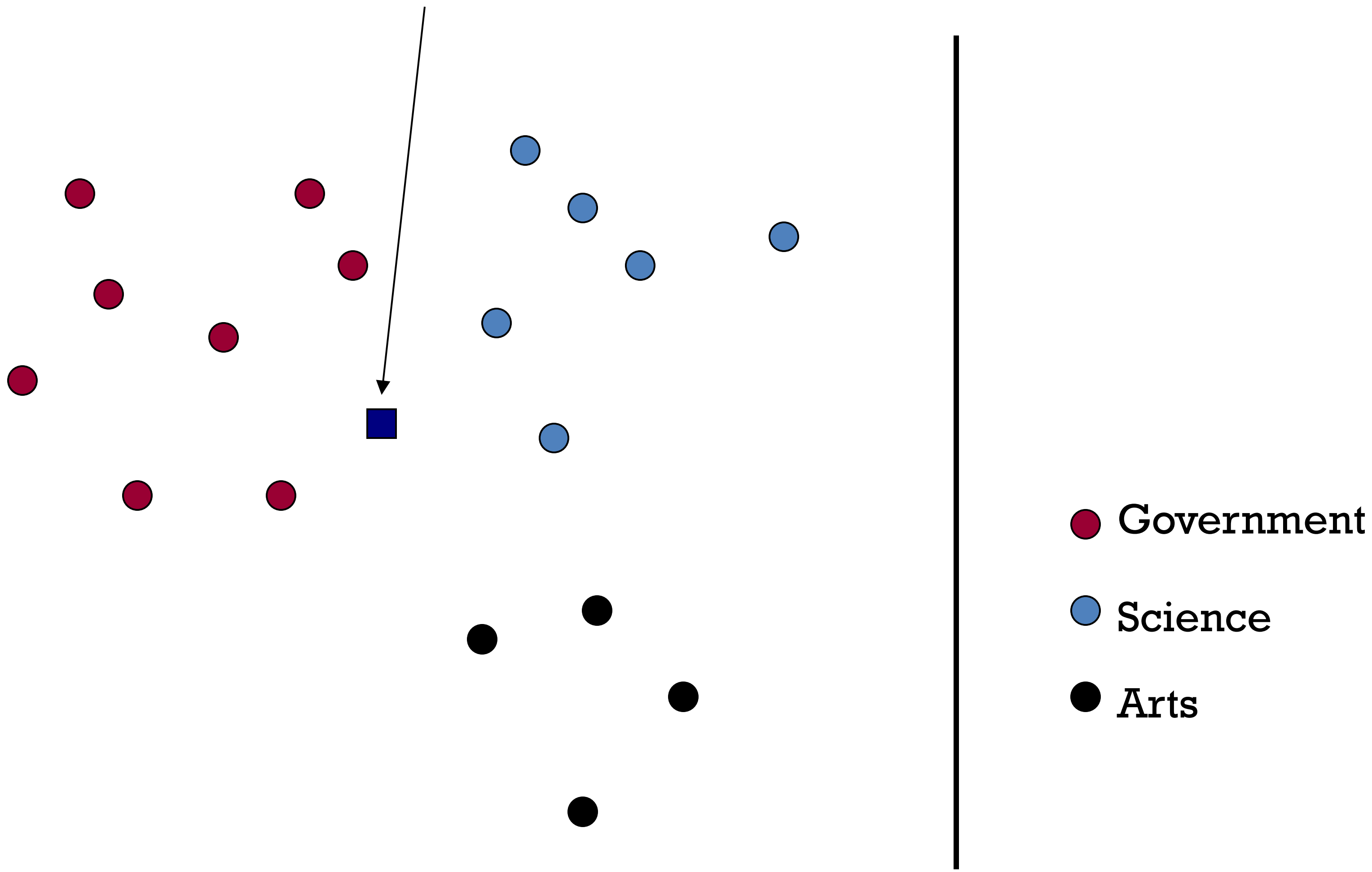
  build surfaces to delineate classes in the space

