Features and classification

CSC401/2511 – Natural Language Computing –Spring 2019
Lecture 4 – Frank Rudzicz and Chloé Pou-Prom
University of Toronto

Lecture 4 overview

- Today:
- **Feature extraction** from text.
  - How to pick the right features?
  - Grammatical ‘parts-of-speech’.
    - (which don’t require spoken language)
- **Classification** overview

- Some slides *may* be based on content from Bob Carpenter, Dan Klein, Roger Levy, Josh Goodman, Dan Jurafsky, and Christopher Manning.

Features

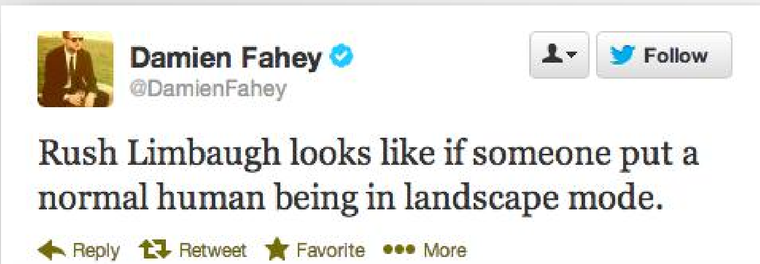
- **Feature**: *n.* A measurable **variable** that is (rather, *should be*) **distinctive** of something we want to model.

- We usually choose features that are useful to **identify** something, i.e., to do **classification**.
  - E.g., an emotional, whiny **tone** is likely to indicate that its source is not legal, or scientific, or political.

- We often need **several** features to adequately model something – *but not too many*!

Feature vectors

- **Values** for **several** features of an **observation** can be put into a single **vector**.



# proper nouns	# 1st person pronouns	# commas
2	0	0
5	0	0
0	1	1

UNIVERSITY OF TORONTO

Feature vectors

- Features should be useful in **discriminating** between categories.

Table 3: Features to be computed for each text

- Counts:
  - First person pronouns
  - Second person pronouns
  - Third person pronouns
  - Coordinating conjunctions
  - Past-tense verbs
  - Future-tense verbs
  - Commas
  - Colons and semi-colons
  - Dashes
  - Parentheses
  - Ellipses
  - Common nouns
  - Proper nouns
  - Adverbs
  - *wh*-words
  - Modern slang acroynms
  - Words all in upper case (at least 2 letters long)
- Average length of sentences (in tokens)
- Average length of tokens, excluding punctuation tokens (in characters)
- Number of sentences

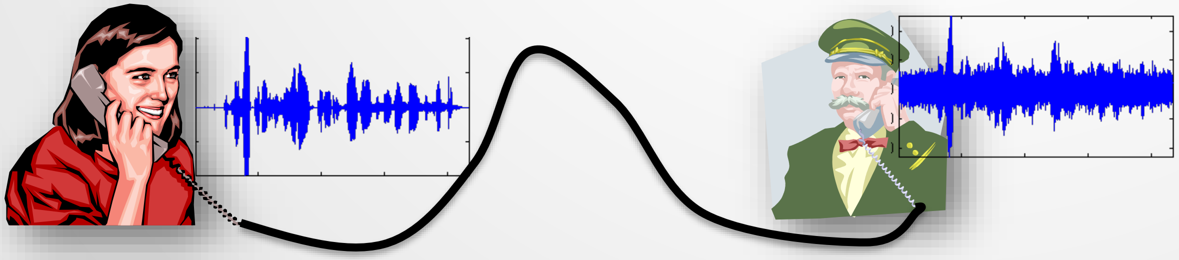
Higher values → this person is referring to themselves (to their opinion, too?)

Higher values → looking forward to (or dreading) some future event?

Lower values → this tweet is more formal. Perhaps not overly sentimental?

Quick comment on noise

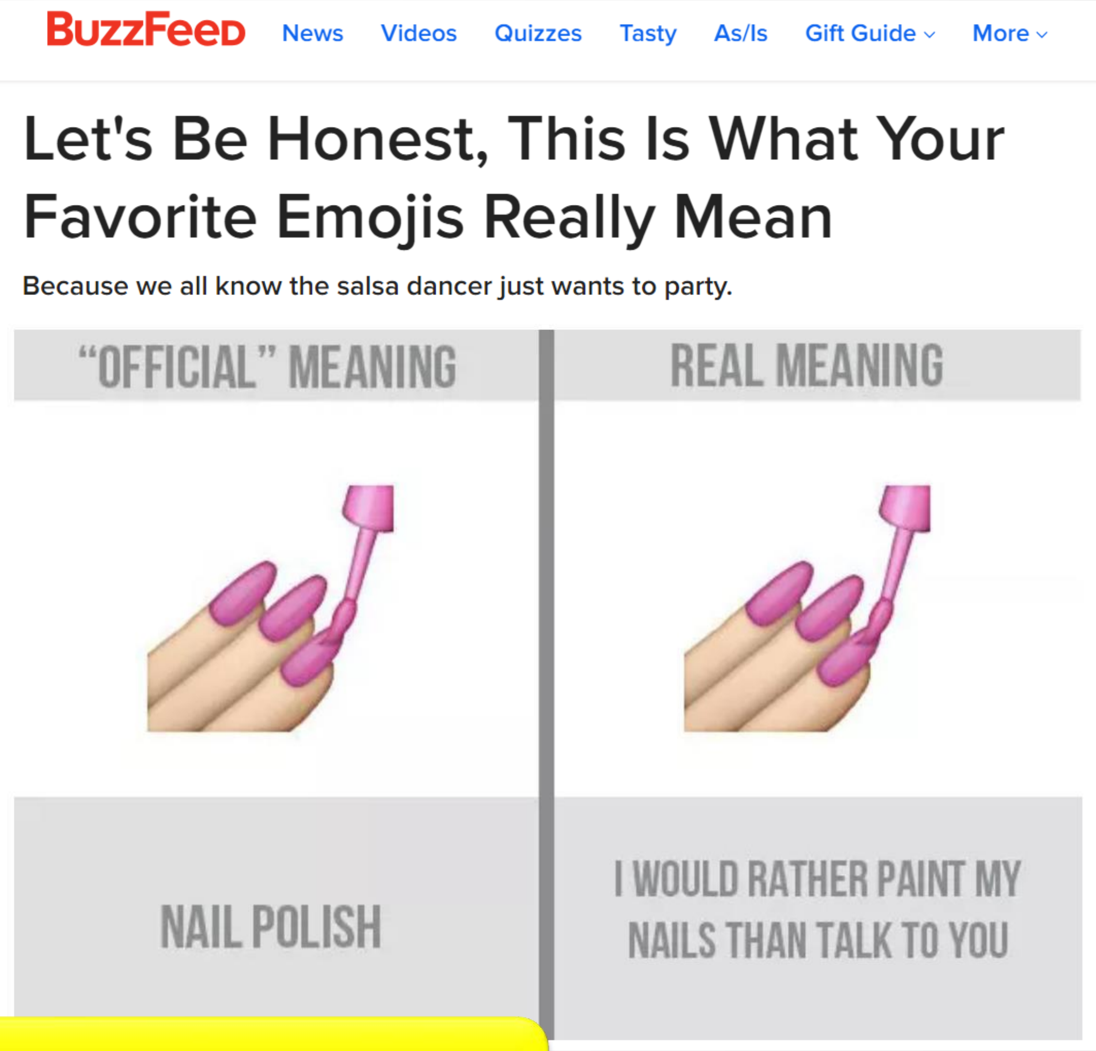
- **Noise** is generally any **artifact** in your received '**signal**' that **obfuscates** (hides) the features you want.
  - E.g., in **acoustics**, it can be a loud buzzing sound that washes out someone's voice.
  - E.g., in **tweets**, it can be text that invalidates your counts.
    - E.g., The semi-colon in "… `octopus ;)`" is part of an **emoticon**; will it confuse our **classifier** if we count it as punctuation?



UNIVERSITY OF TORONTO

Quick comment on noise

- E.g., in **tweets**, it can be text that invalidates your counts.
  - The semi-colon in "… `octopus ;)`" is part of an **emoticon**; will it confuse our **classifier** if we count it as punctuation?



Note: you don't have to deal with emoticons in A1.

Pre-processing

- **Pre-processing** involves **preparing** your data to make feature extraction easier or more valid.
  - E.g., **punctuation** likes to press up against words. The sequence “ *example,* ” should be counted as **two** tokens – not one.
    - We separate the punctuation, as in “ *example ,* ”.

- ***<u>There is no perfect pre-processor</u>***.
  Mutually exclusive approaches can often **both** be justified.
  - E.g., Is *Newfoundland-Labrador* **one** word type or **two**?
    Each answer has a unique implication for splitting the dash.
  - Often, **noise-reduction** removes *some* information.
  - Being **consistent** is important.

Different features for different tasks

- **Alzheimer's disease** involves atrophy in the brain.
  - Excessive **pauses** (acoustic disfluencies),
  - Excessive **word type repetition**, and
  - Simplistic or **short** sentences.
    - '**function words**' like *the* and *an* are often **dropped**.

- To **diagnose** Alzheimer's disease, one might measure:
  - **Proportion** of utterance spent in **silence**.
  - **Entropy** of **word type** usage.
  - **Number** of word **tokens** in a sentence.
    - **Number** of **prepositions** and **determiners** (explained shortly).

UNIVERSITY OF TORONTO

Features in Sentiment Analysis

- Sentiment analysis can involve detecting:
  - **Stress** or **frustration** in a conversation.
  - **Interest**, **confusion**, or **preferences**. Useful to marketers.
    - e.g., *'omg got pickle rick 4xmas wanted #botw fml'*
  - **Lies**. e.g., *'Let's watch Netflix and chill.'*

What does this mean?

- Complicating factors include **sarcasm**, **implicitness**, and a **subtle** spectrum from **negative** to **positive** opinions.

- **Useful features** for sentiment analyzers include:
  - Trigrams.
  - First-person pronouns.

Pronouns? Prepositions? Determiners?

Parts of Speech

Parts of speech (PoS)

- Linguists like to group words according to their **structural function** in building sentences.
  - This is similar to grouping Lego by their shapes.

- **Part-of-speech**: $\boxed{\textit{n.}}$ lexical category or morphological class.

Nouns collectively constitute a part of speech (called *Noun*)

UNIVERSITY OF TORONTO

Example parts of speech

Part of Speech	Description	Examples
Noun	is usually a **person**, **place**, **event**, or **entity**.	*chair, pacing, monkey, breath.*
Verb	is usually an **action** or **predicate**.	*run, debate, explicate.*
Adjective	modifies a **noun** to further describe it.	*orange, obscene, disgusting.*
Adverb	modifies a **verb** to further describe it.	*lovingly, horrifyingly, often*

Example parts of speech

Part of Speech	Description	Examples
Preposition	Often specifies aspects of **space**, **time**, or **means**.	*around, over, under, after, before, with*
Pronoun	Substitutes for nouns; referent typically understood in context.	*I, we, they*
Determiner	logically **quantify** words, usually nouns.	*the, an, both, either*
Conjunction	**combines** words or phrases.	*and, or, although*

Other parts of speech

- **Particles**: *up, down, on, off*
  - e.g., *throw her coat* ***off***
    ≡ throw ***off*** *her coat*
- **Auxiliaries**: *can, may, should, is, have*
- **Numerals**: *one, \$19.99,* 6.02×10^{23}
- **Punctuation**: *), (, :, ,, .*
- **Symbols**: *+, %, &*
- **Interjection**: *uh, hmmm, duh, aaah*
- …

UNIVERSITY OF TORONTO

Contentful parts-of-speech

- Some PoS convey more **meaning.**
  - Usually nouns, verbs, adjectives, adverbs.
  - **Contentful** PoS usually contain more words.
    - e.g., there are more **nouns** than **prepositions**.

  - **New** contentful words are continually **added**

    e.g., an *app, to google, to misunderestimate*.
  - **Archaic** contentful words go **extinct**.

    e.g., *fumificate, v., (1721-1792),*
    frenigerent, adj., (1656-1681),
    melanochalcographer, n., (c. 1697).

Functional parts-of-speech

- Some PoS are ‘**glue**’ that holds others together.
  - E.g., prepositions, determiners, conjunctions.
  - **Functional** PoS usually cover a **small** and **fixed** number of word types (i.e., a ‘**closed class**’).

  - Their **semantics** depend on the contentful words with which they’re used.
    - E.g., *I’m* ***on*** *time* vs. *I’m* ***on*** *a boat*

UNIVERSITY OF TORONTO

Grammatical features

- There are several **grammatical features** that can be associated with words:
  - **Case**
  - **Person**
  - **Number**
  - **Gender**
- These features can **restrict** other words in a sentence.

(Aside) Grammatical features – case

- **Case**: *n.* the **grammatical** form of a **noun** or **pronoun**.

- E.g.,

  nominative: the **subject** of a verb (e.g., "***We*** *remember*")

  accusative: the **direct object** of a verb (e.g., "*You remember* ***us***"))

  dative: the **indirect object** of a verb (e.g. "*I gave your* ***mom*** *the book*")

  genitive: indicates **possession** (e.g., "*your* ***mom****'s book")*

  …

(Aside) Grammatical features – person

- **Person**: *n.* typically refers to a participant in an event, especially with **pronouns** in a conversation.

