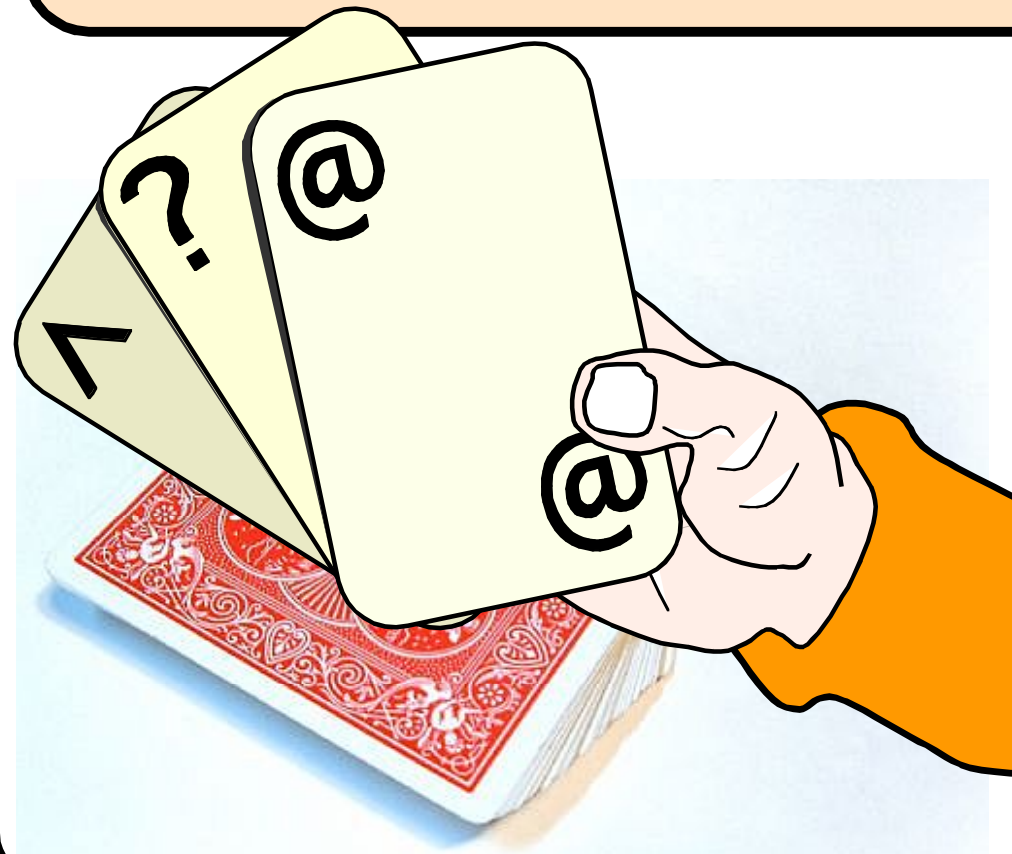


# *Word Knowledge vs. World Knowledge*

*Augmenting WordNets with (Un)Common Sense  
for Robust Web Applications*



**Tony Veale (+ students)**

**School of Computer Science  
and informatics**

**University College Dublin**

**[Tony.Veale@UCD.ie](mailto:Tony.Veale@UCD.ie)**

**<http://Afflatus.UCD.ie>**

*Helsinki, November 2011*

## The Missing Link: Bridging “Word” and “World” Knowledge

*Real Texts (in Real Life) use words and categories in unexpected ways ...*



The Rutting Chimpanzee



The Rational Animal

## **WordNets vs. WorldNets: Lightweight vs. Heavyweight**

**Dictionaries & WordNets are just one part of a language-processing solution**

We must be realistic about what WordNets can and cannot offer the user

**WordNets are simultaneously aimed at very different kinds of user**

*Linguists & language scholars | dictionary users | AI/NLP computer scientists*

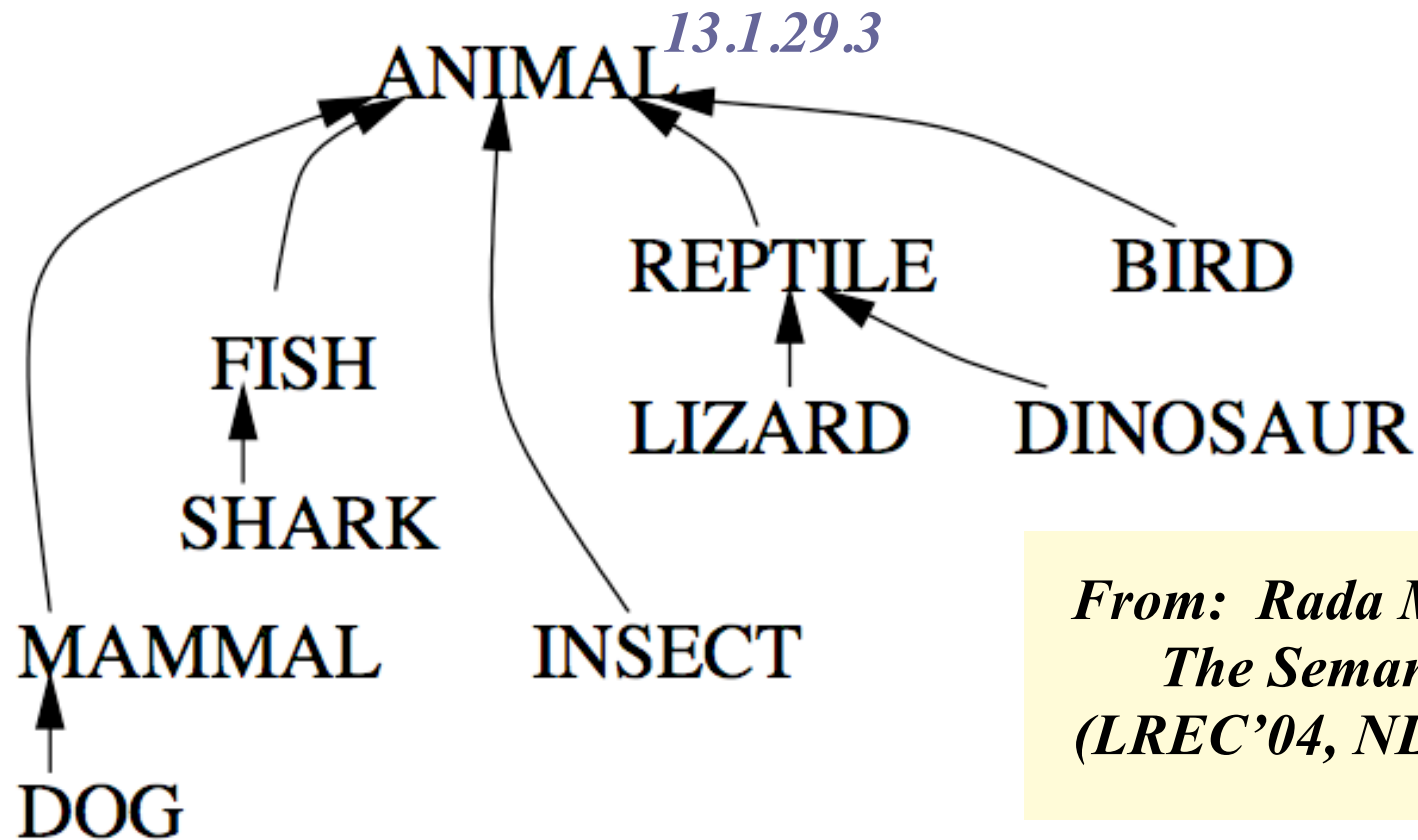
**WordNets are *lightweight* ontologies. WordNets are not WorldNets**