Documents in a Vector Space

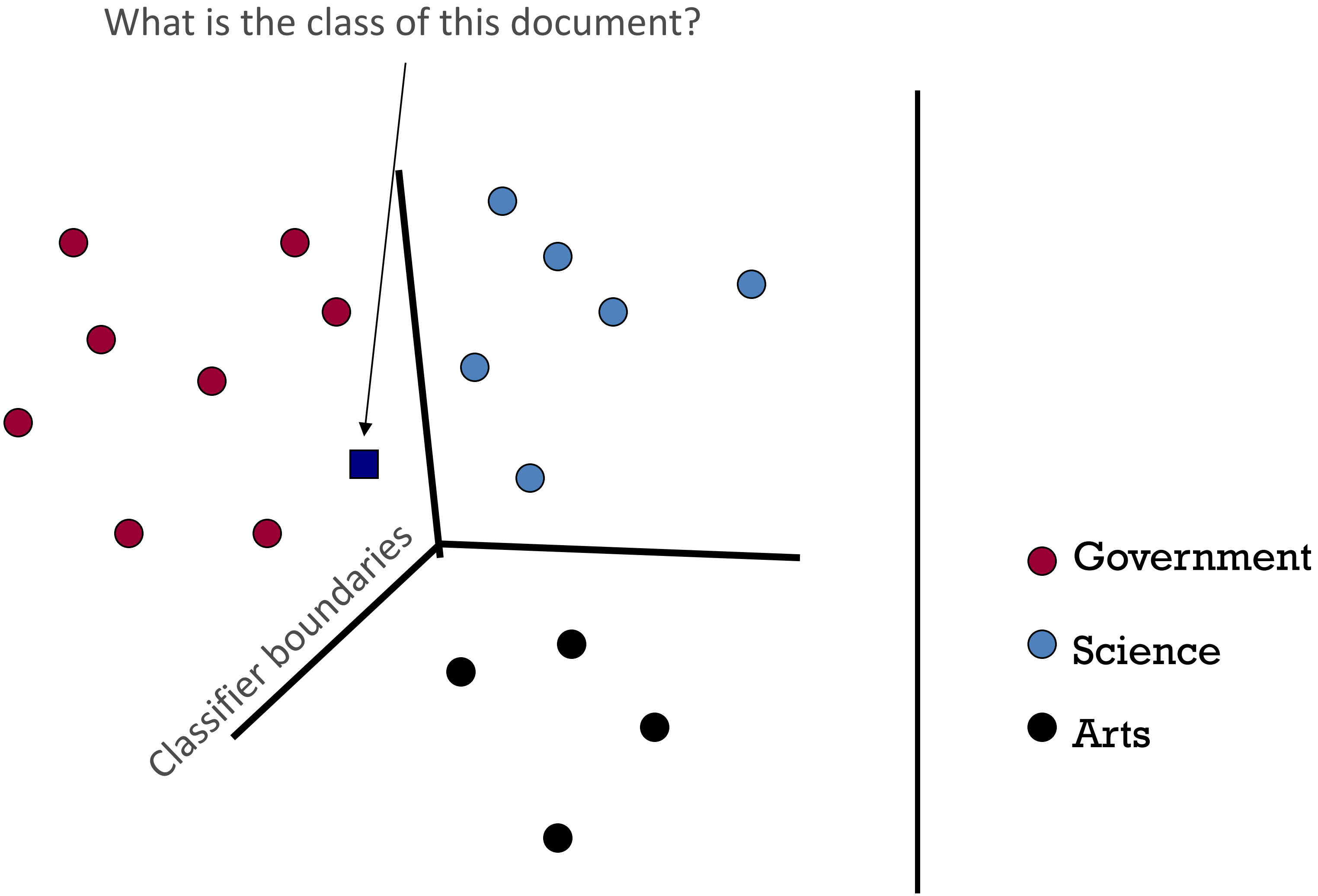


Documents in a Vector Space

What is the class of this document?