- E.g.,

first:	The speaker/author. Can be either inclusive ("***we***") or exclusive of hearer/reader ("***I***").
second:	The hearer/reader, exclusive of speaker ("***you***")
third:	Everyone else ("***they***")

UNIVERSITY OF TORONTO

(Aside) Grammatical features – number

- **Number**: *n.* Broad numerical distinction.

- E.g.,

  singular: Exactly one ("one ***cow***")
  plural: More than one ("two ***cows***")
  dual: Exactly two (e.g., **ان -** in Arabic).
  paucal: Not *too* many (e.g., in Hopi).
  collective: Countable (e.g., Welsh "***moch***" for 'pigs' as opposed to "***mochyn***" for vast 'pigginess').
  …

UNIVERSITY OF TORONTO

(Aside) Grammatical features – gender

- **gender**: *n.* typically partitions **nouns** into classes associated with biological gender. **Not** typical in English.
  - Gender alters neighbouring words **regardless** of speaker/hearer.

- E.g.,

feminine:	Typically **pleasant** things (not always). (e.g., *la* ***France***, *eine* ***Brücke***, *une* ***poubelle***).
masculine:	Typically **ugly** or **rugged** things (not always). (e.g., *le* ***Québec***, *un* ***pont***).
neuter:	Everything else.

(***Brücke***: German bridge; ***pont***: French bridge; ***poubelle***: French garbage)

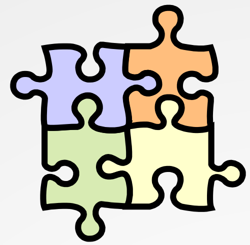
Other features of nouns

- **Proper noun**: **named** things (e.g., "*they've killed **Bill**!*")
- **Common noun**: **unnamed** things (e.g., "*they've killed the **bill**!*")
- **Mass noun**: **divisible** and **uncountable** (e.g., "***butter***" split in two gives two piles of butter – not two '*butters*')
- **Count noun**: **indivisible** and **countable**. (e.g., a "***pig***" split in two does not give two pigs)

(Aside) Some features of prepositions

- *By*
  - **Alongside**: *a cottage by the lake*
  - **Agentive**: *Chlamydia was given to Mary by John*
- *For*
  - **Benefactive**: *I have a message for your mom*
  - **Purpose**: *have a friend (over) for dinner*
- *With*
  - **Sociative**: *watch a film with a friend*
  - **Instrumental**: *hit a nail with a hammer*

Agreement



- Parts-of-speech ***should*** match (i.e., **agree**) in certain ways.

- **Articles** ‘have’ to **agree** with the **number** of their **noun**
  - e.g., “*these pretzels are making me thirsty*”
  - e.g., “*a winters are coming*”

- **Verbs** ‘have’ to **agree** (at least) with their **subject** (in English)
  - e.g., “*the dogs eats the gravy*” **no number** agreement
  - e.g., “*Yesterday, all my troubles seem so far away*”
    bad tense – should be past tense *seemed*
  - e.g., “*Can you handle me the way I are?*”

UNIVERSITY OF TORONTO

Tagging

PoS tagging

- **Tagging**: *v.g.* the process of **assigning** a **part-of-speech** to each word in a sequence.

- E.g., using the '**Penn treebank**' tag set (see appendix):

Word	The	nurse	put	the	angry	koala	to	sleep
Tag	DT	NN	VBD	DT	JJ	NN	IN	NN

Ambiguities in parts-of-speech

- Words can belong to many parts-of-speech.
  - E.g., *back:*
    - *The back/JJ door* (adjective)
    - *On its back/NN* (noun)
    - *Win the voters back/RB* (adverb)
    - *Promise to back/VB you in a fight* (verb)
- We want to decide the **appropriate** tag given a particular sequence of tokens.

Why is tagging useful?

- First step towards practical purposes.
  - E.g.,
    - **Speech synthesis**: how to pronounce text
      - *I'm **conTENT**/JJ* vs. *the **CONtent**/NN*
      - *I **obJECT**/VBP* vs. *the **OBJect**/NN*
      - *I **lead**/VBP ("l iy d")* vs. *it's **lead**/NN ("l eh d")*
    - **Information extraction**:
      - Quickly finding names and relations.
    - **Machine translation**:
      - Identifying grammatical 'chunks' is useful

Tagging as classification

- We have access to a **sequence of observations** and are expected to decide on the best assignment of a **hidden variable**, i.e., the PoS

Hidden variable				NN		
				VB		
		VBN		JJ		**NN**
	PRP	**VBD**	**TO**	RB	**DT**	VB
Observation	**she**	**promised**	**to**	**back**	**the**	**bill**

Rule-based tagging

1. **Start** with a **dictionary**
2. **Assign all** possible tags to words from the dictionary.
3. **Write rules** (‘by hand’) to selectively **remove** tags

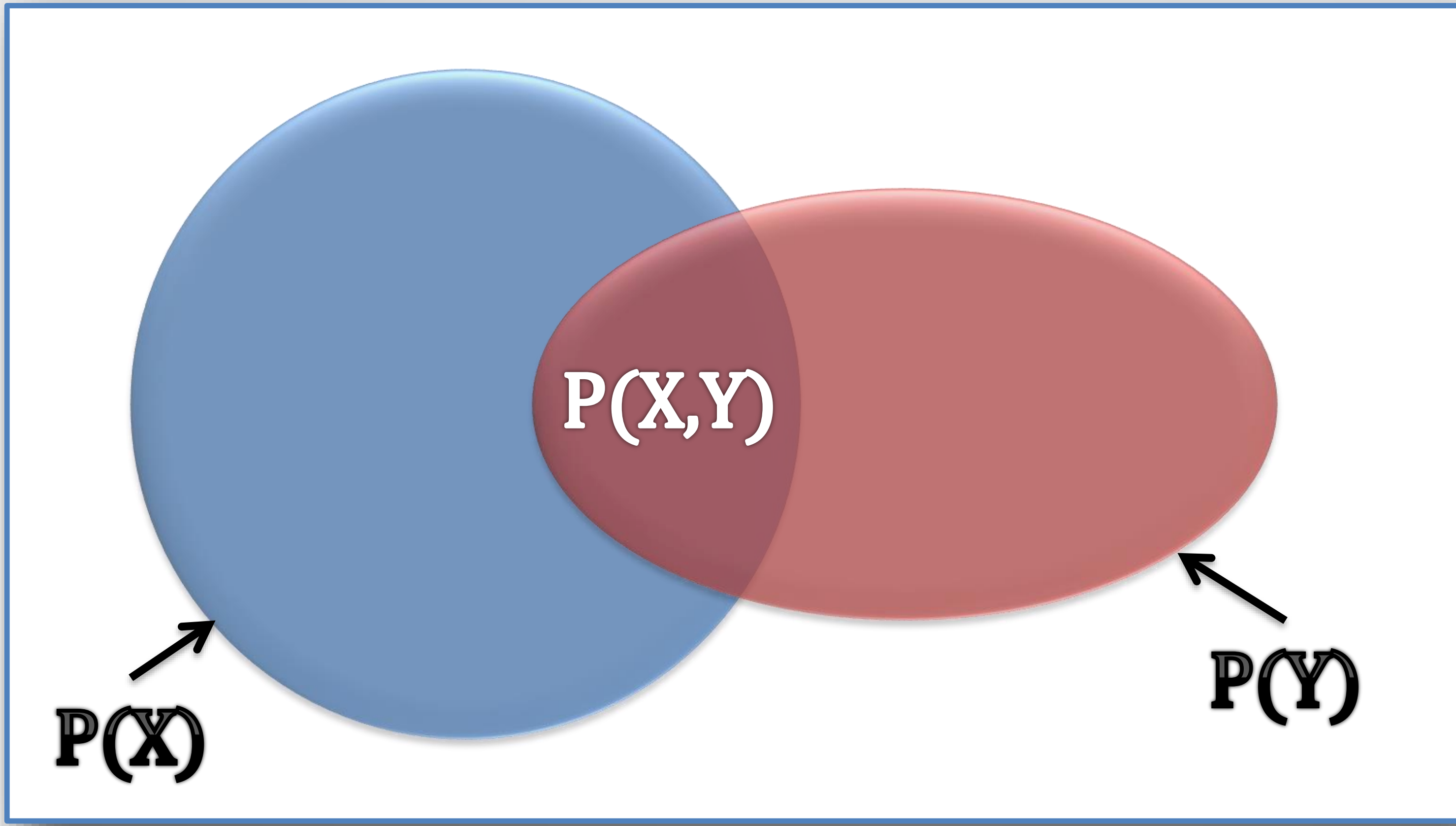
Rule-based tagging example

- Eliminate VBN (past participle) if VBD (past tense) is an option when (VBN|VBD) follows "<s> PRP (personal pronoun)"
- These kinds of rules become **unwieldy** and force **determinism** where there may not be any.

			NN		
			VB		
	~~VBN~~		JJ		NN
PRP	VBD	TO	RB	DT	VB
she	**promised**	**to**	**back**	**the**	**bill**

Can we use statistics instead?

Reminder: Bayes' Rule



$$P(X,Y) = P(X)P(Y|X)$$
$$P(X,Y) = P(Y)P(X|Y)$$

$$\boxed{P(X|Y) = \frac{P(X)}{P(Y)}P(Y|X)}$$

Statistical PoS tagging

- Determine the **most likely** tag sequence $t_{1:n}$ by:

$$\underset{t_{1:n}}{\operatorname{argmax}} P(t_{1:n}|w_{1:n}) = \underset{t_{1:n}}{\operatorname{argmax}} \frac{P(w_{1:n}|t_{1:n})P(t_{1:n})}{P(w_{1:n})}$$

By Bayes' Rule

$$= \underset{t_{1:n}}{\operatorname{argmax}} \frac{P(w_{1:n}|t_{1:n})P(t_{1:n})}{\cancel{P(w_{1:n})}}$$

Only maximize numerator

$$\boxed{\approx \underset{t_{1:n}}{\operatorname{argmax}} \prod_{i}^{n} \underbrace{P(w_i|t_i)}_{\text{Assuming independence}} \underbrace{P(t_i|t_{i-1})}_{\text{Assuming Markov}}}$$

UNIVERSITY OF TORONTO

Word likelihood probability $P(w_i|t_i)$

- **VBZ** (verb, 3rd person singular present) is likely ***is***.
- Compute $P(\boldsymbol{is}|VBZ)$ by **counting** in a corpus that has **already** been **tagged**:

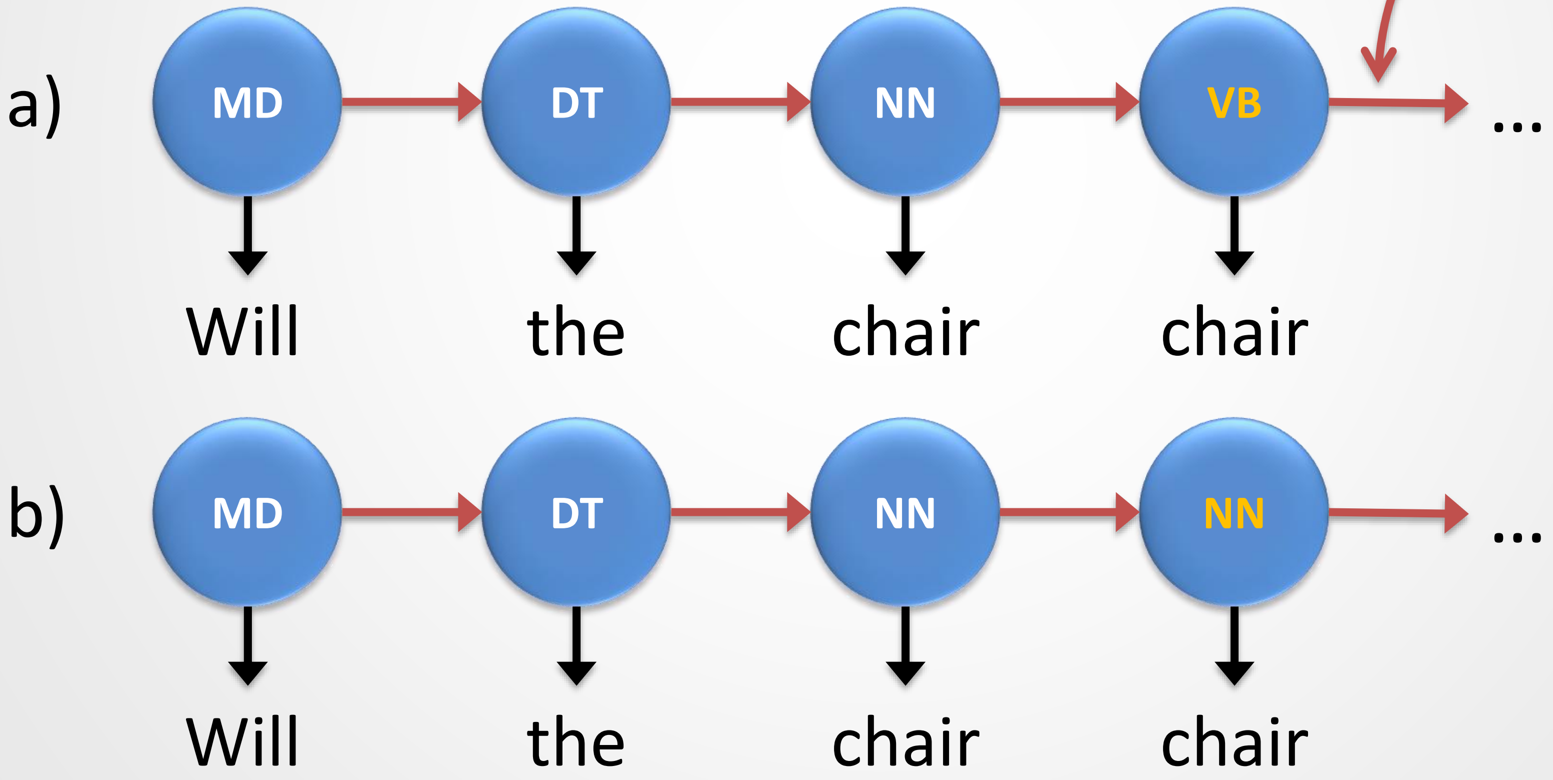
$$P(w_i|t_i) = \frac{Count(w_i \text{ tagged as } t_i)}{Count(t_i)}$$

e.g.,

$$P(\boldsymbol{is}|\boldsymbol{VBZ}) = \frac{Count(\boldsymbol{is} \text{ tagged as } \boldsymbol{VBZ})}{Count(\boldsymbol{VBZ})} = \frac{10{,}073}{21{,}627} = 0.47$$

Tag-transition probability $P(t_i|t_{i-1})$

- *Will/MD the/DT* ***chair/NN chair/??*** *the/DT meeting/NN from/IN that/DT* ***chair/NN****?*



Those are hidden Markov models!