*We can integrate WNs with sources of encyclopaedic knowledge [ Cyc | SUMO ]*

**Natural language processing requires word knowledge and world knowledge**

WordNets provide most of the former, some of the latter. But we need more ...

## Using WordNets for Semantics: Rada Mihalcea's Semantic Wildcard

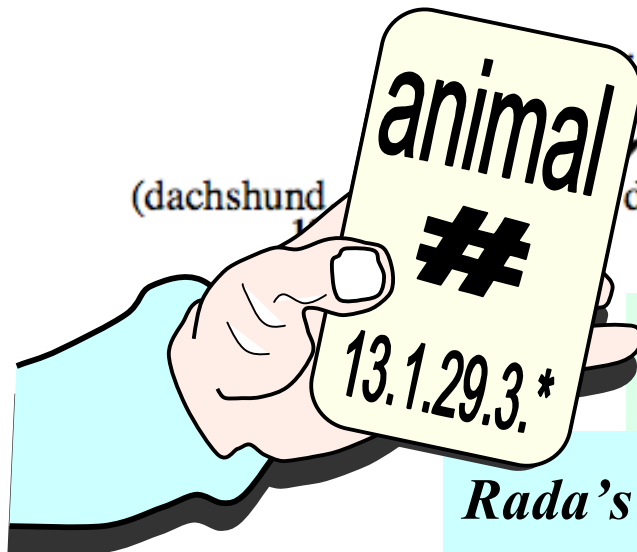
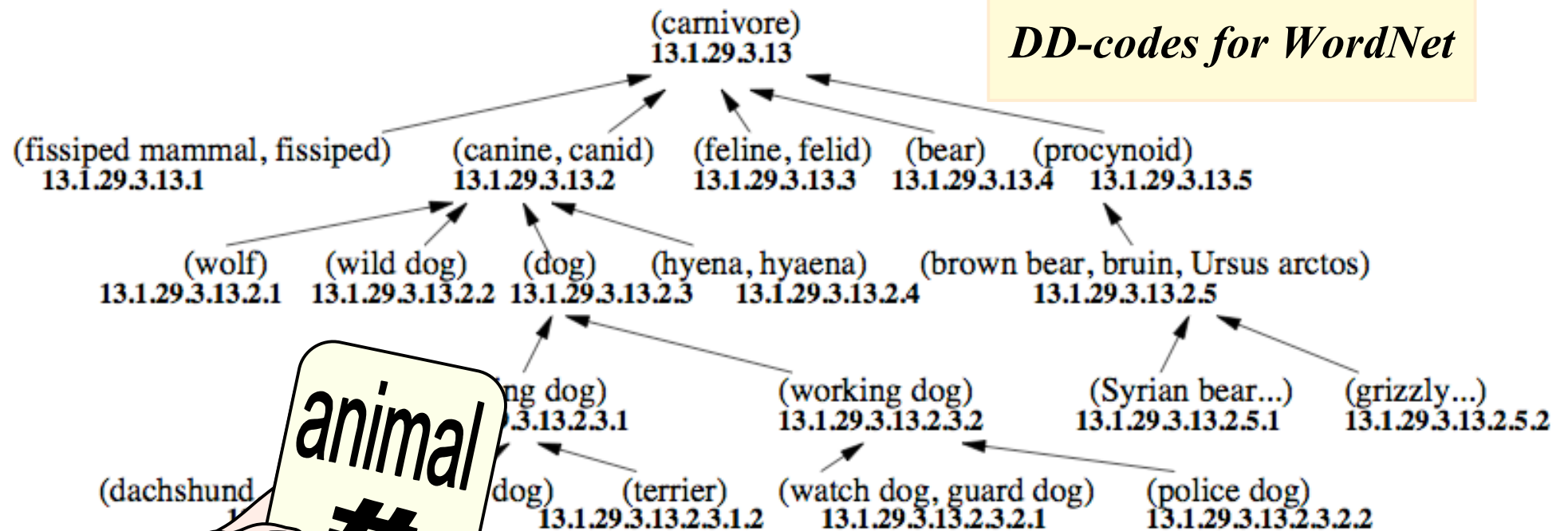


