Text processing

••

USC Viterbi
School of Engineering

Standing Queries

- The path from IR to text classification:
  - You have an information need to monitor, say:
    - Unrest in the Niger delta region
  - You want to rerun an appropriate query periodically to find new news items on this topic
  - You will be sent new documents that are found
    - I.e., it's not ranking but classification (relevant vs. not relevant)
- Such queries are called standing queries
  - Long used by "information professionals"
  - A modern mass instantiation is Google Alerts
- Standing queries are (hand-written) text classifiers

• •

From: Google Alerts
Subject: **Google Alert - stanford -neuro-linguistic nlp OR "Natural Language Processing" OR parser OR tagger OR ner OR "named entity" OR segmenter OR classifier OR dependencies OR "core nlp" OR corenlp OR phrasal**
Date: May 7, 2012 8:54:53 PM PDT
To: Christopher Manning

Web — 3 new results for **stanford -neuro-linguistic nlp OR "Natural Language Processing" OR parser OR tagger OR ner OR "named entity" OR segmenter OR classifier OR dependencies OR "core nlp" OR corenlp OR phrasal**

Twitter / **Stanford NLP** Group: @Robertoross If you only n ...
@Robertoross If you only need tokenization, java -mx2m edu.**stanford**.**nlp**. process.PTBTokenizer file.txt runs in 2MB on a whole file for me.... 9:41 PM Apr 28th ...
twitter.com/stanfordnlp/status/196459102770171905

[Java] LexicalizedParser lp = LexicalizedParser.loadModel("edu ...
loadModel("edu/**stanford**/**nlp**/models/lexparser/englishPCFG.ser.gz");. String[] sent = { "This", "is", "an", "easy", "sentence", "." };. Tree parse = lp.apply(Arrays.
pastebin.com/az14R9nd

More Problems with Statistical **NLP** || kuro5hin.org
Tags: nlp, ai, coursera, **stanford**, **nlp**-class, cky, nltk, reinventing the wheel, ... Programming Assignment 6 for **Stanford's** **nlp**-class is to implement a CKY parser .
www.kuro5hin.org/story/2012/5/5/11011/68221

Tip: Use quotes ("like this") around a set of words in your query to match them exactly. Learn more.

Delete this alert.
Create another alert.
Manage your alerts.