- We'll see these soon…

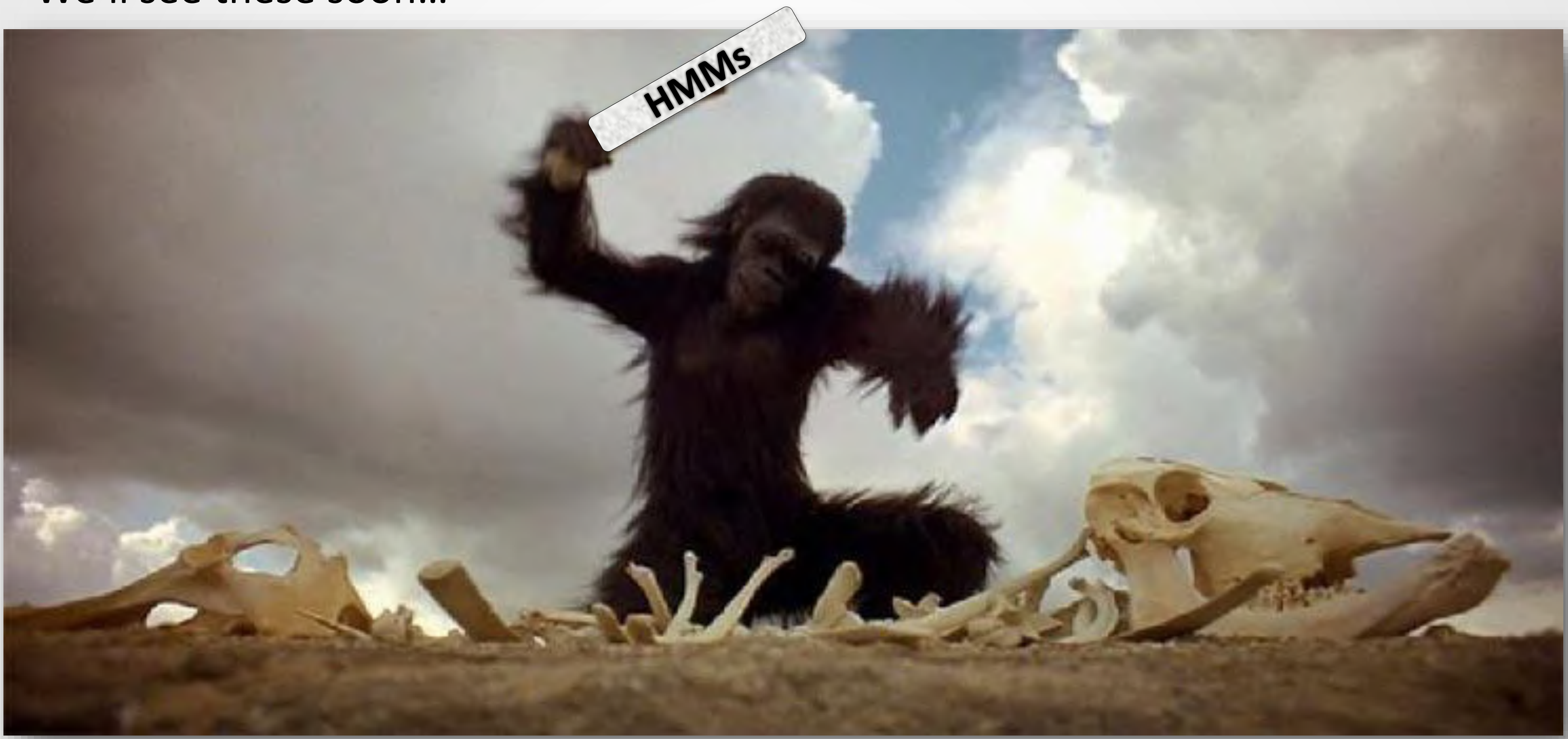


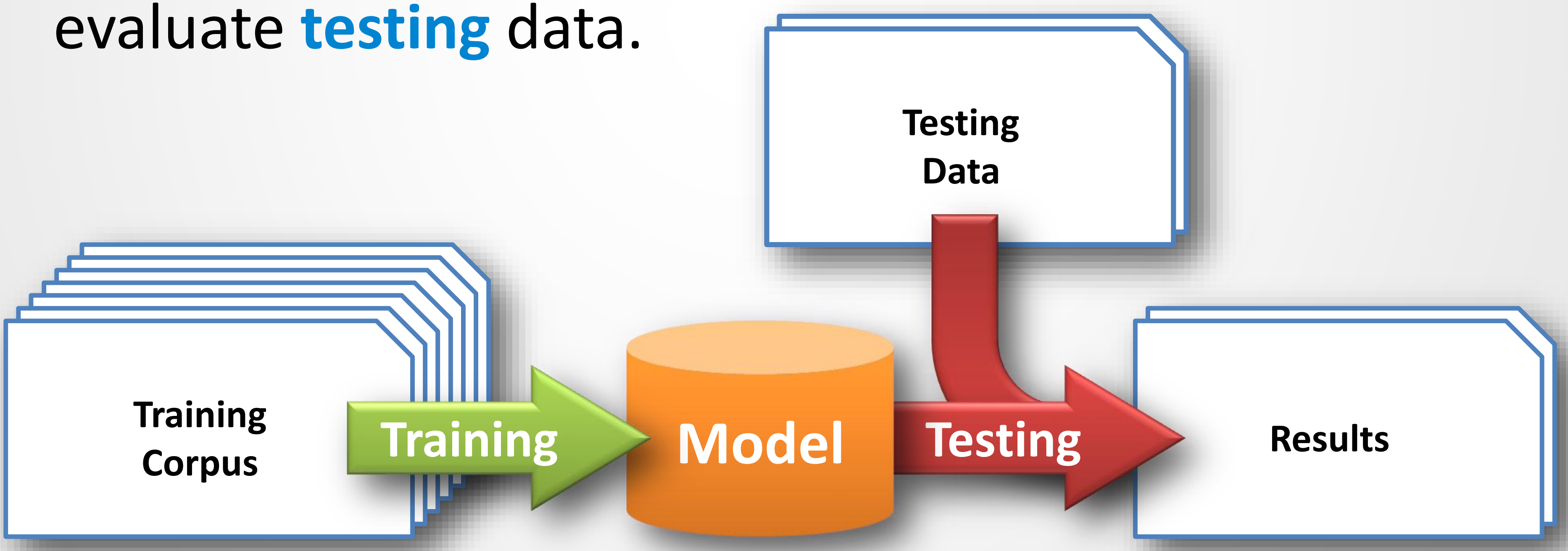
Image sort of from *2001:A Space Odyssey*
by MGM pictures

UNIVERSITY OF TORONTO

Classification

General process

1. We gather a big and relevant **training** corpus.
2. We learn our **parameters** (e.g., probabilities) from that corpus to build our **model**.
3. Once that model is fixed, we use those probabilities to evaluate **testing** data.



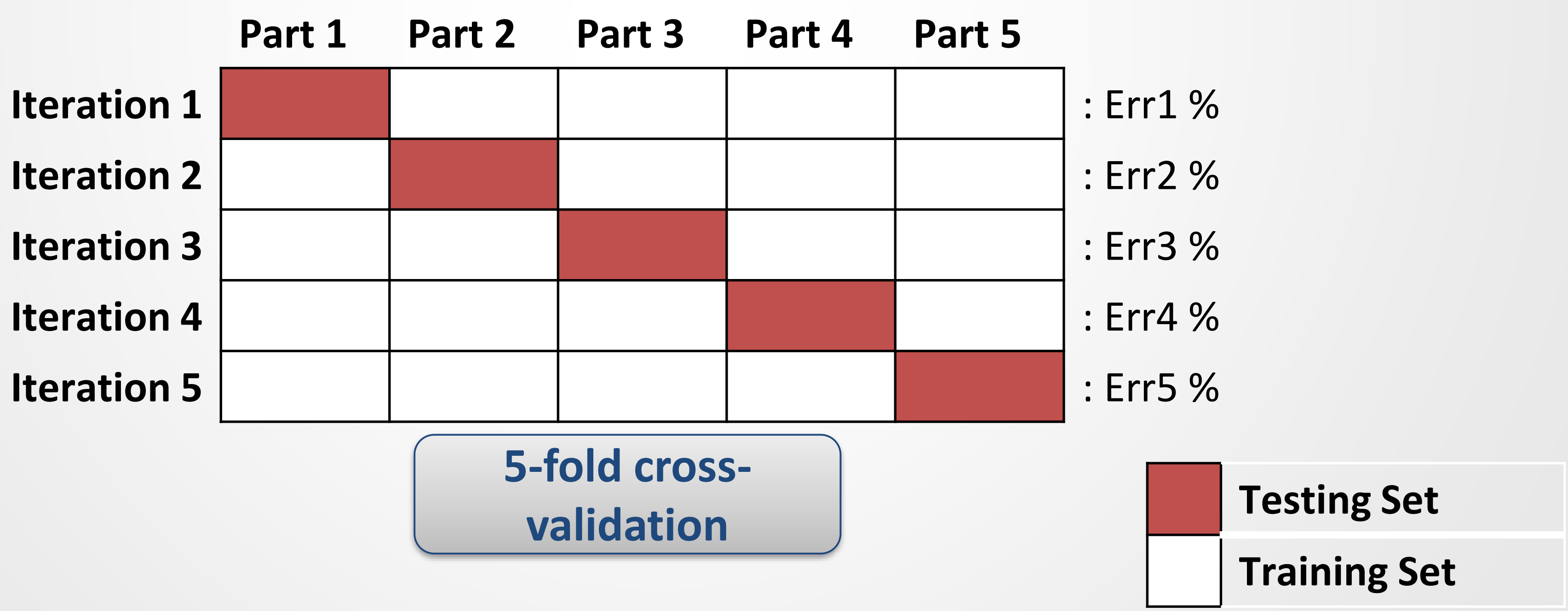
General process

- Often, **training data** consists of 80% to 90% of the available data.
  - Often, some subset of *this* is used as a **validation/development set**.

- **Testing data** is ***not*** used for training but comes from the same *corpus*.
  - It often consists of the remaining available data.
  - Sometimes, it's important to **partition** speakers/writers so they **don't** appear in both training and testing.
  - *But what if we just randomized (un)luckily??*

Better process: *K*-fold cross-validation

- ***K*-fold cross validation**: *n.* splitting all data into *K* **partitions** and iteratively testing on each after training on the rest (report means and variances).



(Some) Types of classifiers

- **Generative** classifiers model the world.
  - Parameters set to maximize likelihood of training data.
  - We can *generate* new observations from these.
    - e.g., hidden Markov models

Vs.