*From: Rada Mihalcea (2004):  
The Semantic Wildcard  
(LREC'04, NLP/IR workshop)*

*Trec Q: What was the largest dinosaur? A: Diplodocus? Argentinosaurus?*

## Using WordNet for Answer Retrieval: **Semantic Wildcard Matching**

*DD-codes for WordNet*

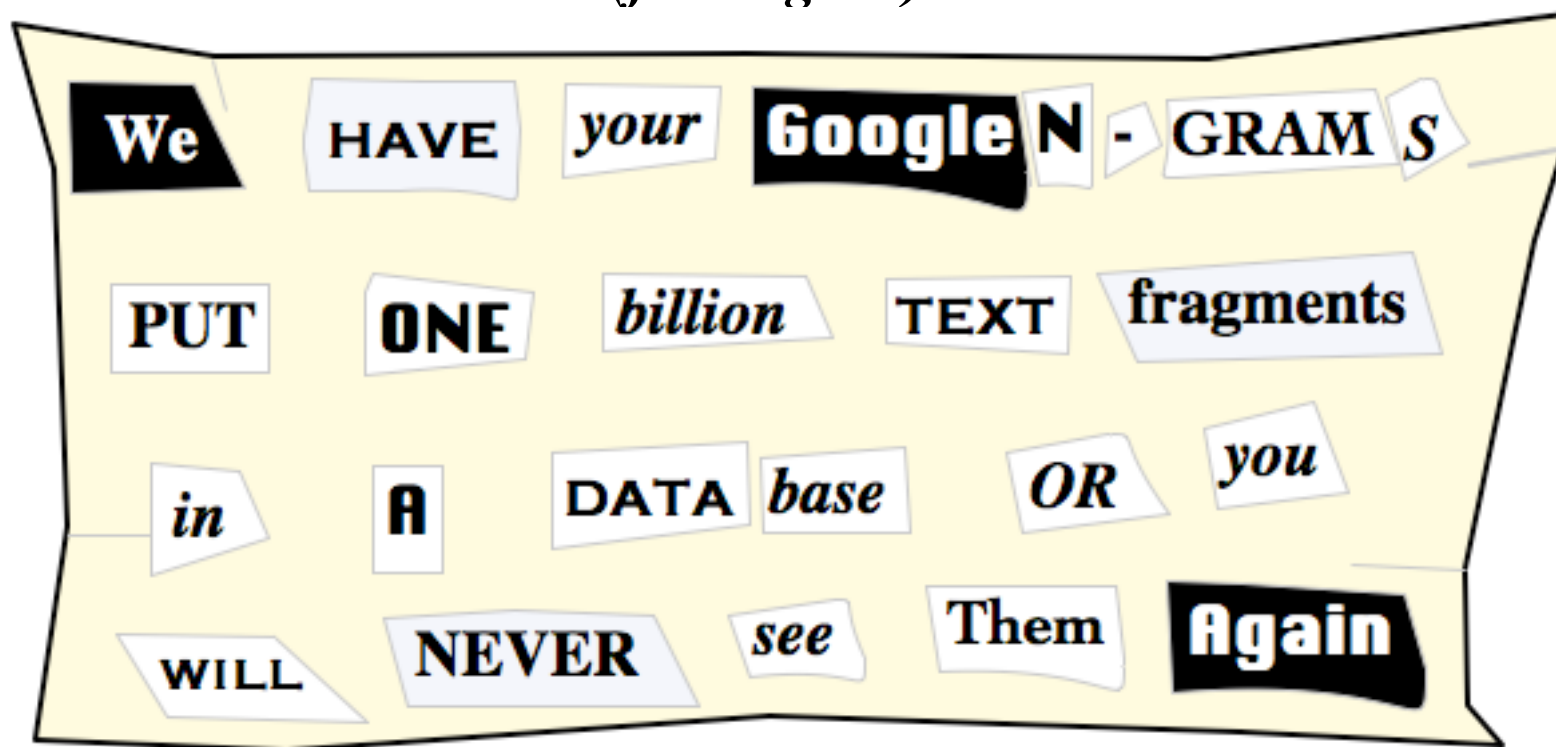


*Trec Q: What animal do Buffalo wings come from?*

*Rada's Semantic Query: animal# "Buffalo wing"*

## A More Fluid View of Semantic Categories: **Large Corpora**

*The Google N-Grams is vast database of recurring text fragments on the Web  
(for English)*



*Web n-grams: suited to mining knowledge from recurring small text fragments*



# Mining Collocations from Corpora: *Robust Category Structures*

## 5-grams

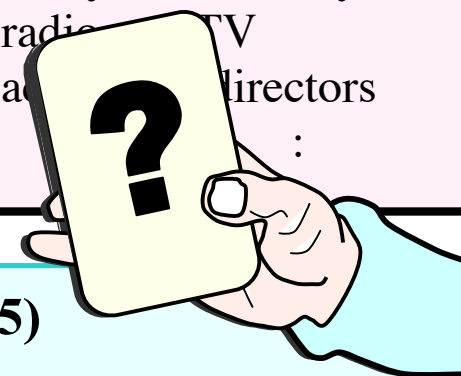
*Roger Federer , Tiger Woods*  
 Rafael Nadal and Roger Federer  
 Roger Federer, Andy Roddick  
*Thierry Henry , Roger Federer*  
*Tiger Woods , Roger Federer*  
 David Beckham, Thierry Henry  
 Tom Cruise & David Beckham  
 Tom Cruise and Katie Holmes  
 Steven Spielberg, Tom Cruise  
 Tom Hanks / Steven Spielberg  
 Dan Brown and Tom Hanks  
*Tiger Woods vs. Thierry Henry*  
 : : : : :