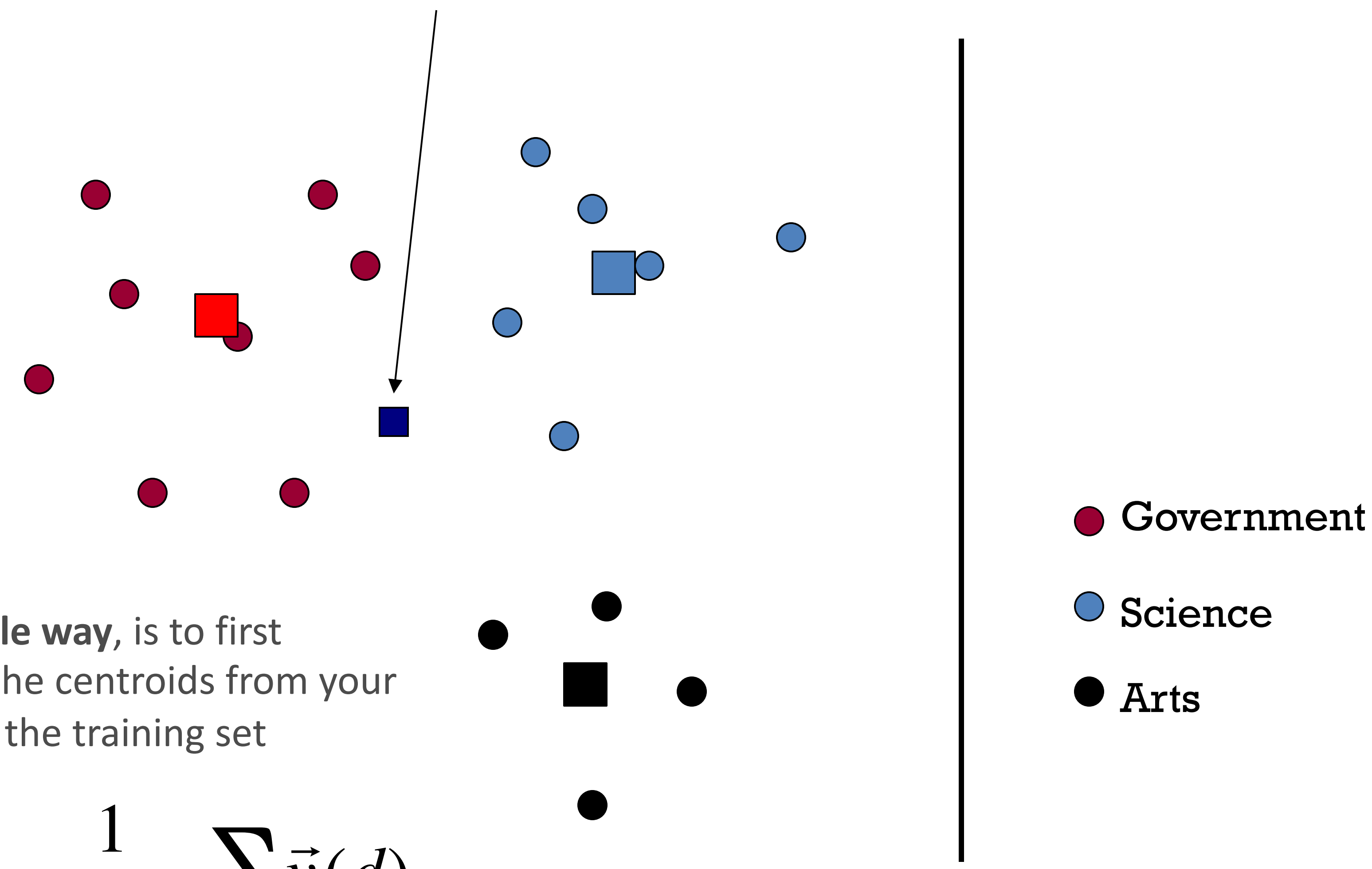
Documents in a Vector Space



But how can you define these boundaries?

Documents in a Vector Space : simple classifier

What is the class of this document?