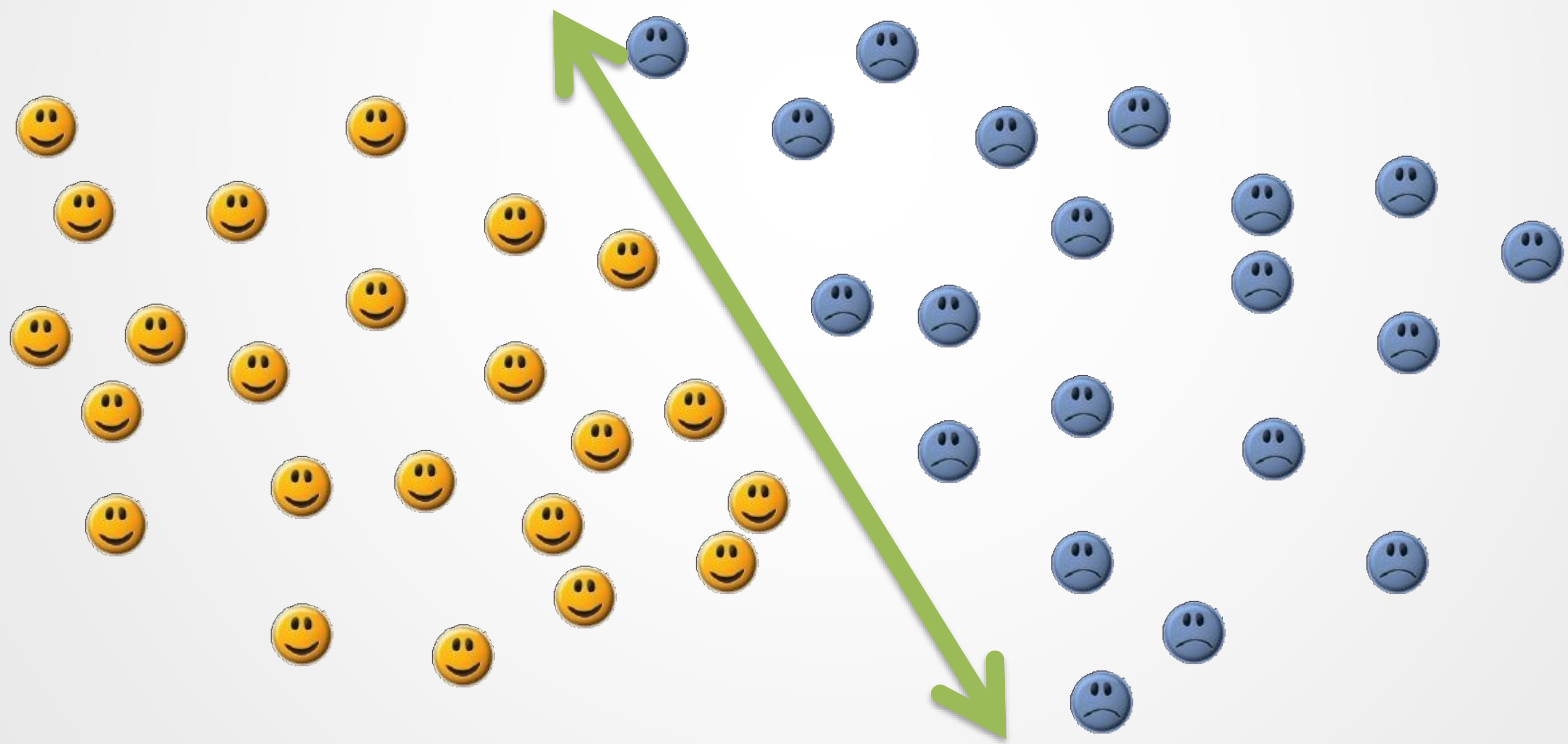
- **Discriminative** classifiers emphasize **class boundaries**.
  - Parameters set to minimize error on training data.
    - e.g., support vector machines, decision trees.

- …

What do class boundaries look like in the data?

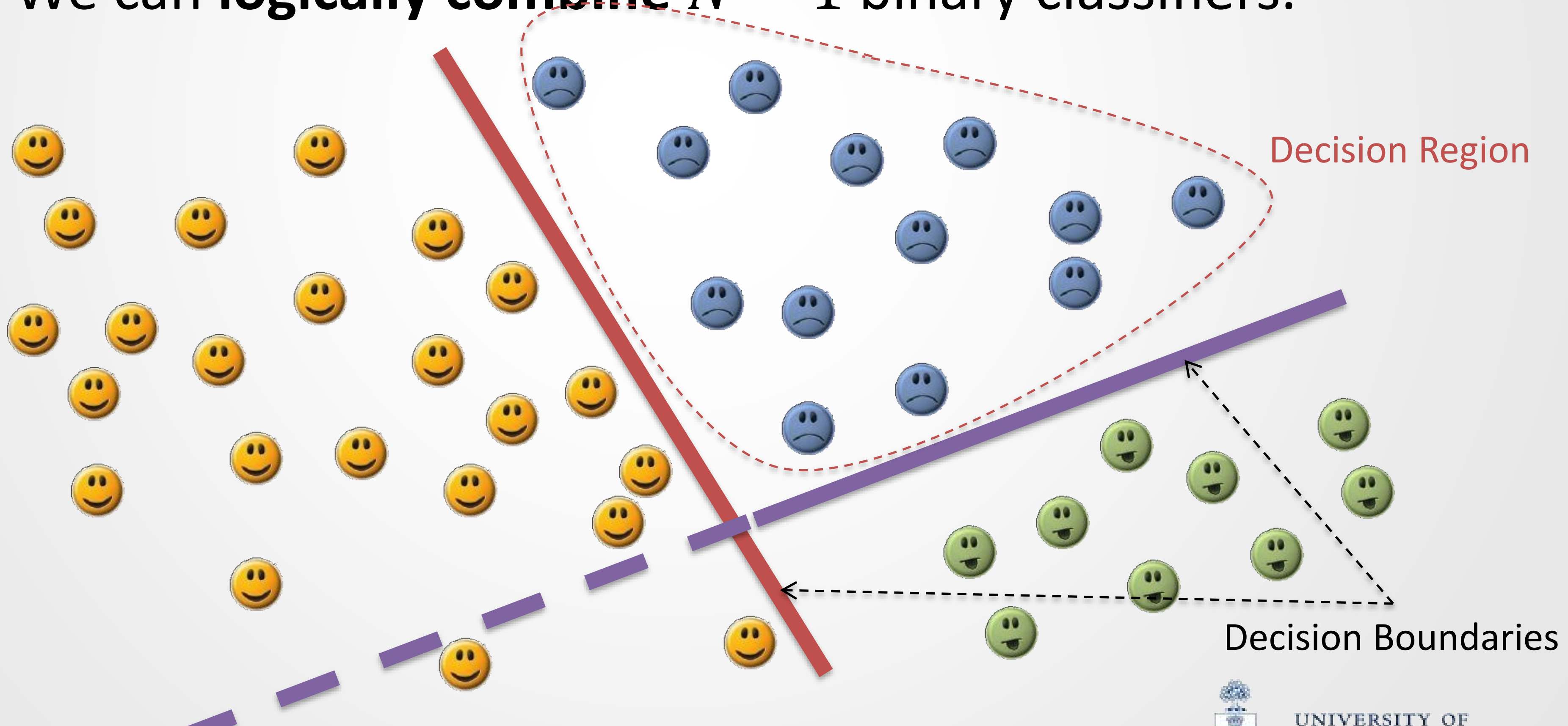
Binary and linearly separable

- Perhaps the easiest case.
  - Extends to dimensions $d \geq 3$, line becomes (hyper-)plane.



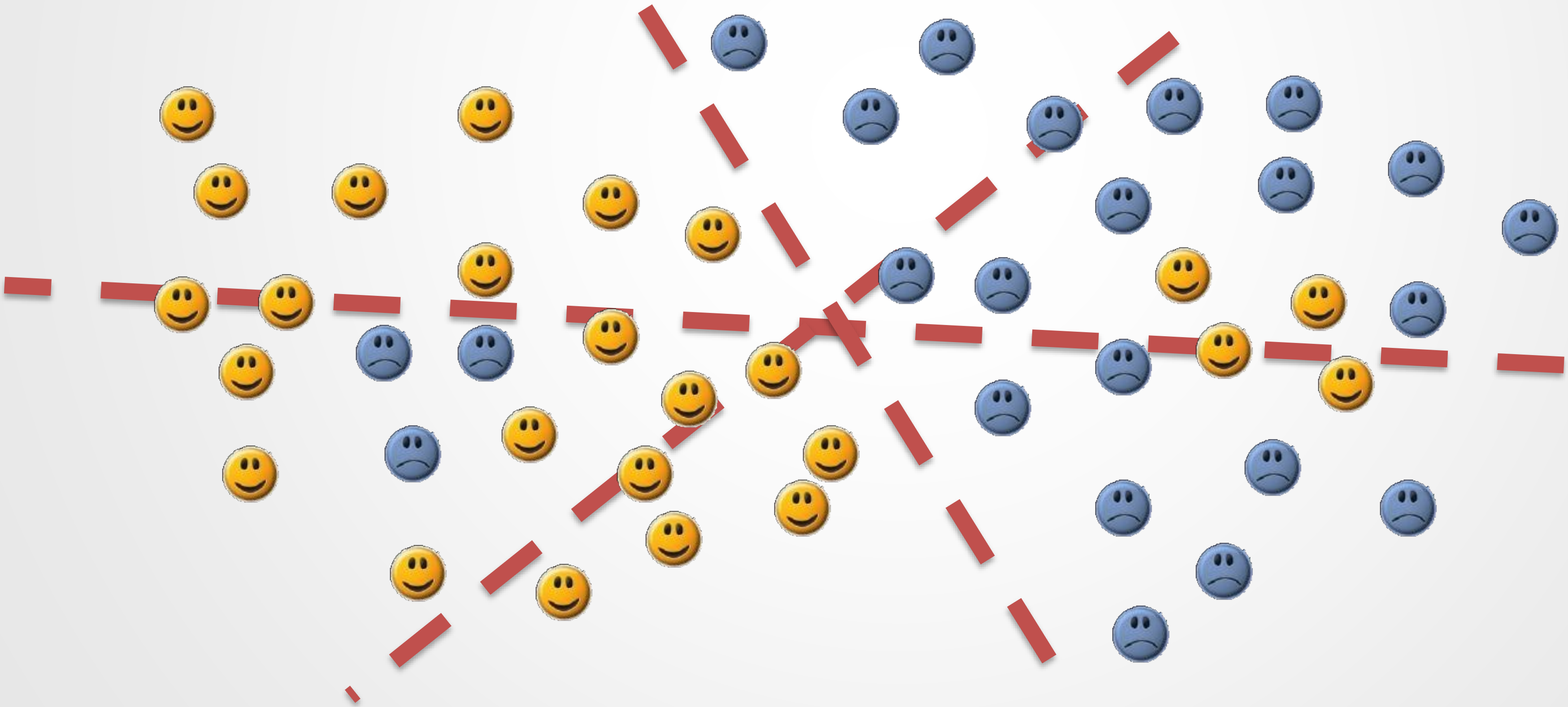
N-ary and linearly separable

- A bit harder – random guessing gives $\frac{1}{N}$ accuracy (given equally likely classes).
  - We can **logically combine** $N-1$ binary classifiers.



Class holes

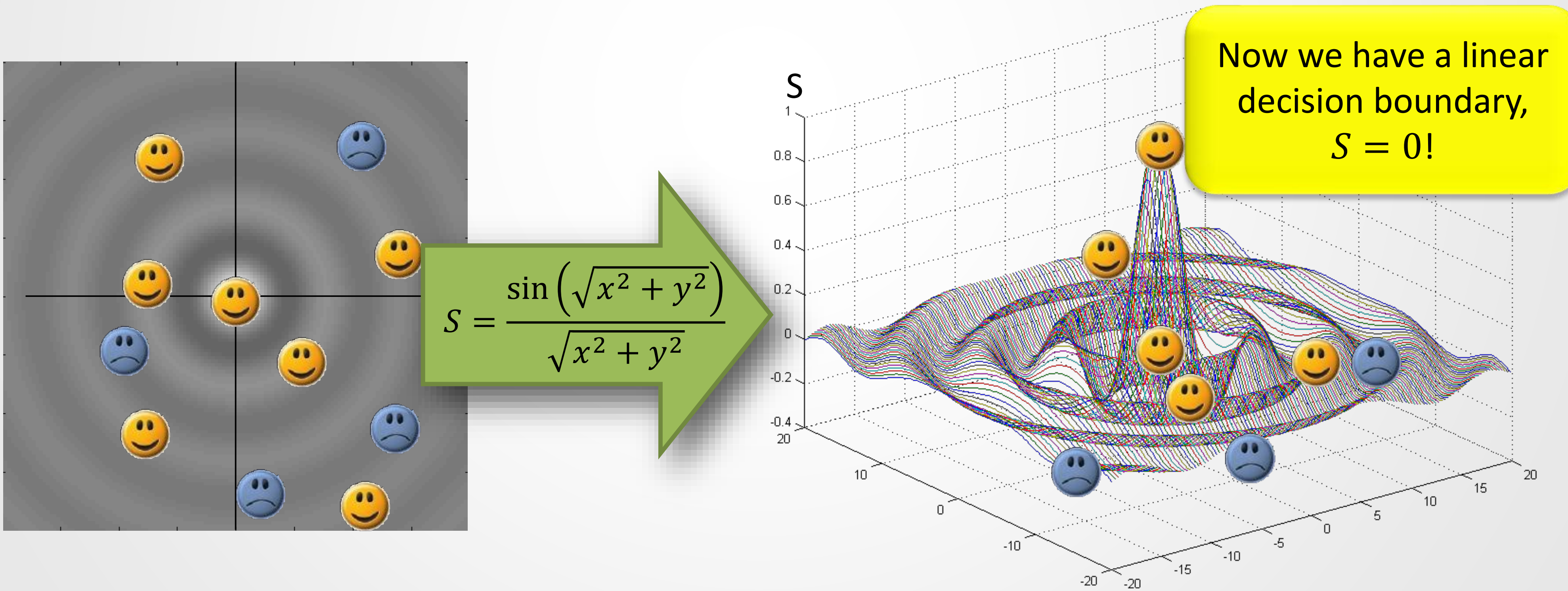
- Sometimes it can be impossible to draw *any* lines through the data to separate the classes.
  - *Are those troublesome points noise or real phenomena?*



The kernel trick

- We can sometimes linearize a non-linear case by moving the data into a higher dimension with a **kernel function**.