••

Spam filtering
Another text classification task

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
===

••

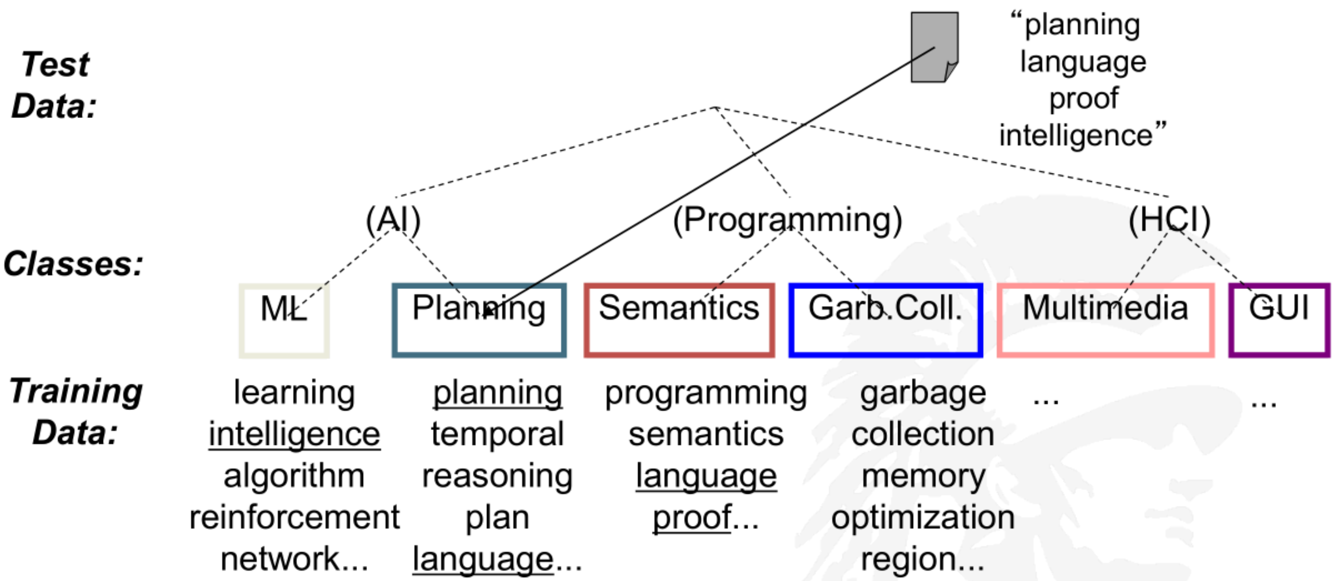
Categorization/Classification

- **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* by generating a classification function, say $\gamma(d)$**
  - **We want to build classification functions ("classifiers").**

USC Viterbi School of Engineering

Document Classification



••

Classification Methods (1)

- Manual classification
  - Used by the original Yahoo! Directory
  - Looksmart, about.com, ODP, PubMed
  - Accurate when job is done by experts
  - Consistent when the problem size and team is small
  - Difficult and expensive to scale
    - Means we need automatic classification methods for big problems

••

Classification Methods (2)

- **Hand-coded rule-based classifiers**
  - **One technique used by news agencies, intelligence agencies, etc.**
  - **Widely deployed in government and enterprises**
  - **Vendors provide "IDE" for writing such rules**

••

USC Viterbi School of Engineering

Classification Methods (2)

- **Hand-coded rule-based classifiers**
  - **Commercial systems have complex query languages**
  - **Accuracy is can be high if a rule has been carefully refined over time by a subject expert**
  - **Building and maintaining these rules is expensive**

••

Classification Methods (3): 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 test document d, we assign it the class**

$\gamma(d) \in C$

••

USC Viterbi School of Engineering

Classification Methods (3)

- Supervised learning
  - Naive Bayes (simple, common)
  - k-Nearest Neighbors (simple, powerful)
  - Support-vector machines (newer, generally more powerful)
  - … plus many other methods
  - No free lunch: requires hand-classified training data
  - But data can be built up (and refined) by amateurs
- Many commercial systems use a mixture of methods

••

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
...	...

$)=c$

••

Features

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

••

Feature Selection: Why?