## 4-grams

*tennis and golf players*  
 tennis / squash players  
 soccer and hockey moms  
 polo and tennis teams  
 squash and tennis courts  
 soccer and rugby fields  
 tennis and soccer fans  
*soccer and tennis players*  
 polo and lacrosse teams  
*soccer vs. golf players*  
 TV and movie stars  
 radio and TV stars  
 : : : :

## 3-grams

*tennis and golf*  
 polo and tennis  
 artists and scientists  
 apples and oranges  
 players and fans  
 coaches and players  
*golf vs. soccer*  
 terrorism and extremism  
*soccer versus tennis*  
 Hollywood / Bollywood  
 radio and TV  
 actors and directors  
 :



E.g., we use the Google N-Grams 1T Web Corpus ( $N \leq 5$ )

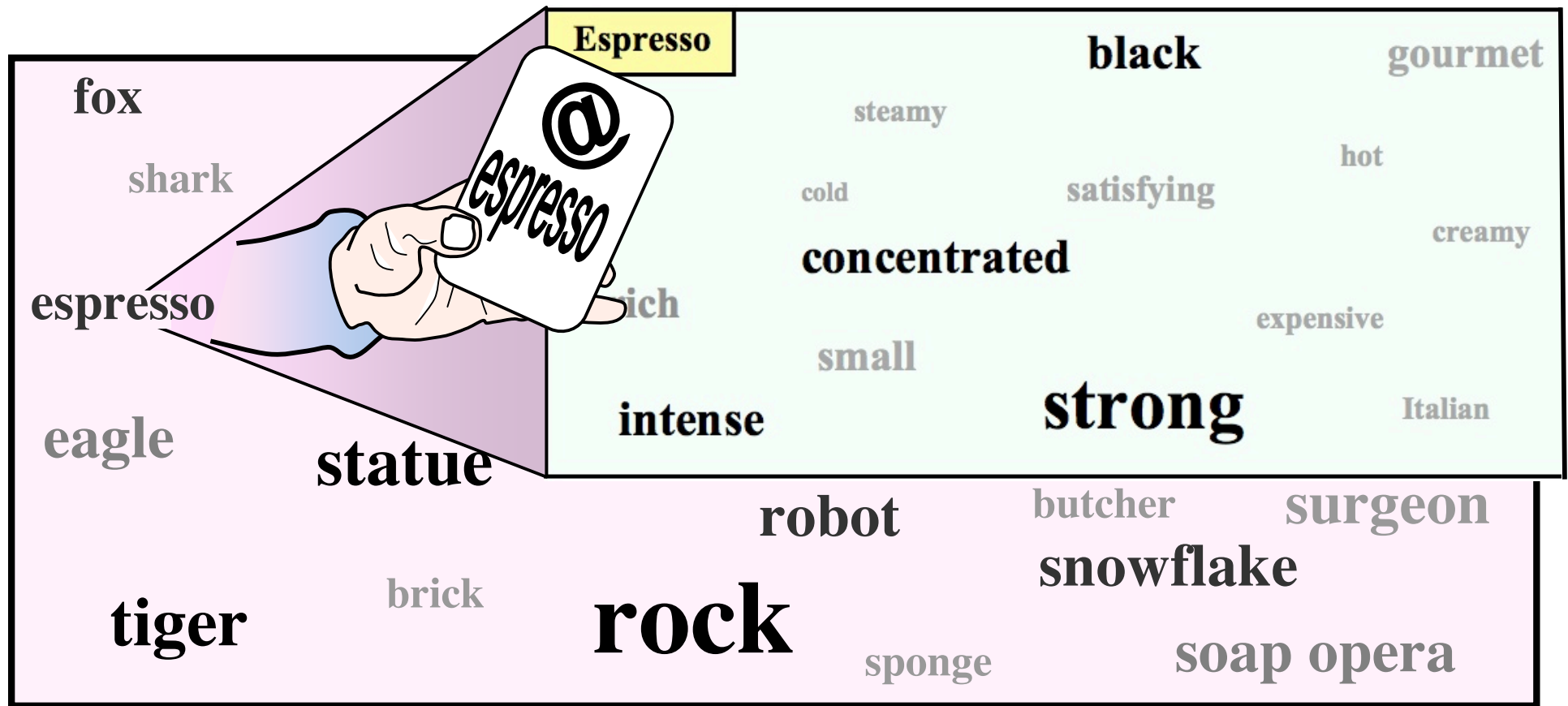
## Pragmatic Comparability Versus Semantic Similarity

disaster		terrorist		beast	
<i>tragedy</i>	99	<i>extremist</i>	90	<i>savage</i>	97
<i>catastrophe</i>	99	<i>radical</i>	88	<i>animal</i>	96
<i>calamity</i>	98	<i>anarchist</i>	83	<i>brute</i>	95
<i>destruction</i>	90	<i>subversive</i>	83	<i>wolf</i>	94
<i>famine</i>	89	<i>revolutionary</i>	82	<i>vulture</i>	86
<i>hardship</i>	89	<i>insurgent</i>	72	<i>pet</i>	83
<i>plague</i>	89	<i>separatist</i>	72	<i>plant</i>	73
<i>misfortune</i>	88	<i>guerrilla</i>	71	<i>thief</i>	73
<i>mishap</i>	85	<i>tyrant</i>	71	<i>bird</i>	70
<i>affliction</i>	84	<i>hacker</i>	70	<i>reptile</i>	64
<i>death</i>	80	<i>rebel</i>	70	<i>bandit</i>	63
<i>explosion</i>	80	<i>liberal</i>	70	<i>insect</i>	63
:		:		:	