  E.g.,

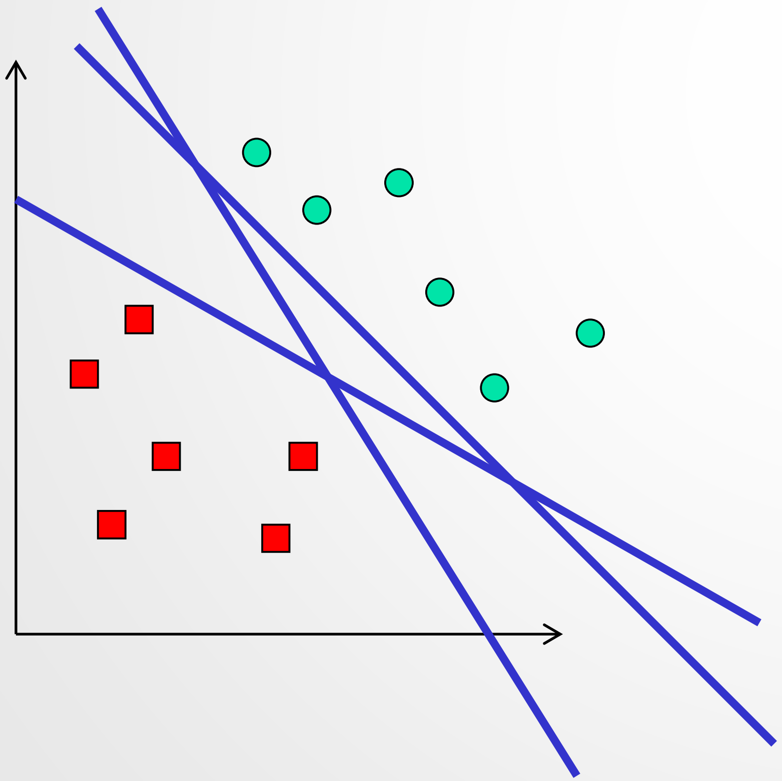


```
from sklearn.SVM import SVC
```

Support Vector Machines (SVMs)

Support vector machines (SVMs)

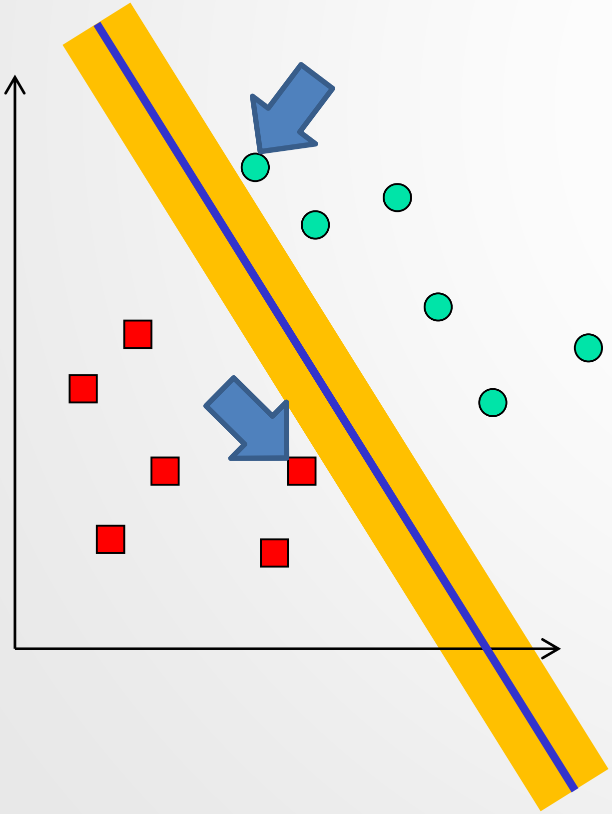
- In binary linear classification, two classes are assumed to be separable by a line (or plane). However, many possible separating planes might exist.



- Each of these blue lines separates the training data.
  - *Which line is the best?*

Support vector machines (SVMs)

- The **margin** is the width by which the boundary could be **increased** before it hits a training datum.

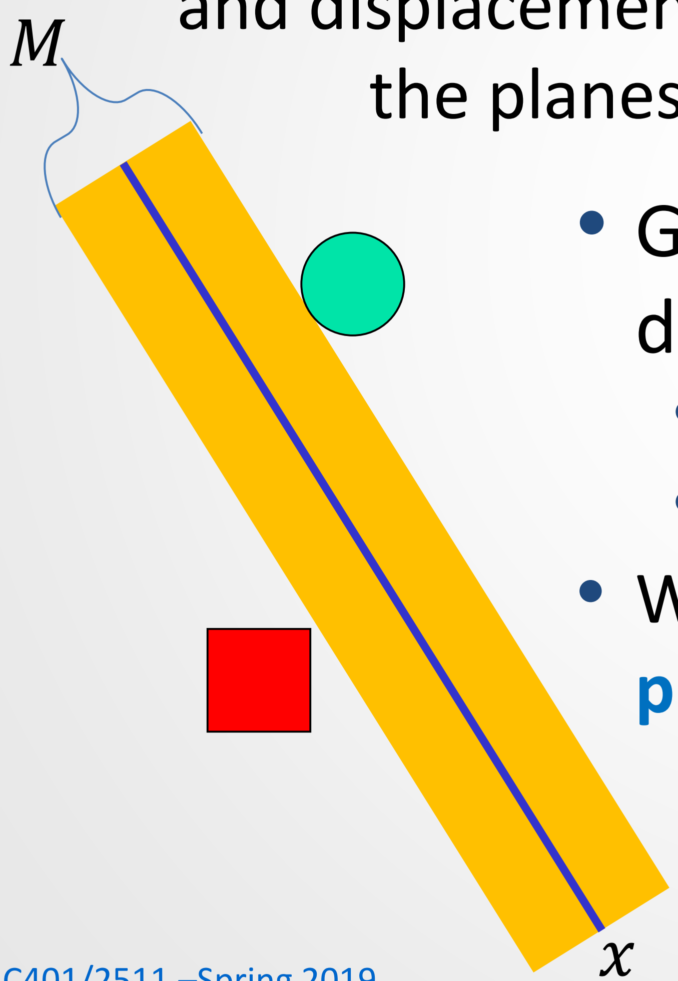


- The **maximum margin linear classifier** is ∴ the linear classifier with the maximum margin.
- The **support vectors** (indicated) are those data points against which the margin is pressed.
- The bigger the margin – the less sensitive the boundary is to error.

UNIVERSITY OF TORONTO

Support vector machines (SVMs)

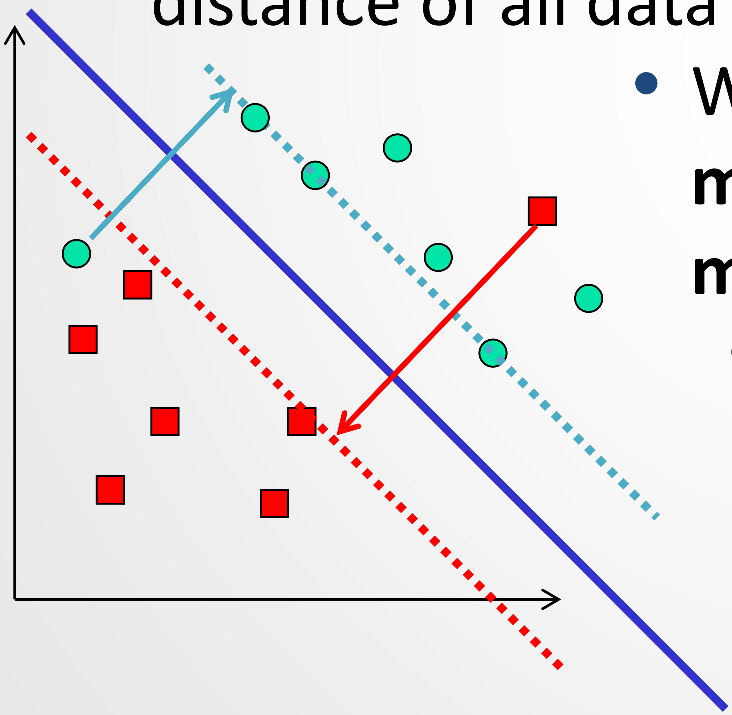
- The width of the margin, M, can be computed by the angle and displacement of the planar boundary, x, as well as the planes that touch data points.



- Given an initial guess of the angle and displacement of x we can compute:
  - whether all data is correctly classified,
  - The width of the margin, M.
- We update our guess by **quadratic programming**, which is semi-analytic.

Support vector machines (SVMs)

- The maximum margin helps SVMs **generalize** to situations when it's **impossible** to linearly separate the data.
  - We introduce a parameter that allows us to measure the distance of all data not in their correct 'zones'.



- We simultaneously **maximize the margin** while **minimizing the misclassification error**.
  - There is a straightforward approach to solving this system based on **quadratic programming**.

UNIVERSITY OF TORONTO

Support vector machines (SVMs)

- SVMs generalize to higher-dimensional data and to systems in which the data is non-linearly separable (e.g., by a circular decision boundary).
  - Using the **kernel trick** (from before) is common.

- Many binary SVM classifiers can also be combined to simulate a multi-category classifier.

- (Still) one of the most popular off-the-shelf classifiers.

Support vector machines (SVMs)

- SVMs are empirically **very accurate** classifiers.
  - They perform well in situations where data are **static**, i.e., don't change over time, e.g.,
    - **genre classification** given **fixed statistics** of documents

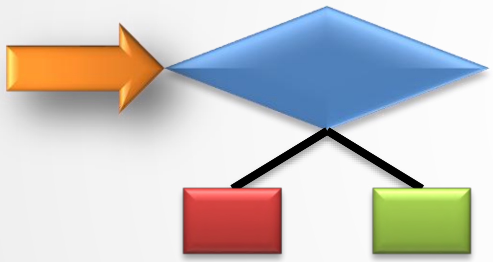
- SVMs do **not generalize** as well to **time-variant** systems.
  - Kernel functions tend to not allow for observations of **different lengths** (i.e., all data points have to be of the same dimensionality).

Trees!

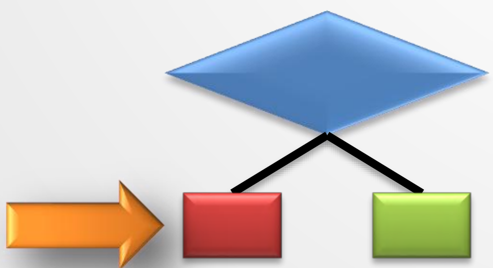
(The larch.)

Decision trees

- Consists of **rules** for classifying data that have many **attributes** (features).
  - **Decision nodes**: **Non-terminal**. Consists of a *question* asked of one of the attributes, and a *branch* for each possible answer.