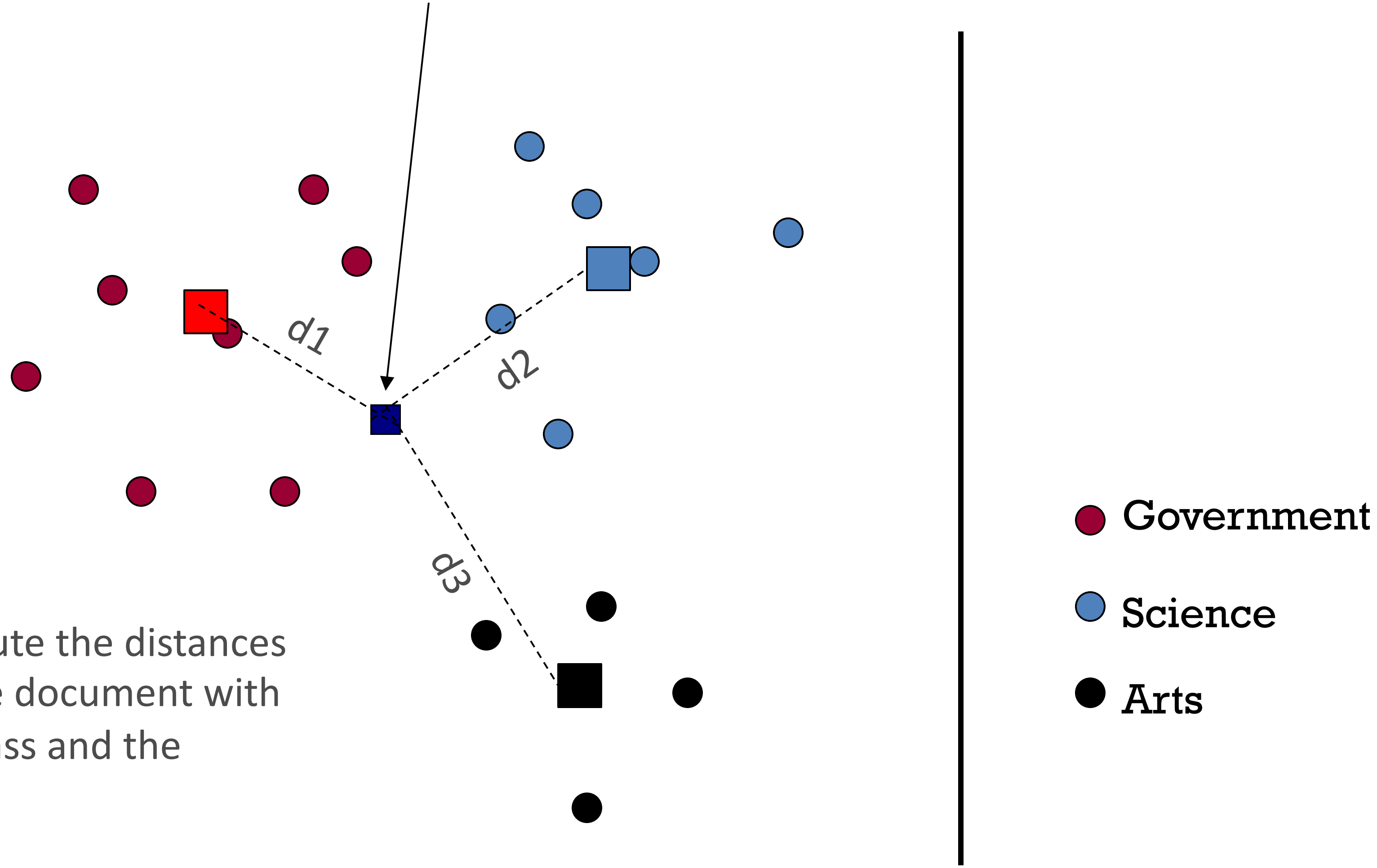


One **simple way**, is to first discover the centroids from your classes in the training set