Calculate WordNet-based semantic similarity for each coordination



## Stereotypical Associations: Mine *Simile* patterns from the WWW



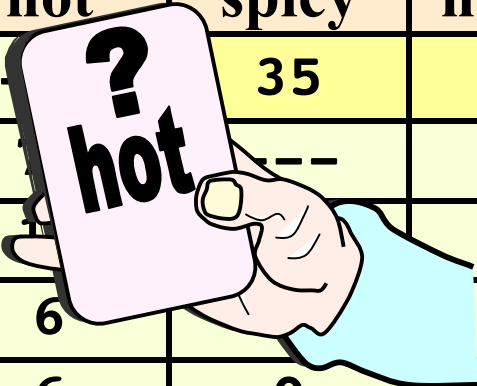
Use Web query pattern

“ as \* a | an as \* ”

to harvest 1000's of similes

## Stereotypical **Properties** co-occur in pragmatic clusters

Adjacency matrix of *mutually-reinforcing* properties acquired from WWW:



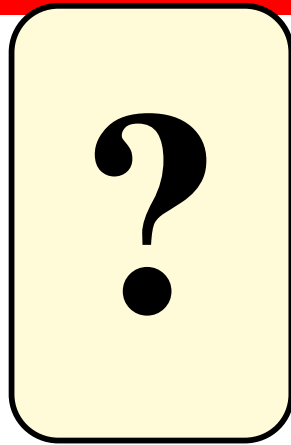
	hot	spicy	humid	fiery	dry	sultry	...
hot	---	35	39	6	34	11	...
spicy	---	---	0	15	1	1	...
humid	---	---	---	0	1	0	...
fiery	6	---	0	---	0	0	...
dry	6	0	0	0	---	0	...
sultry	11	1	0	2	0	---	...
...	...	...	...	...	...	...	...

Use the Google query

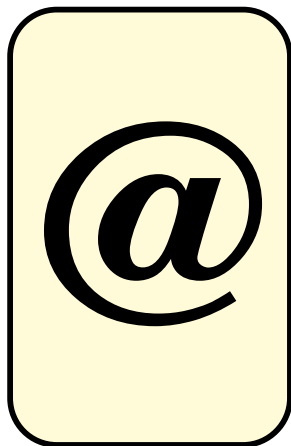
**“as \* and \* as”**

to acquire associations

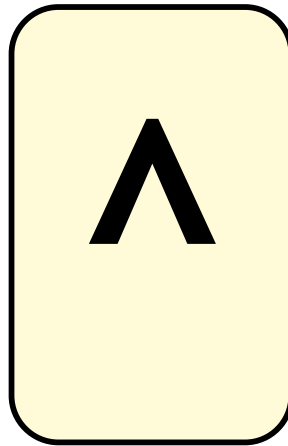
# Creative Information Retrieval with Pragmatic Wildcards



*Pragmatic Neighbourhood*



*Stereotype*



*Named Category*



*A mix of wildcards that can  
combined with each other.  
Derived from **WordNet**,  
**Wikipedia** and large corpora  
(e.g., **Google ngrams**)*



*Matches: savage, animal, brute...*



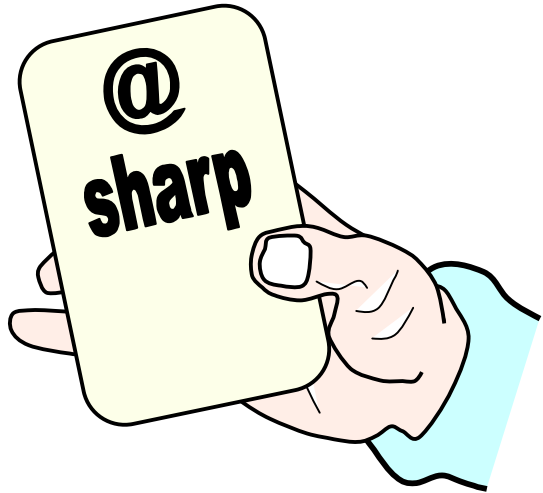
*Matches: painful, nasty, depraved...*



*Matches: brute, barbarian, bully,  
cannibal, criminal ...*



*Matches: violent, dirty, vile,  
embarrassing, humiliating...*



*Matches: sword, razor, laser...*



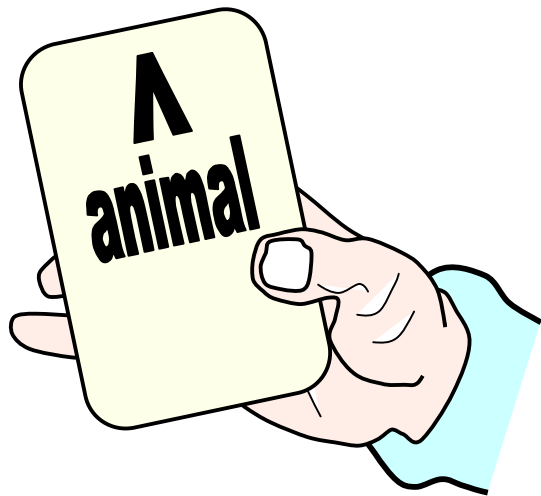
*Matches: strong, ugly, wild...*



*Matches: stiff, lethal, piercing...*



*Matches: tiger, toad, ape, bear...*



*Matches: dog, cat, wolf, ape,...*



*Matches: wild, strong, hairy,...*



*Matches: violent, ugly, vile,  
muscular, scary, ...*



*Matches: bull, bear, gorilla,  
hog, warthog, dog...*

Afflatus.UCD.ie/aristotle

## Generating Metaphor

Input the concept you would like to metaphorically describe (called *the tenor*).  
Then choose the property of the tenor you would like to accentuate.

Retrieve for P:  $(?P \cap @@P) @P$

Tenor:

speaker

Property:

dangerous  
dangerous

Category:

person

Click on any vehicle to see the properties and facets that it imparts to a tenor, and further click properties and facets to see the relevant ones.