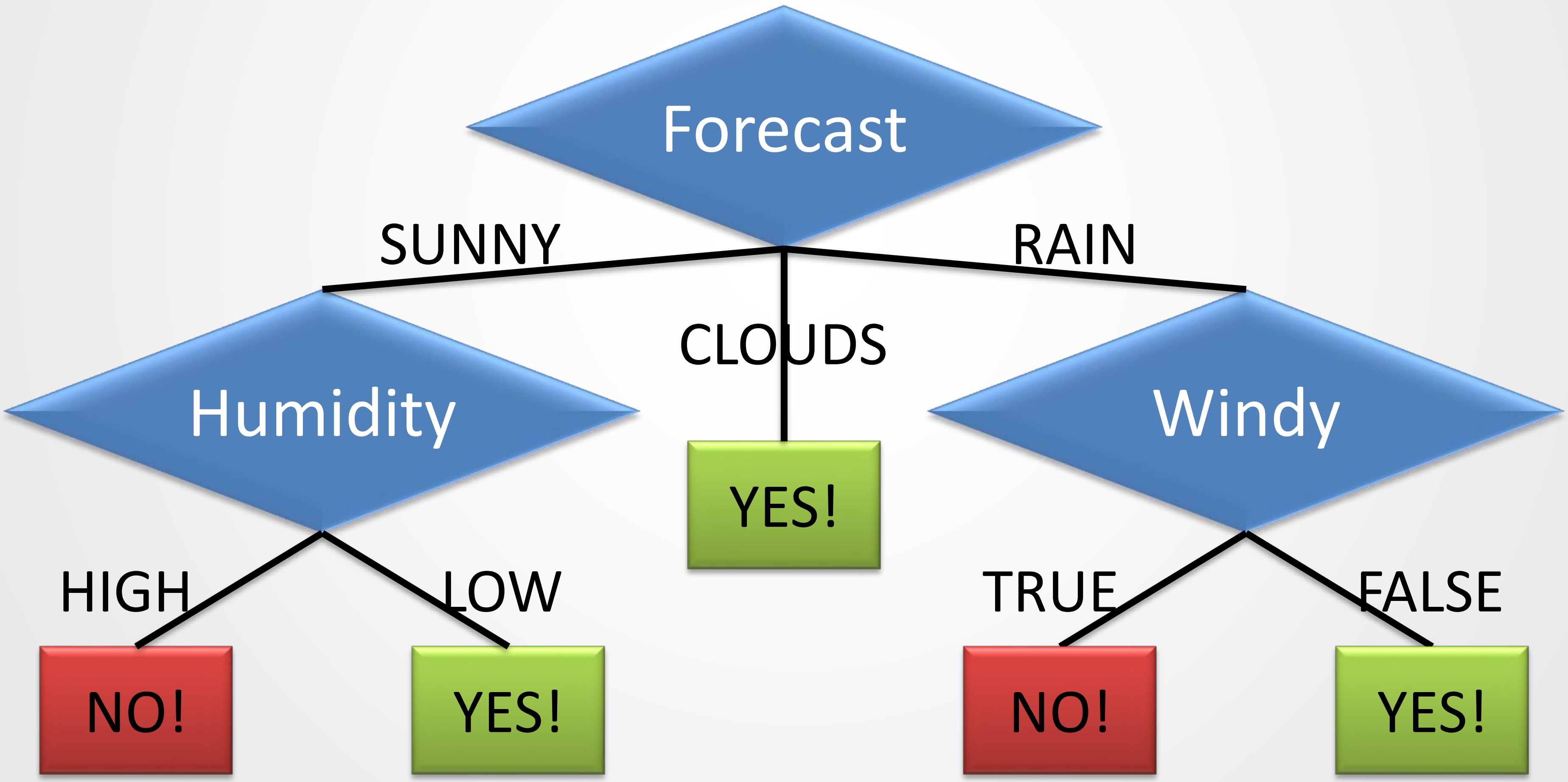
  - **Leaf nodes**: **Terminal**. Consists of a single class/category, so no further testing is required.



UNIVERSITY OF TORONTO

Decision tree example

- Shall I go for a walk?



UNIVERSITY OF TORONTO

Decision tree algorithm: ID3

- **ID3** (iterative dichotomiser 3) is an algorithm invented by Ross Quinlan to produce decision trees from data.

- Basically,
  1. Compute the **entropy** of asking about **each attribute**.
  2. Choose the attribute which **reduces** the most entropy.
  3. **Make a node** asking a question of that attribute.
  4. Go to step 1, **minus** the chosen attribute.

- Example attribute vectors (observations):

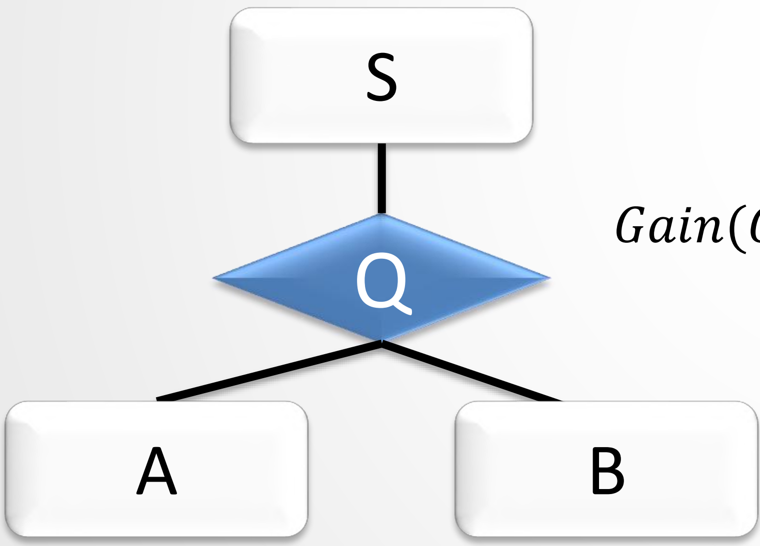
Forecast	Humidity	Wind

Avg. token length	Avg. sentence length	Frequency of nouns	…

UNIVERSITY OF TORONTO

Information gain

- The **information gain** is based on the expected decrease in entropy after a set of **training** data is split on an attribute.
  - We prefer the attribute that removes the most entropy.



$$Gain(Q) = H(S) - \sum_{child\ set} p(child\ set) H(child\ set)$$

$$S = A \cup B$$

$$\emptyset = A \cap B$$

Each of S, A, and B consist of examples from the data

So $p(child\ set)$ is computed by the proportion of examples in that set

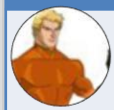
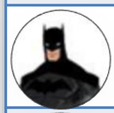
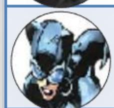
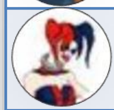
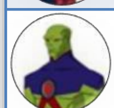
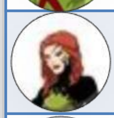
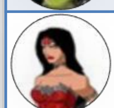
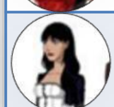
Information gain and ID3

- When a node in the decision tree is **generated** in which **all members** have the **same** class,
  - that node has 0 entropy,
  - that node is a leaf node.

- Otherwise, we need to (try to) split that node with another question.

Example – Hero classification

Training data

	Hero	Hair length	Height	Age	Hero Type
	Aquaman	2”	6’2”	35	Hero
	Batman	1”	5’11”	32	Hero
	Catwoman	7”	5’9”	29	Villain
	Deathstroke	0”	6’4”	28	Villain
	Harley Quinn	5”	5’0”	27	Villain
	Martian Manhunter	0”	8’2”	128	Hero
	Poison Ivy	6”	5’2”	24	Villain
	Wonder Woman	6”	6’1”	108	Hero
	Zatanna	10”	5’8”	26	Hero

Test data

	Red Hood	2”	6’0”	22	?

Example – Hero classification

- How do we split?
  - Split on *hair length?*
  - Split on *height?*
  - Split on *age?*

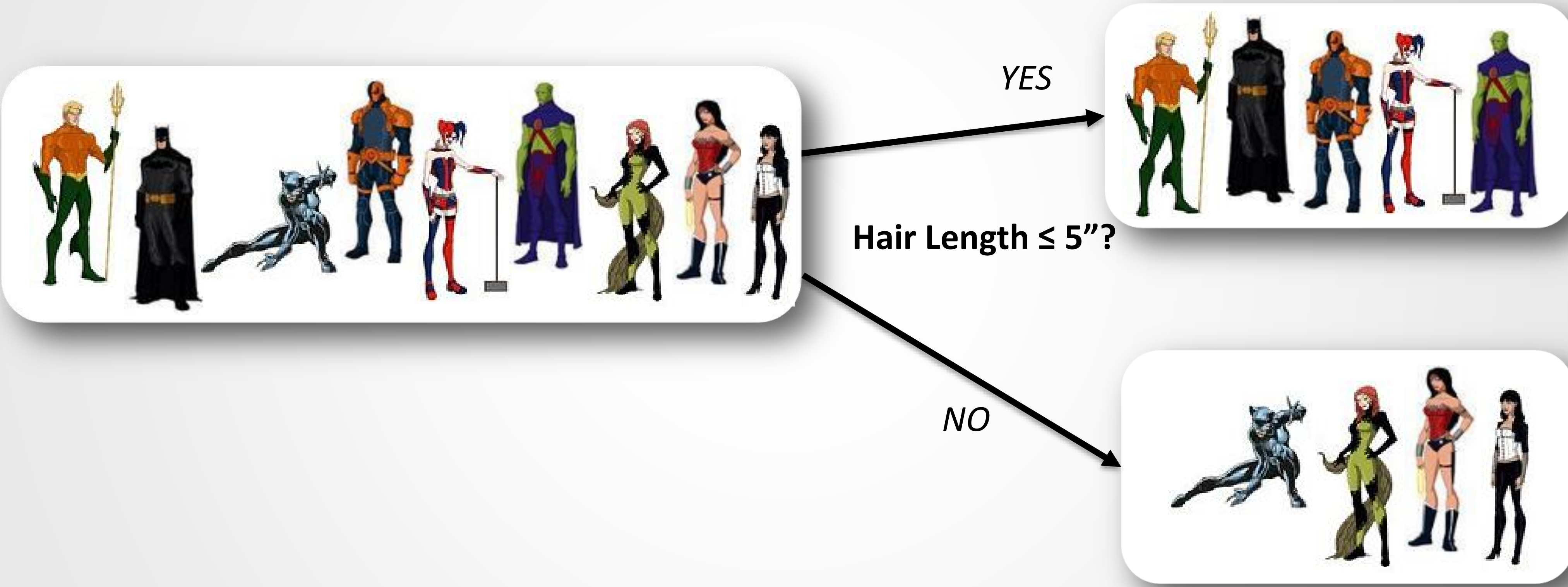
- Let's compute the **information gain** for each:

$$Gain(Q) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

UNIVERSITY OF TORONTO

Split on hair length?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$



Split on hair length?

$$Gain(Question) = \boxed{H(S)} - \sum_{child\ set} p(child\ set) H(child\ set)$$

YES

Hair Length ≤ 5”?

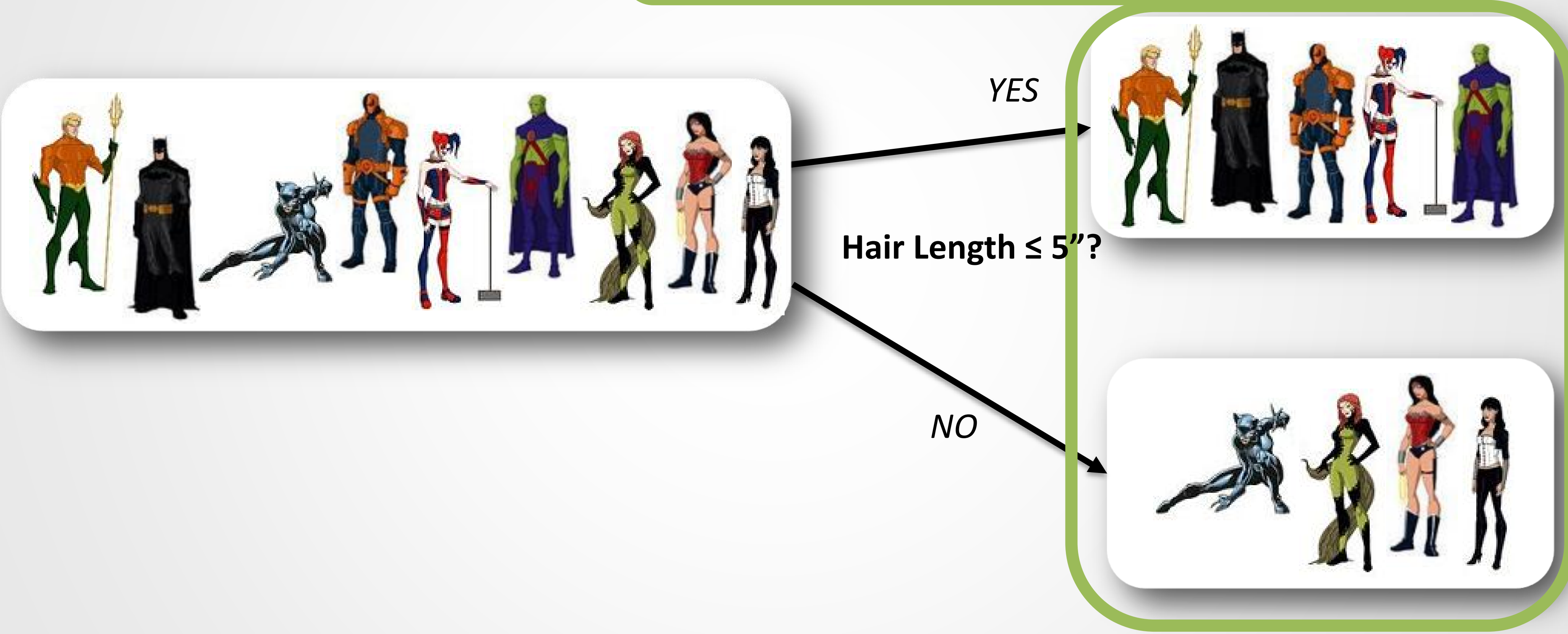
NO

$$H(S) = \frac{h}{h+v}\log_2\left(\frac{h+v}{h}\right) + \frac{v}{h+v}\log_2\left(\frac{h+v}{v}\right)$$

$$H(5h, 4v) = \frac{5}{9}\log_2\left(\frac{9}{5}\right) + \frac{4}{9}\log_2\left(\frac{9}{4}\right) = \mathbf{0.9911}\ \text{bits}$$

Split on hair length?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$



Split on hair length?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

YES

Hair Length ≤ 5"?