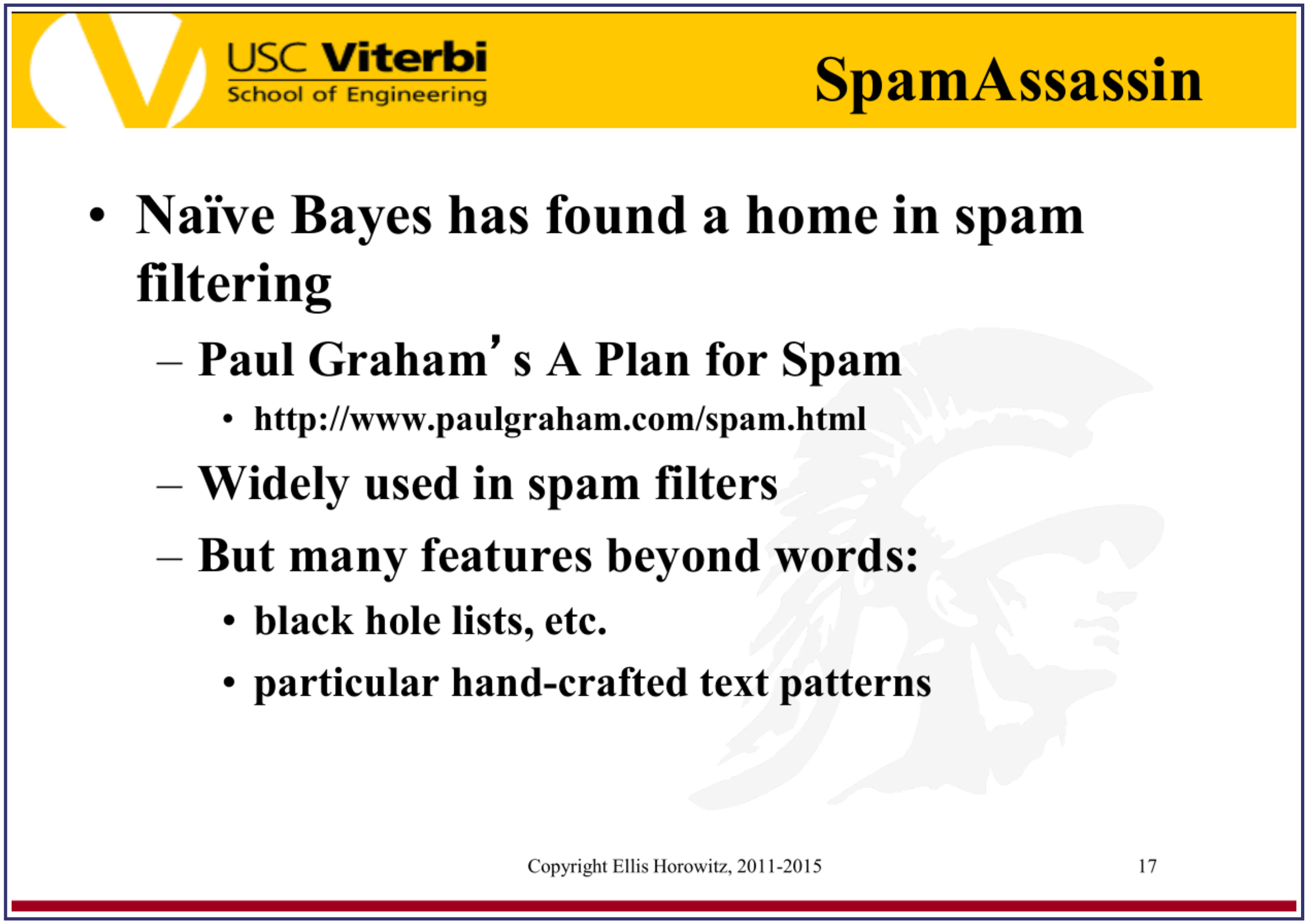
- Text collections have a large number of features
  - 10,000 – 1,000,000 unique words … and more
- Selection may make a particular classifier feasible
  - Some classifiers can't deal with 1,000,000 features
- Reduces training time
  - Training time for some methods is quadratic or worse in the number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

••

USC Viterbi School of Engineering

Feature Selection: Frequency

- The simplest feature selection method:
  - Just use the most common terms
  - No particular foundation
  - But it make sense why this works
    - They are the words that can be well-estimated and are most often available as evidence
  - In practice, this is often 90% as good as better methods
  - Smarter feature selection – future lecture



Here is Paul's page, on his spam filtering expts.

Here is SpamAssassin, and this is an old page with results of tests performed.

••

Naive Bayes is Not So Naive

- Very fast learning and testing (basically just count words)
- Low storage requirements
- Very good in domains with many equally important features
- More robust to irrelevant features than many learning methods

  Irrelevant features cancel each other without affecting results

••

Evaluating Categorization

- **Measures: precision, recall, F1, classification accuracy**
- **Classification accuracy: *r/n* where *n* is the total number of test docs and *r* is the number of test docs correctly classified**

••

Recall: 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?