Vehicles:	Other Properties:	Aspects of predator	deadly:teeth is like
lioness	dangerous(66)	deadly:teeth	sharp teeth
samurai	strong(58)	dangerous:nature	strong teeth
tank	formidable(53)	formidable:nature	sharp teeth
predator	tough(39)	deadly:nature	ferocious teeth
wolf	deadly(32)	strong:nature	fierce teeth
shark	heinous(2)	heinous:nature	poisonous teeth
blade		tough:nature	strong teeth
panther			

"A speaker is deadly" could imply that a speaker is:

evil(5)	effective(4)	powerful(3)	significant(2)	provocative(2)	competent(2)
expensive(1)	tiny(1)	advanced(1)	professional(1)	new(1)	

Emphatic Uses:





*George Orwell expressed a deep and persuasive distrust of re-use & readymades in language*



*The Readymade / Originality debate had already been placed centre-stage by Marcel Duchamp*  
*His infamous “Fountain” showed that art required neither true originality nor manual craft*





**Creativity** arises as much from **intentions** as from **meanings**, and from **decisions** as much as **actions**

**Everyday objects**, wrenched from their **conventional** contexts of use, can acquire **resonant** new meanings in **new** contexts.

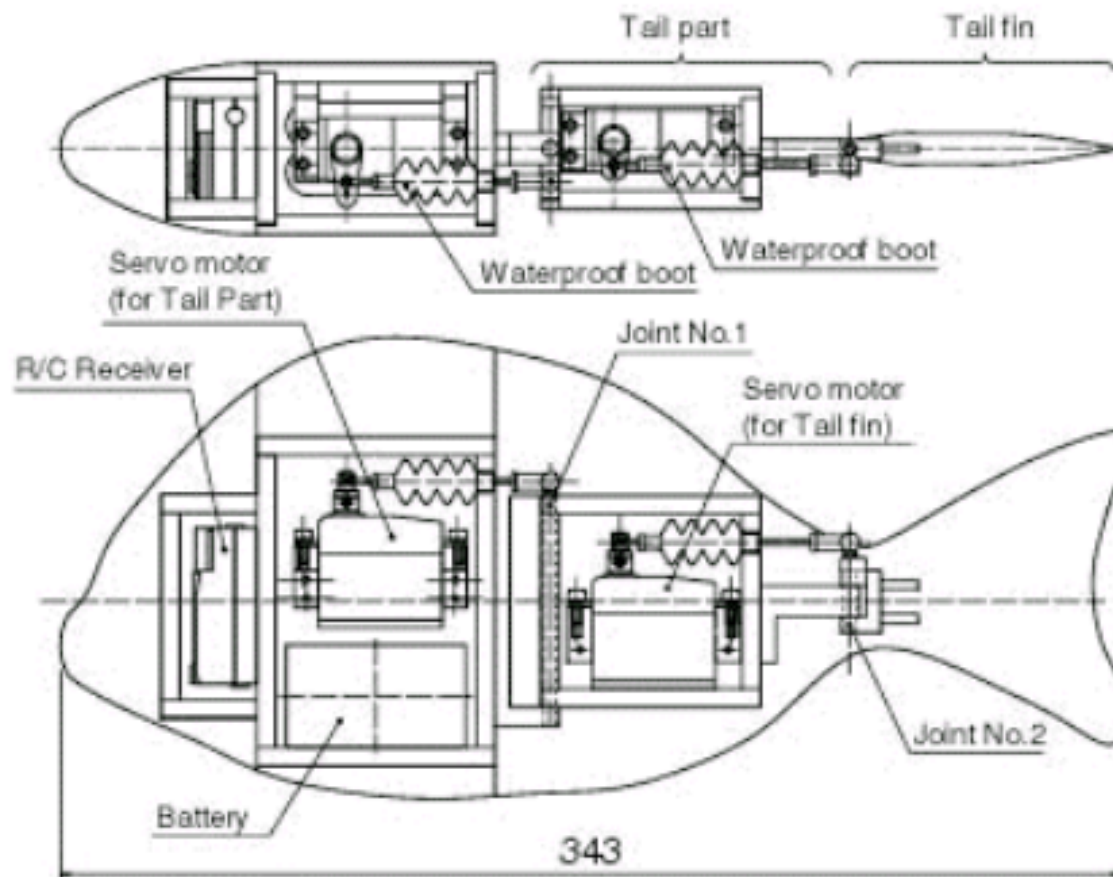
And so it is for **familiar phrases**. If wrenched from their **common** contexts of use, they can acquire **creative** new meanings.

Example: The Google 2-gram

**“robot fish”**

*Stereotype for coldness*

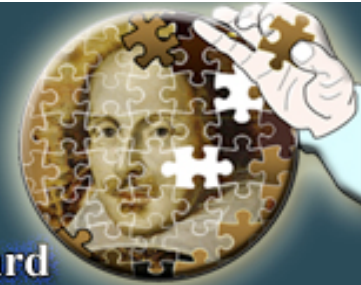
*Stereotype for coldness*



Input an adjectival property

as cold

as Go



Afflatus.UCD.ie/jigsaw

## The Jigsaw Bard

Phrases in blue are computer-generated; all other phrases are automatically mined from large corpora.