NO

YES $H(\mathbf{4}h, \mathbf{1}v) = \frac{4}{5}\log_2\left(\frac{5}{4}\right) + \frac{1}{5}\log_2\left(\frac{5}{1}\right) = \mathbf{0.7219}$

UNIVERSITY OF TORONTO

Split on hair length?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

Hair Length ≤ 5"?

YES

NO

YES $H(\mathbf{4}h, \mathbf{1}v) = \frac{4}{5}\log_2\left(\frac{5}{4}\right) + \frac{1}{5}\log_2\left(\frac{5}{1}\right) = \mathbf{0.7219}$

NO $H(\mathbf{2}h, \mathbf{2}v) = \frac{2}{4}\log_2\left(\frac{4}{2}\right) + \frac{2}{4}\log_2\left(\frac{4}{2}\right) = \mathbf{1}$

UNIVERSITY OF TORONTO

Split on hair length?

$$Gain(Question) = \boxed{H(S)} - \boxed{\sum_{child\ set} p(child\ set)H(child\ set)}$$

YES

Hair Length ≤ 5”?

NO

$$Gain(HairLength \leq 5") = \underbrace{0.9911}\ \underbrace{-\frac{5}{9}\mathbf{0.7219} - \frac{4}{9}\mathbf{1}} = \mathbf{0.00721}$$

UNIVERSITY OF TORONTO

Example – Hero classification

- How do we split?
  - Split on *hair length?* $Gain(HairLength \leq 5") = \mathbf{0.00721}$
  - Split on *height?*
  - Split on *age?*

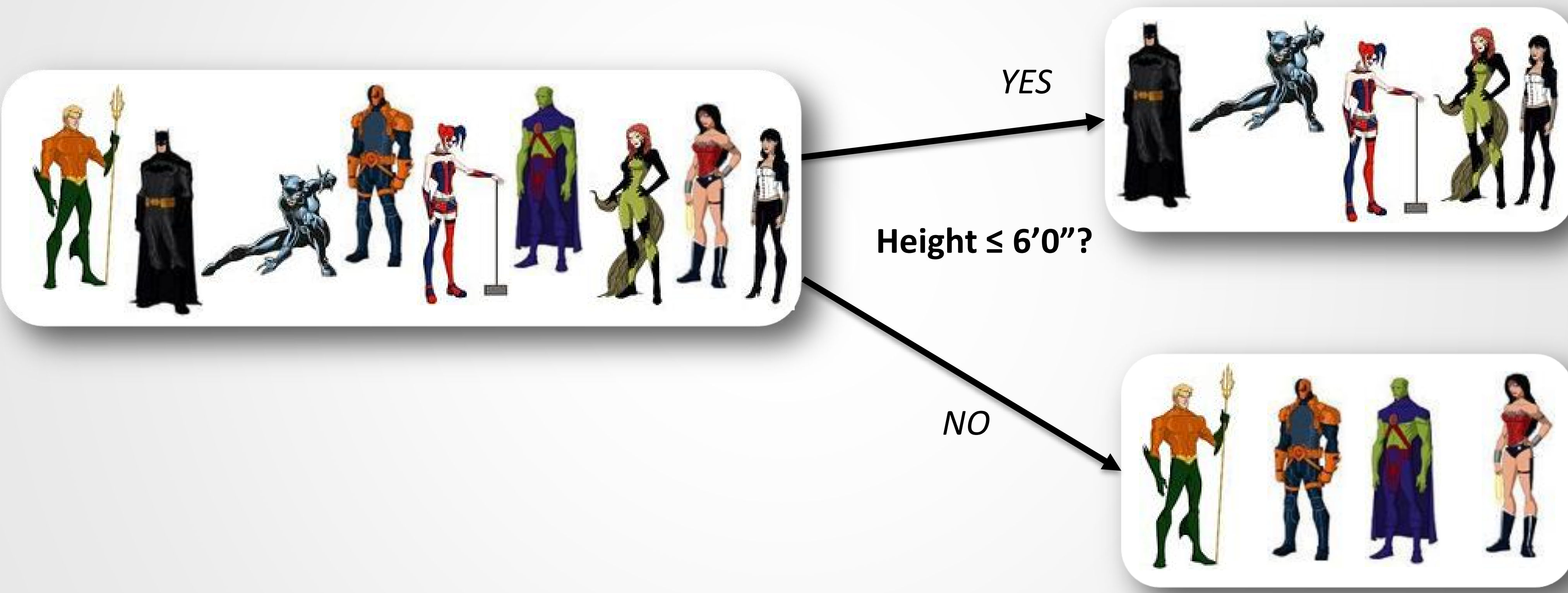
- Let's compute the **information gain** for each:

$$Gain(Q) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

UNIVERSITY OF TORONTO

Split on height?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$



UNIVERSITY OF TORONTO

Split on height?

$$Gain(Question) = \boxed{H(S)} - \sum_{child\ set} p(child\ set)H(child\ set)$$

YES

Height ≤ 6’0”?

NO

$$H(S) = \frac{h}{h+v}\log_2\left(\frac{h+v}{h}\right) + \frac{v}{h+v}\log_2\left(\frac{h+v}{v}\right)$$

$$H(5h, 4v) = \frac{5}{9}\log_2\left(\frac{9}{5}\right) + \frac{4}{9}\log_2\left(\frac{9}{4}\right) = \mathbf{0.9911}\ \text{bits}$$

Split on height?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

YES

Height ≤ 6'0"

NO

YES $H(\mathbf{2}h, \mathbf{3}v) = \frac{2}{5}\log_2\left(\frac{5}{2}\right) + \frac{3}{5}\log_2\left(\frac{5}{3}\right) = \mathbf{0.971}$

NO $H(\mathbf{3}h, \mathbf{1}v) = \frac{3}{4}\log_2\left(\frac{4}{3}\right) + \frac{1}{4}\log_2\left(\frac{4}{1}\right) = \mathbf{0.813}$

UNIVERSITY OF TORONTO

Split on height?

$$Gain(Question) = \boxed{H(S)} - \boxed{\sum_{child\ set} p(child\ set)H(child\ set)}$$

YES

Height ≤ 6’0”

NO

$$Gain(Height \leq 6'0") = 0.9911 - \frac{5}{9}[0.971] - \frac{4}{9}[0.813] = 0.0903$$

Example – Hero classification

- How do we split?
  - Split on *hair length?* $Gain(HairLength \leq 5") = \mathbf{0.00721}$
  - Split on *height?*
  - Split on *age?*

- Let's compute the **information gain** for each:

$$Gain(Q) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

Split on age?

$$Gain(Question) = \boxed{H(S)} - \sum_{child\ set} p(child\ set)H(child\ set)$$

YES

Age ≤ 30?

NO

$$H(S) = \frac{h}{h+v}\log_2\left(\frac{h+v}{h}\right) + \frac{v}{h+v}\log_2\left(\frac{h+v}{v}\right)$$

$$H(5h, 4v) = \frac{5}{9}\log_2\left(\frac{9}{5}\right) + \frac{4}{9}\log_2\left(\frac{9}{4}\right) = \mathbf{0.9911}\text{ bits}$$

UNIVERSITY OF TORONTO

Split on age?

$$Gain(Question) = H(S) - \sum_{child\ set} p(child\ set)H(child\ set)$$

YES

Age ≤ 30?

NO

YES $H(\mathbf{1}h, \mathbf{4}v) = \frac{1}{5}\log_2\left(\frac{5}{1}\right) + \frac{4}{5}\log_2\left(\frac{5}{4}\right) = \mathbf{0.722}$

NO $H(\mathbf{4}h, \mathbf{0}v) = \frac{4}{4}\log_2\left(\frac{4}{4}\right) + \frac{0}{4}\log_2(\infty) = \mathbf{0}$

UNIVERSITY OF TORONTO

Split on age?

$$Gain(Question) = H(S) - \boxed{\sum_{child\ set} p(child\ set)H(child\ set)}$$

$$Gain(Age \leq 30) = 0.9911 - \frac{5}{9}[0.722] - \frac{4}{9}[0] = 0.590$$

Example – Hero classification

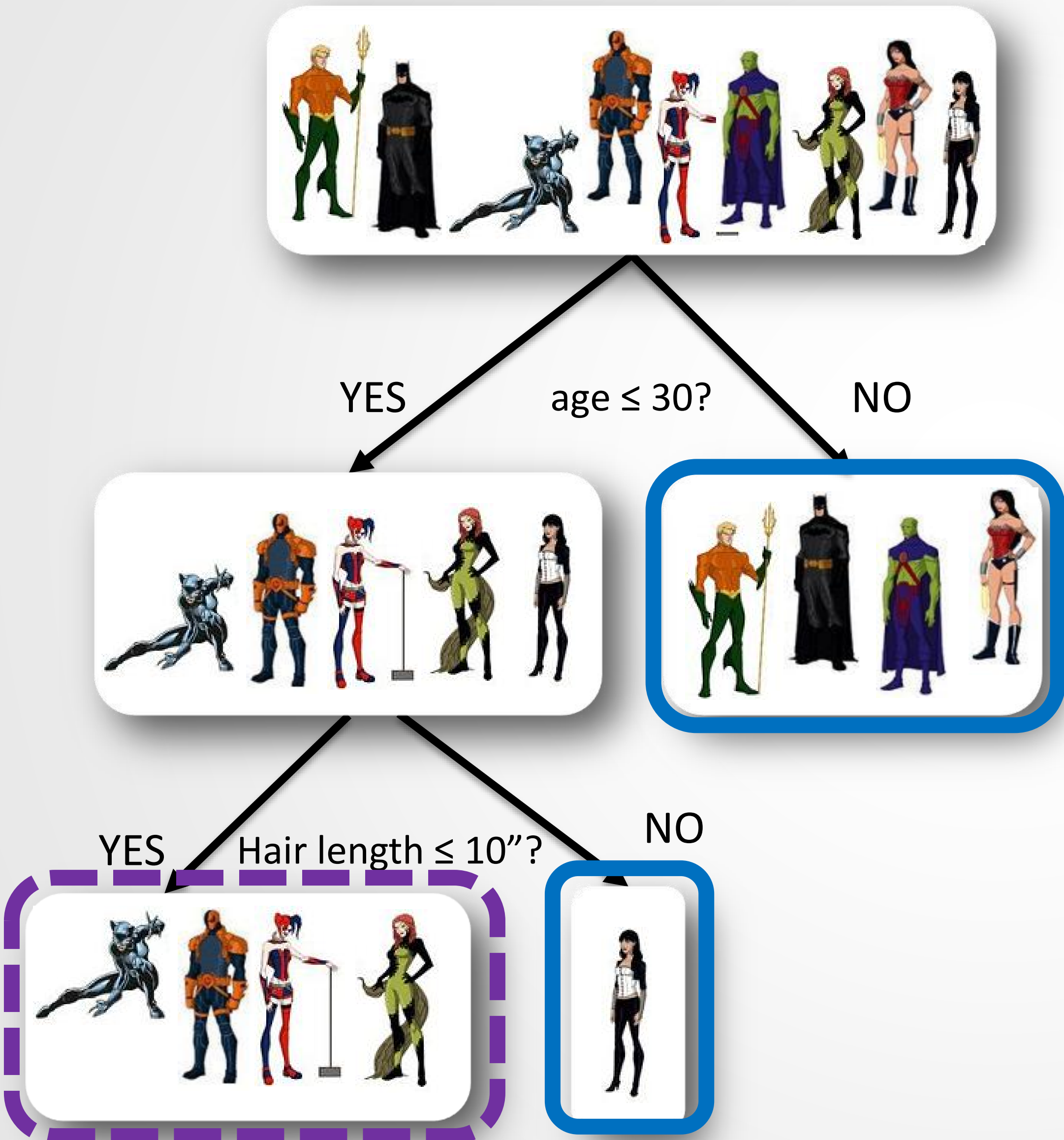
- How do we split?
  - Split on *hair length?* $Gain(HairLength \leq 5") = \mathbf{0.00721}$
  - Split on *height?* $Gain(Height \leq 6'0") = \mathbf{0.0903}$
  - Split on *age?* $Gain(Age \leq 30) = \mathbf{0.590}$