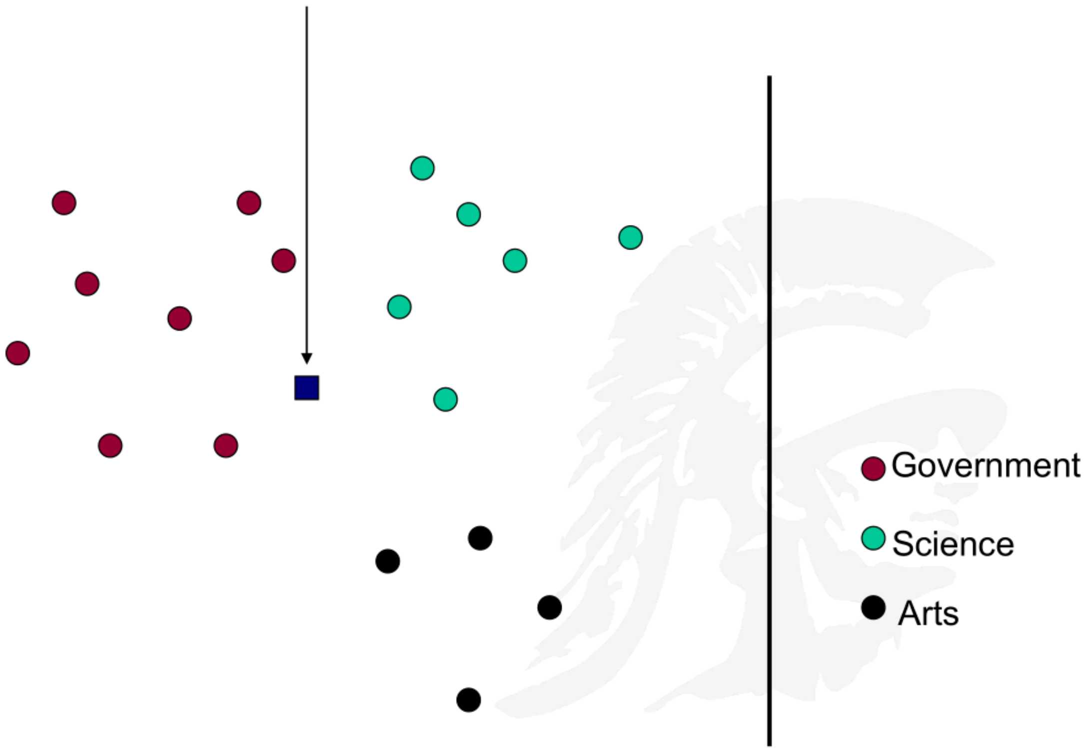
••

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

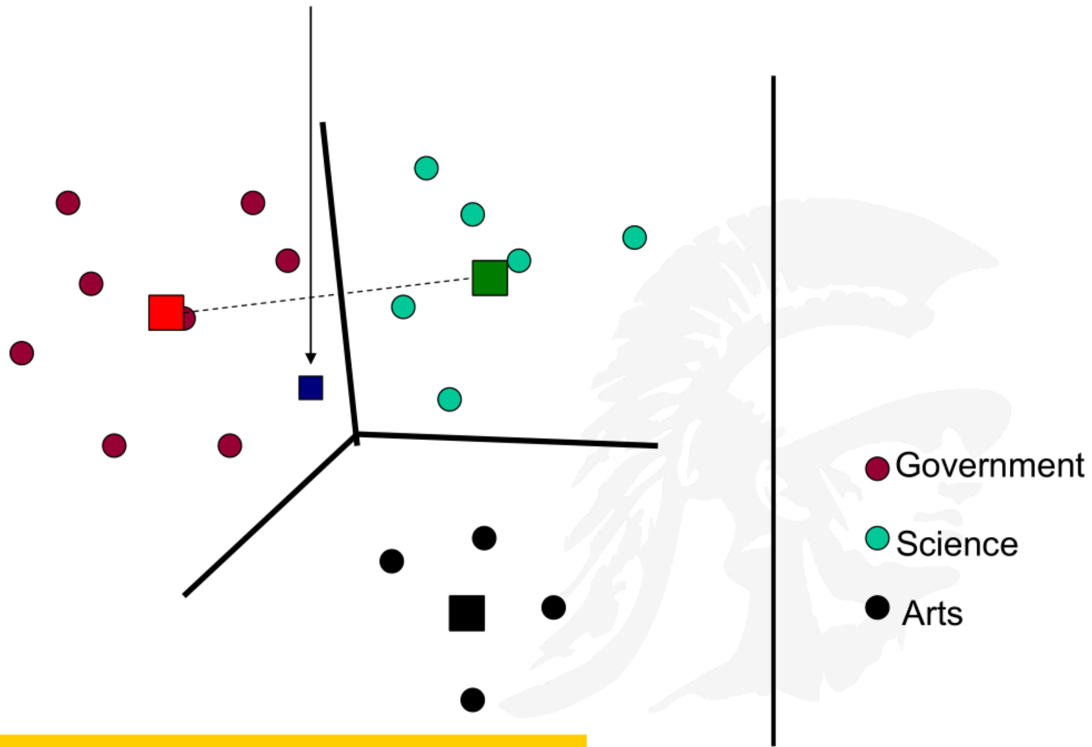
USC Viterbi School of Engineering

Test Document of what class?