$$\vec{\mu}(c) = \frac{1}{|D_c|}\sum_{d \in D_c} \vec{v}(d)$$

Documents in a Vector Space : simple classifier

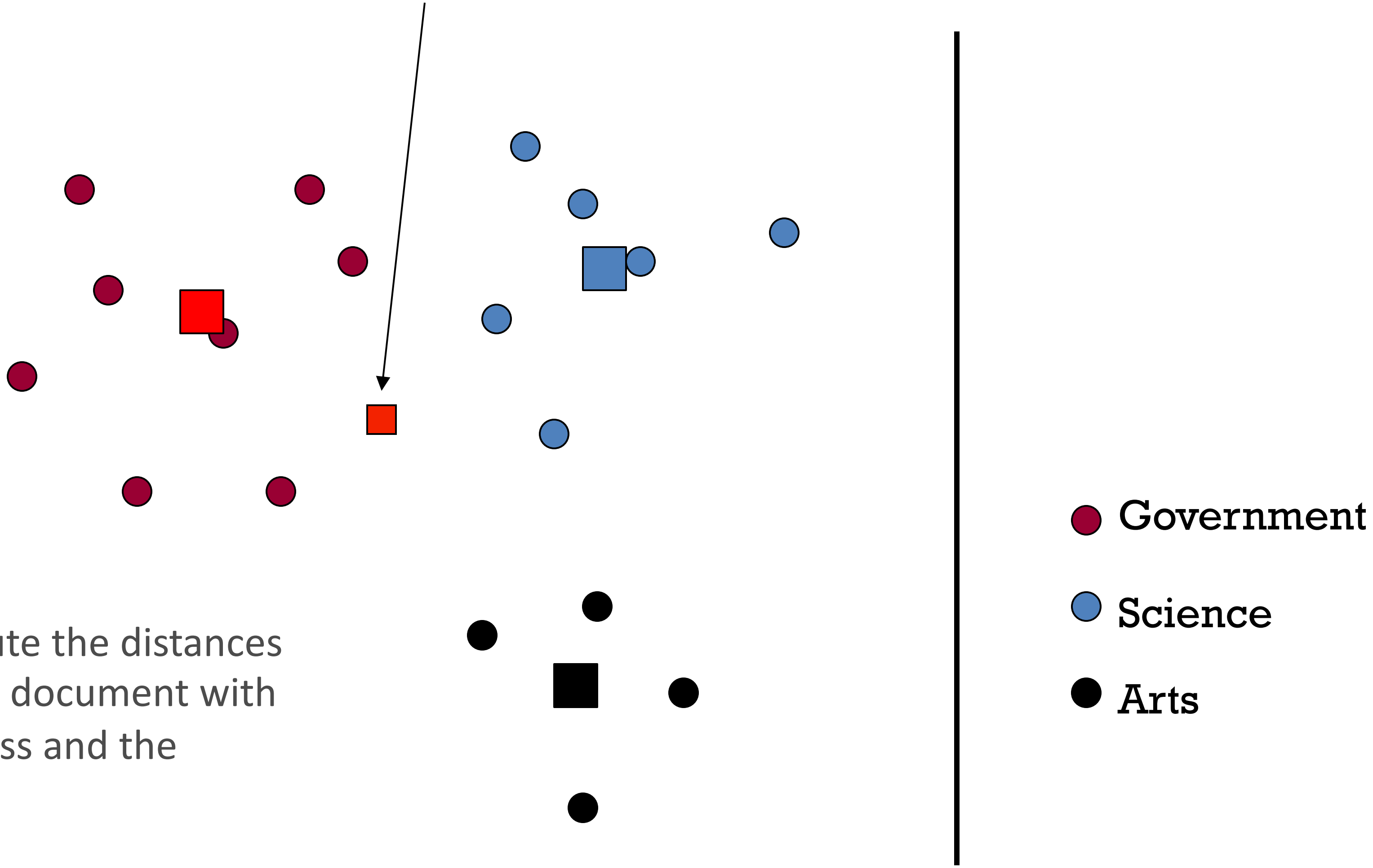
What is the class of this document?



Then, compute the distances between the document with unknown class and the centroids.

Documents in a Vector Space : simple classifier

What is the class of this document? : Government



Then, compute the distances between the document with unknown class and the centroids.

Finally, assign the class as the same class as the nearest centroid.