- Let's compute the **information gain** for each:

$$Gain(Q) = H(S) - \sum_{child\ set} p(child\ set) H(child\ set)$$

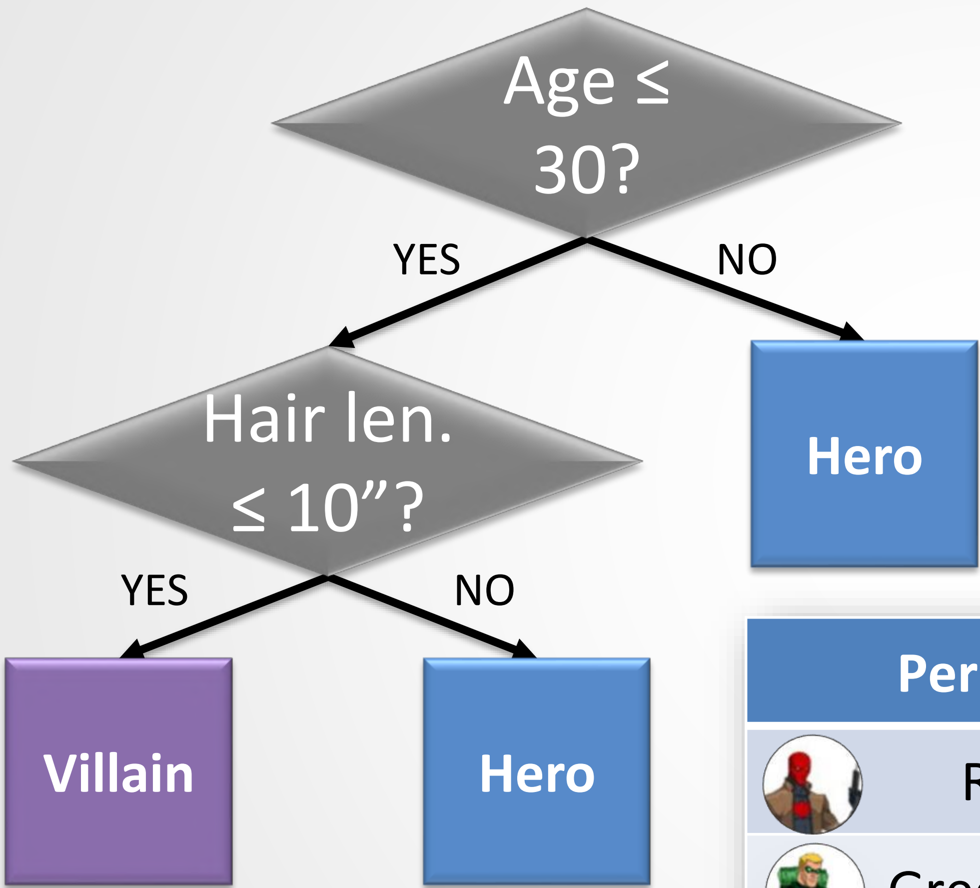
UNIVERSITY OF TORONTO

The resulting tree



- Splitting on *age* resulted in the greatest information gain.
- We’re left with one heterogeneous set, so we recurse and find that *hair length* results in a complete classification of the training data.

Testing



- We just need to keep track of the attribute questions – not the training data.
- How are the following characters classified?

Person	Hair length	Height	Age
Red Hood	2”	6’0”	22
Green Arrow	1”	6’2”	38
Bane	0”	5’8”	29

- Inspired from Allan Neymark’s (San Jose State University) Simpsons example.

Aspects of ID3

- ID3 tends to build **short trees** since at each step we are removing the maximum amount of entropy possible.
- ID3 trains on the **whole training set** and does not succumb to issues related to **random initialization**.

- ID3 can **over-fit** to training data.
- Only **one attribute is used at a time** to make decisions
- It can be difficult to use **continuous** data, since many trees need to be generated to see where to break the continuum.

```
from sklearn.ensemble import RandomForestClassifier
```

Random Forests

- **Random forests** *n.pl.* are **ensemble** classifiers that produce K decision trees, and output the **mode** class of those trees.
  - Can support continuous features.
  - Can support non-binary decisions.
  - Support cross-validation.

- The component trees in a random forest must differ.
  - Sometimes, decision trees are **pruned** randomly.
  - Usually, different trees accept different **subsets of features**.

That's a good idea – can we choose the best features in a reasonable way?

Feature selection

Determining a good set of features

- **Restricting** your feature set to a proper subset quickens **training** and reduces **overfitting**.

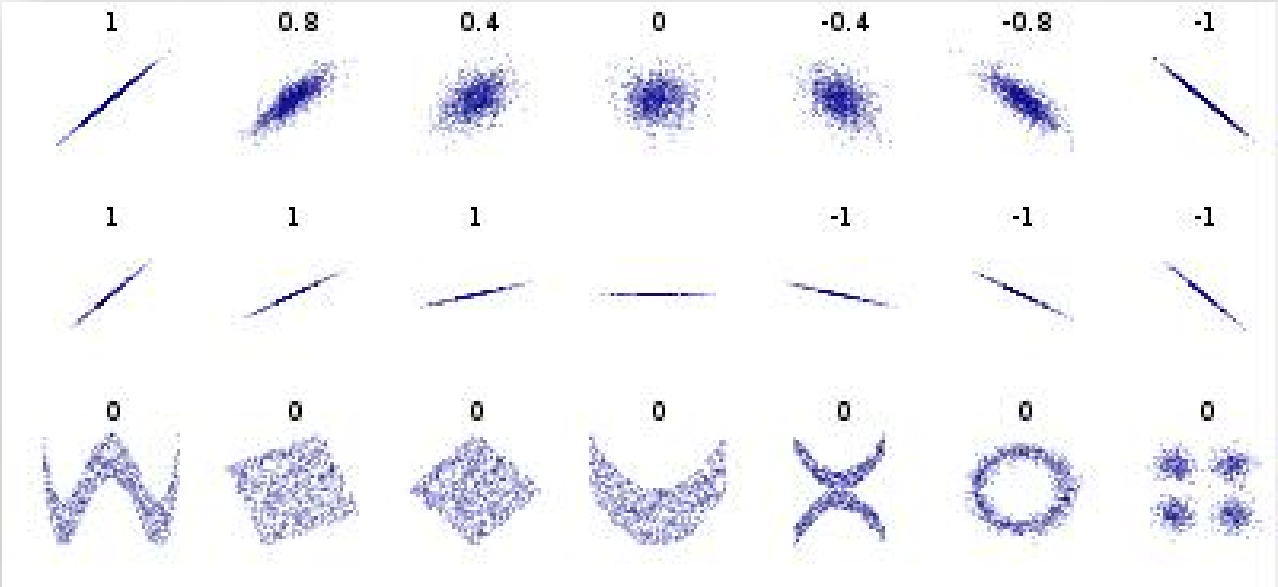
- There are a few methods that select good features, e.g.,
  1. Correlation-based feature selection
  2. Minimum Redundancy, Maximum Relevance
  3. χ^2

1. Pearson's correlation

- **Pearson** is a measure of **linear** dependence

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- Does not measure '**slope**' nor **non-linear** relations.



UNIVERSITY OF TORONTO

1. Spearman's correlation

- **Spearman** is a non-parametric measure of **rank** correlation, $r_{cX} = r(c, X)$.
  - It is basically Pearson's correlation, but on 'rank variables' that are monotonically increasing integers.
  - If the class c can be **ordered** (e.g., in any binary case), then we can compute the correlation between a feature X and that class.

1. Correlation-based feature selection

- 'Good' features should correlate **strongly** (+ or -) with the ***predicted variable*** but **not** with other ***features***.

- S_{CFS} is some set S of k features f_i that maximizes this ratio, given class c:

$$S_{CFS} = \underset{S}{\operatorname{argmax}} \frac{\sum_{f_i \in S} r_{c f_i}}{\sqrt{k + 2\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \rho_{f_i f_j}}}$$

2. mRMR feature selection

- **Minimum-redundancy-maximum-relevance (mRMR)** can use **correlation**, **distance** scores (e.g., D_{KL}) or **mutual information** to select features.

- For feature set S of features f_i, and class c,
  $\boldsymbol{D(S,c)}$: a measure of **relevance** S has for c, and
  $\boldsymbol{R(S)}$: a measure of the **redundancy** within S,

$$\boxed{\mathrm{S}_{mRMR} = \underset{S}{\operatorname{argmax}} \left[D(S,c) - R(S)\right]}$$

2. mRMR feature selection

- Measures of **relevance** and **redundancy** can make use of our familiar measures of *mutual information*,

- $D(S, c) = \frac{1}{\|S\|} \sum_{f_i \in S} I(f_i; c)$

- $R(S) = \frac{1}{\|S\|^2} \sum_{f_i \in S} \sum_{f_j \in S} I(f_i; f_j)$

- mRMR is **robust** but doesn't measure **interactions** of features in estimating c (for that we could use ANOVAs).

3. χ^2 method

- We adapt the χ^2 method we saw when testing whether distributions were significantly different:

$$\chi^2 = \sum_{c=1}^{C} \frac{(O_c - E_c)^2}{E_c} \quad \Rightarrow \quad \chi^2 = \sum_{c=1}^{C} \sum_{f_i=f}^{F} \frac{\left(O_{c,f} - E_{c,f}\right)^2}{E_{c,f}}$$

where $O_{c,f}$ and $E_{c,f}$ are the observed and expected number, respectively, of times the class c occurs together with the (discrete) feature f.
  - The expectation $E_{c,f}$ assumes c and f are **independent**.

- Now, **every feature has a *p*-value**. A lower *p*-value means c and f are *less likely* to be independent.
- Select the *k* features with the lowest *p*-values.

Multiple comparisons

- If we're just **ordering** features, this χ^2 approach is (mostly) fine.
- But what if we get a 'significant' *p*-value (e.g., $p < 0.05$)? Can we claim a significant effect of the class on that feature?

- Imagine you're flipping a coin to see if it's fair. You claim that if you get 'heads' in 9/10 flips, it's biased.
- Assuming H_0, the coin is fair, the probability that a fair coin would come up heads $\geq$ 9 out of 10 times is:

$$(10 + 1) \times 0.5^{10} = 0.0107$$

Number of ways 9 flips are heads

Number of ways all 10 flips are heads

Multiple comparisons

- But imagine that you're simultaneously testing **173** coins – you're doing **173** (**multiple**) **comparisons**.
- If you want to see if *a specific chosen* coin is fair, you still have only a 1.07% chance that it will give heads $\geq \frac{9}{10}$ times.
- **But** if you don't preselect a coin, what is the probability that *none* of these fair coins will accidentally appear biased?

$$(1 - 0.0107)^{\mathbf{173}} \approx 0.156$$

- If you're testing 1000 coins?

$$(1 - 0.0107)^{\mathbf{1000}} \approx 0.0000213$$

UNIVERSITY OF TORONTO

Multiple comparisons

- The more features you evaluate with a statistical test (like χ^2), the more likely you are to accidentally find spurious (incorrect) significance **accidentally**.

- Various compensatory tactics exist, including **Bonferroni correction**, which basically divides your level of significance required, by the number of comparisons.
  - E.g., if $\alpha = 0.05$, and you're doing **173** comparisons, each would need $p < \frac{0.05}{173} \approx 0.00029$ to be considered significant.

UNIVERSITY OF TORONTO

Readings

- J&M: 5.1-5.5 (2^{nd} edition)
- M&S: 16.1, 16.4

Features and classification

- We talked about:
  - How preprocessing can effect feature extraction.
  - What parts-of-speech are, and how to identify them.
  - How to prepare data for classification
  - SVMs
  - Decision trees (which are parts of random forests)
  - Feature selection
    - By correlation
    - By mRMR
    - By χ^2
- Again, we've only taken our first step into the water…

Appendix – prepositions from CELEX

of	540,085	through	14,964	worth	1,563	pace	12
in	331,235	after	13,670	toward	1,390	nigh	9
for	142,421	between	13,275	plus	750	re	4
to	125,691	under	9,525	till	686	mid	3
with	124,965	per	6,515	amongst	525	o'er	2
on	109,129	among	5,090	via	351	but	0
at	100,169	within	5,030	amid	222	ere	0
by	77,794	towards	4,700	underneath	164	less	0
from	74,843	above	3,056	versus	113	midst	0
about	38,428	near	2,026	amidst	67	o'	0
than	20,210	off	1,695	sans	20	thru	0
over	18,071	past	1,575	circa	14	vice	0

Appendix – particles

aboard	aside	besides	forward(s)	opposite	through
about	astray	between	home	out	throughout
above	away	beyond	in	outside	together
across	back	by	inside	over	under
ahead	before	close	instead	overhead	underneath
alongside	behind	down	near	past	up
apart	below	east, etc.	off	round	within
around	beneath	eastward(s),etc.	on	since	without

UNIVERSITY OF TORONTO

Appendix – conjunctions

and	514,946	yet	5,040	considering	174	forasmuch as	0
that	134,773	since	4,843	lest	131	however	0
but	96,889	where	3,952	albeit	104	immediately	0
or	76,563	nor	3,078	providing	96	in as far as	0
as	54,608	once	2,826	whereupon	85	in so far as	0
if	53,917	unless	2,205	seeing	63	inasmuch as	0
when	37,975	why	1,333	directly	26	insomuch as	0
because	23,626	now	1,290	ere	12	insomuch that	0
so	12,933	neither	1,120	notwithstanding	3	like	0
before	10,720	whenever	913	according as	0	neither nor	0
though	10,329	whereas	867	as if	0	now that	0
than	9,511	except	864	as long as	0	only	0
while	8,144	till	686	as though	0	provided that	0
after	7,042	provided	594	both and	0	providing that	0
whether	5,978	whilst	351	but that	0	seeing as	0
for	5,935	suppose	281	but then	0	seeing as how	0
although	5,424	cos	188	but then again	0	seeing that	0
until	5,072	supposing	185	either or	0	without	0

Appendix – Penn TreeBank PoS tags

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			