### Co-Occurring Properties of 'cold'

cold and slippy  
cold and dreary  
cold and heartless  
cold and motionless  
cold and miserable  
cold and inorganic  
cold and unsympathetic

### Peotic Elaborations

the eye of a storm (10365)  
the eye and power of a storm (10365)  
the eye and voice of a storm (10365)  
the eye and air of a storm (10365)  
the eye and wake of a storm (10365)  
the power of a storm (2828)  
the power and eye of a storm (2828)  
the power and voice of a storm (2828)  
the power and air of a storm (2828)  
the power and fury of a storm (2828)

### Simple Elaborations

a wet haddock (6155)  
a wet fish (6152)  
a wet snow (6142)  
a wet January (6118)  
a wet storm (6112)  
a wet cucumber (6111)  
a wet mackerel (6109)  
a wet snowball (6106)  
a wet snowstorm (6106)  
an unfeeling robot (2411)  
a heartless robot (2207)  
a gray January (2109)  
a lifeless corpse (2031)  
a lifeless robot (2006)  
a bitter storm (1714)  
a bitter January (1713)  
a bitter snowstorm (1707)  
a pale corpse (1610)

### Complex Elaborations

a fish-eyed storm (10040)  
a glacier with the eye of a fish (10040)  
the belly of a fish (10032)  
the wake of a storm (10032)  
the wall of a cave (10032)  
a snow blizzard (10029)  
a snowy January (10023)  
a fridge with a refrigerator freezer (10023)  
a refrigerator freezer (10023)  
the flesh of a fish (10022)  
the fury of a storm (10020)  
a bullet-riddled corpse (10019)  
the eyes of a fish (10018)  
the power of a storm (10018)  
a robotic fish (10018)  
the surface of a steel (10018)  
the heart of a killer (10017)  
the darkness of a cave (10016)

Afflatus Home Aristotle Sardonius

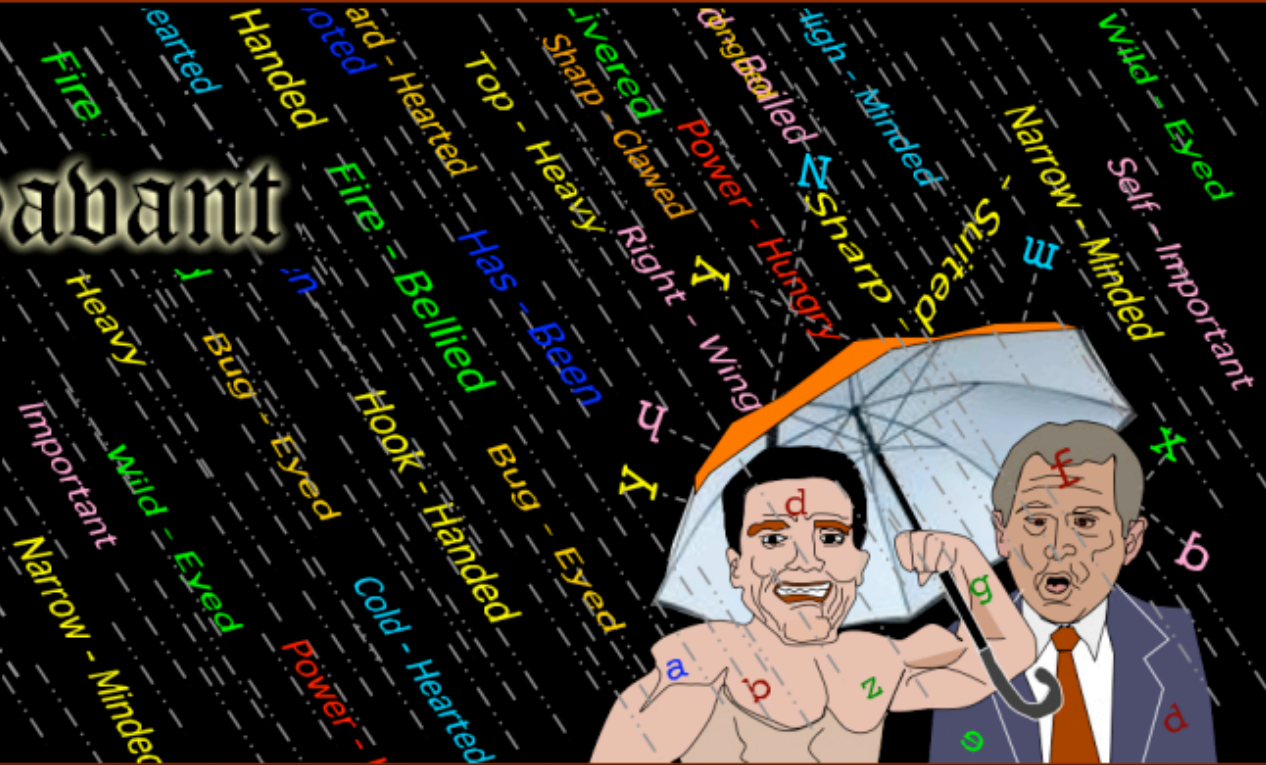
Retrieve: ?P @P

Retrieve: @P @P



critic

# Idiom Savant



[Afflatus.UCD.ie/idiom-savant/](http://Afflatus.UCD.ie/idiom-savant/)

critic

query

# Idiom Sabant



Next Page

## Positive Descriptions

high-profile critic (100)  
well-known critic (100)  
real-life critic (100)  
well-informed critic (100)  
long-standing critic (100)  
long-time critic (100)  
self-made critic (100)  
fair-minded critic (100)  
first-rate critic (100)  
good-natured critic (100)  
well-organized critic (100)  
hard-nosed critic (100)  
self-respecting critic (100)  
top-notch critic (100)  
pre-eminent critic (100)  
well-intentioned critic (100)

^pos-phrase ?X

## Negative Descriptions

sharp-tongued critic (100)  
self-important critic (100)  
acid-tongued critic (100)  
alias-shrouded critic (100)  
knee-jerk critic (100)  
wild-eyed critic (100)  
right-wing critic (100)  
die-hard critic (100)  
hard-line critic (100)  
self-proclaimed critic (100)  
narrow-minded critic (100)  
would-be critic (100)  
hard-hearted critic (100)  
self-serving critic (100)  
serious-minded critic (100)  
far-right critic (100)

^neg-phrase ?X

## Perspectives

reviewer  
reader  
judge  
professional  
cynic  
monster  
educator  
lawyer  
librarian  
practitioner  
publisher  
attorney  
authority  
advocate  
academic  
connoisseur  
detractor