The Information Retrieval course

Problem Space of this course

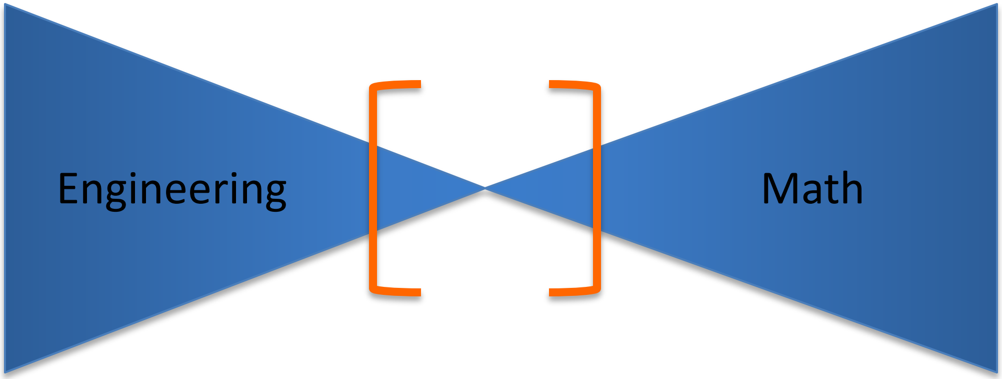
- “Big Data”

- How to
  - collect it
  - index it
  - search it for relevant information

Industry segment of this course

- Search engines
  - Google
  - MS Bing
  - Duck Duck Go
  - nameless others
- Web and social information retrieval = big $$$

Technical content of this course



Goals

- For you to know the fundamentals of information retrieval
- For you to have at your disposal some techniques to find information spread among 10^n documents
- For you to practice Web crawling and **for to build your own simple search engine from scratch**

This course answers three primary questions

- What is information retrieval?
- How to build some of the most important blocks of information retrieval systems?
- How to perform information retrieval in the web?

All together

- Search engines history
- Search & advertising on the Web
- Web corpus
- Characteristics of the Web
- Characteristics of Web search Users
- Search Engine Optimization
- Web crawling
- Index construction
- Map Reduce
- Boolean retrieval
- Scored retrieval
- TF-IDF and corpus-wide statistics
- Text processing
- Link analysis (PageRank, HITS)
- Search Engine Evaluation: Precision, Recall, MAP and NDCG@R

All together

- Search engines history
- Search & advertising on the Web
- Web corpus
- Characteristics of the Web
- Characteristics of Web search Users
- Search Engine Optimization
- Web crawling
- Index construction
- Map Reduce
- Boolean retrieval
- Scored retrieval
- TF-IDF and corpus-wide statistics
- Text processing
- Link analysis (PageRank, HITS)
- Search Engine Evaluation: Precision, Recall, MAP and NDCG@R

You built a complete search engine from the ground up...

Few professionals had this experience!

Big Data jobs

- Plenty…
- Not just traditional search
  - *making sense of data*
- Google it!
- Open your own company
  - Several fields for niche search and recommendation systems
  - Image and video search : nobody is doing it right yet…
  - Deep web search : nobody is doing it right yet…

Where to go from here

- Graphs
- Data mining
- Statistical and Machine learning
- Invest time to learn the fundaments!
  - *Take as many proof-based algorithms courses and proof-based math courses as you can (but avoid taking them at the same time…);*
  - *Always try to do things from scratch (e.g. your own database/webserver/search-engine/etc.);*
  - *Profit from your University years to build your "superpowers"!*

in the summer vacations, watch Karate Kid (the original)
and pay attention to "wax on wax off..."

Where to go from here

- Graphs
- Data mining
- Statistical and Machine learning
- Invest time to learn the fundaments!
  - *Take as many proof-based algorithms courses and proof-based math courses as you can (but avoid taking them at the same time…);*
  - *Always try to do things from scratch (e.g. your own database/webserver/search-engine/etc.);*
  - *Profit from your University years to build your "superpowers"!*

in the summer vacations, watch Karate Kid (the original)
and pay attention to "wax on wax off..."

Whenever you need any help, now and forever, just drop by my office (DBH 5058)!

Thank you for your patience!

and see you around!

Information Retrieval