USC Viterbi School of Engineering

Test Document = Government



Our focus: how to find good separators

••

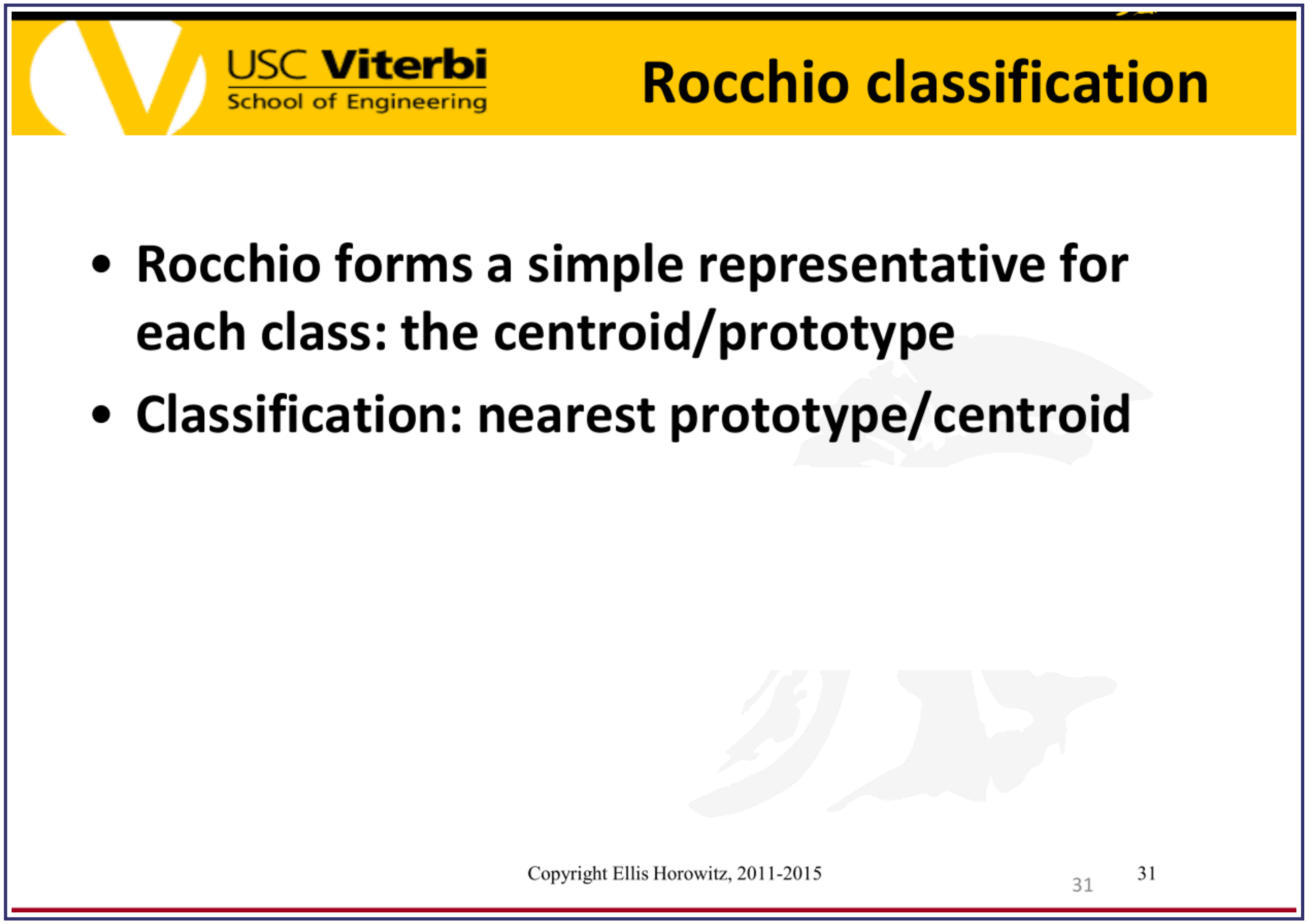
USC Viterbi School of Engineering

Definition of centroid

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

- **Where *D_c* is the set of all documents that belong to class *c* and *v*(*d*) is the vector space representation of *d*.**

- ***Note that centroid will in general* not *be a unit vector even when the inputs are unit vectors.***



In Rocchio classification, centroids (one for each term group) are used to specify regions; lines/planes/hyperplanes between centroids produce convex **Voronoi** regions. The new/incoming term's closest centroid is used to classify the term.

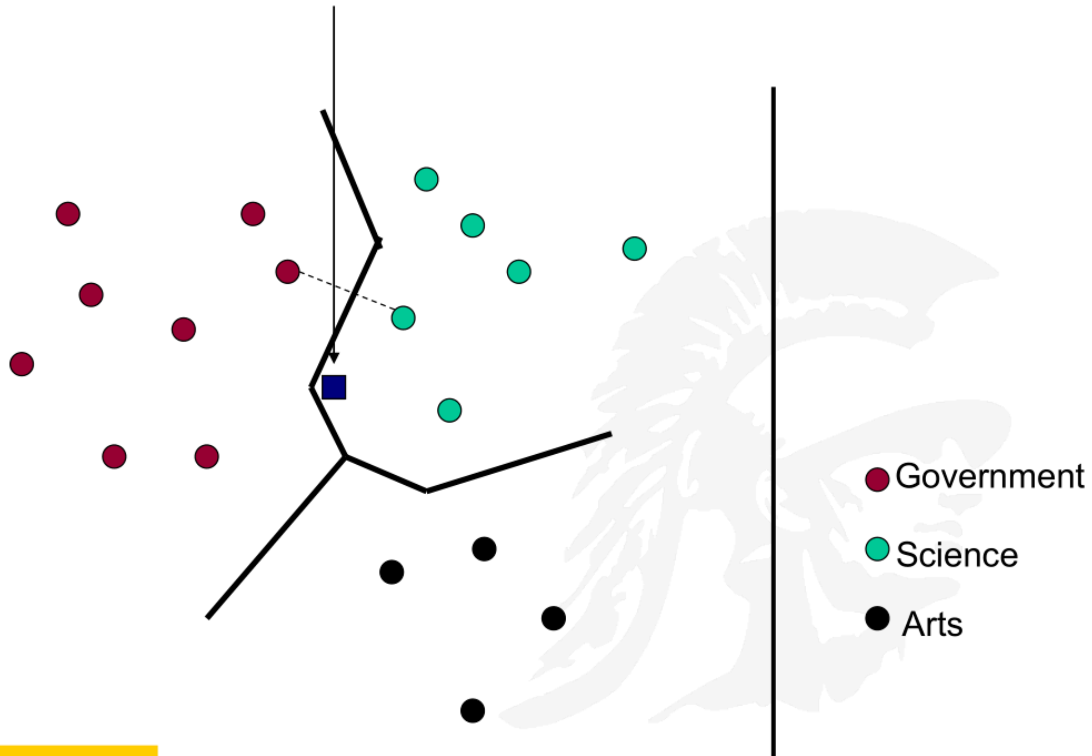
••

k Nearest Neighbor Classification

- kNN = *k* Nearest Neighbor

- To classify a document *d*:
- Define *k*-neighborhood as the *k* nearest neighbors of *d*
- Pick the majority class label in the *k*-neighborhood
- For larger *k* can roughly estimate P(c|*d*) as #(c)/k

Test Document = Science



Voronoi diagram

••

Nearest-Neighbor Learning

- Learning: just store the labeled training examples *D*
- Testing instance *x (under 1NN)*:
  - Compute similarity between *x* and all examples in *D*.
  - Assign *x* the category of the most similar example in *D*.
- Does not compute anything beyond storing the examples
- Also called:
  - Case-based learning
  - Memory-based learning
  - Lazy learning
- Rationale of kNN: contiguity hypothesis

••

USC Viterbi School of Engineering

k Nearest Neighbor

- Using only the closest example (1NN) subject to errors due to:
  - A single atypical example.
  - Noise (i.e., an error) in the category label of a single training example.
- More robust: find the *k* examples and return the majority category of these *k*
- *k* is typically odd to avoid ties; 3 and 5 are most common

••

USC Viterbi School of Engineering

kNN: Discussion

- No feature selection necessary
- No training necessary
- Scales well with large number of classes
  – Don't need to train *n* classifiers for *n* classes
- Classes can influence each other
  – Small changes to one class can have ripple effect
- Done naively, very expensive at test time
- In most cases it's more accurate than NB or Rocchio