# Common Questions On the Web: A Source of World Knowledge

why do cats **purr** × Search

why do cats **purr**  
why do cats **eat grass**  
why do cats **sleep so much**  
why do cats **cry**  
why do cats **have whiskers**

About 1,790,000 results (0.26 seconds) [Go to Google.com](#) [Advanced search](#)

pourquoi les chats **ronronnent** × Rechercher

pourquoi les chats **ronronnent**  
pourquoi les chats **n'aiment pas l'eau**  
pourquoi les chats **mordent**  
pourquoi les chats **miaulent**  
pourquoi les chats **remuent la queue**

[En savoir plus](#)

Environ 144 000 000 résultats (0,05 secondes) [Google.com in English](#) [Recherche avancée](#)

*We “milk” question completions from Google, and parse them into axioms*












Google is a cult

E.g., Scientists as Artists, or, just Scientists

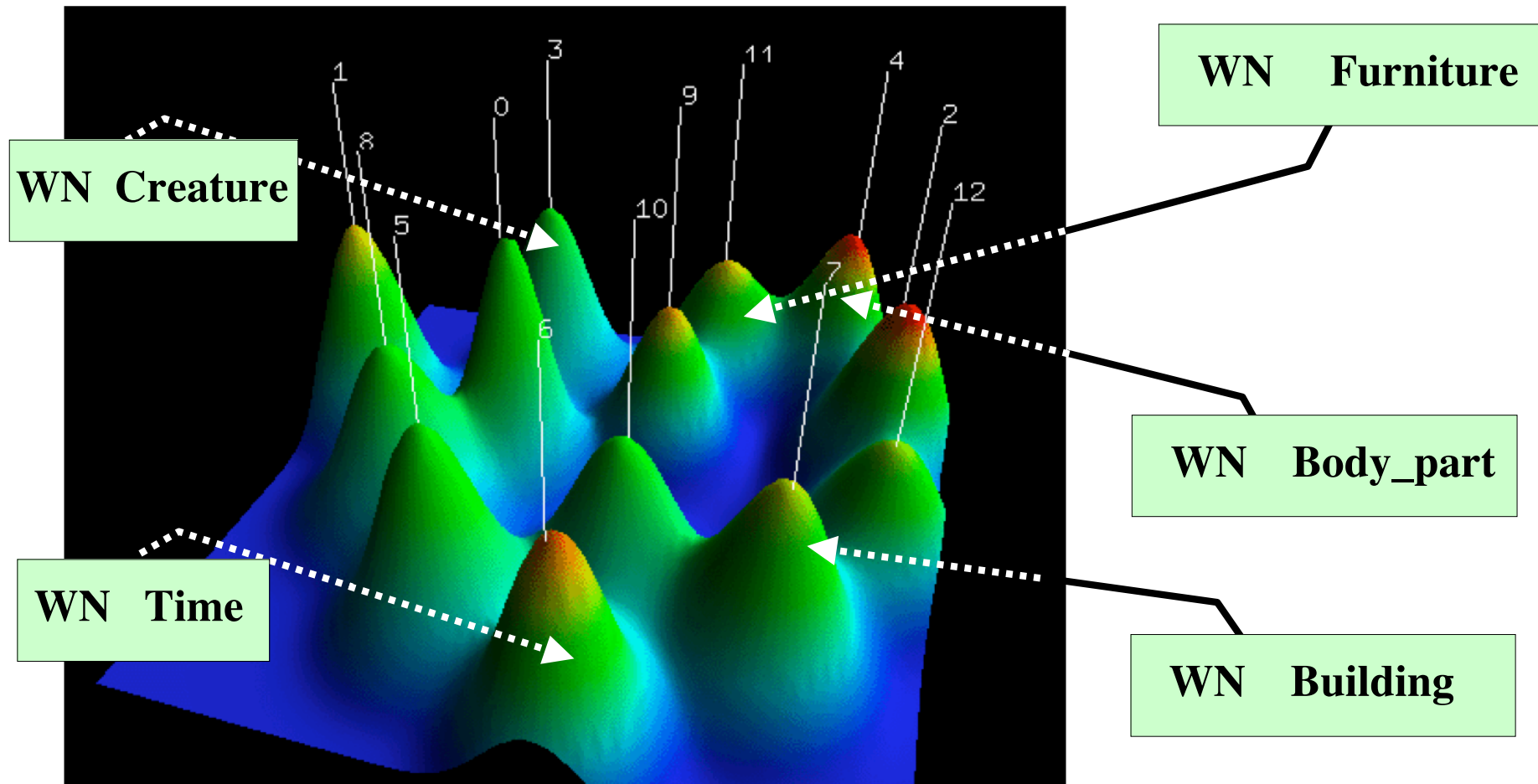
Metaphorize This!

[ngrams.UCD.ie/metaphor-eye/](http://ngrams.UCD.ie/metaphor-eye/)

### Mashups for Google as cult

-  why does Google have apologists like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj apologist) (have ?Subj ?Obj)))))
-  why does Google enforce beliefs like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj belief) (enforce ?Subj ?Obj)))))
-  why does Google promote beliefs like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj belief) (promote ?Subj ?Obj)))))
-  why does Google worship celebrities like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj celebrity) (worship ?Subj ?Obj)))))
-  why does Google worship gods like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj god) (worship ?Subj ?Obj)))))
-  why does Google worship gurus like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj guru) (worship ?Subj ?Obj)))))
-  why is Google led by gurus like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj guru) (led\_by ?Subj ?Obj)))))
-  why does Google follow gurus like cult  score: 100, support: 0, similarity: 0  
(=> (instance ?Subj google) (exists (?Obj) (and (instance ?Obj guru) (follow ?Subj ?Obj)))))
-  why is Google established by gurus like cult  score: 100, support: 0, similarity: 0

# Using **WordNet(s)** as a **Gold-Standard** for **Pragmatic WorldNet(s)**



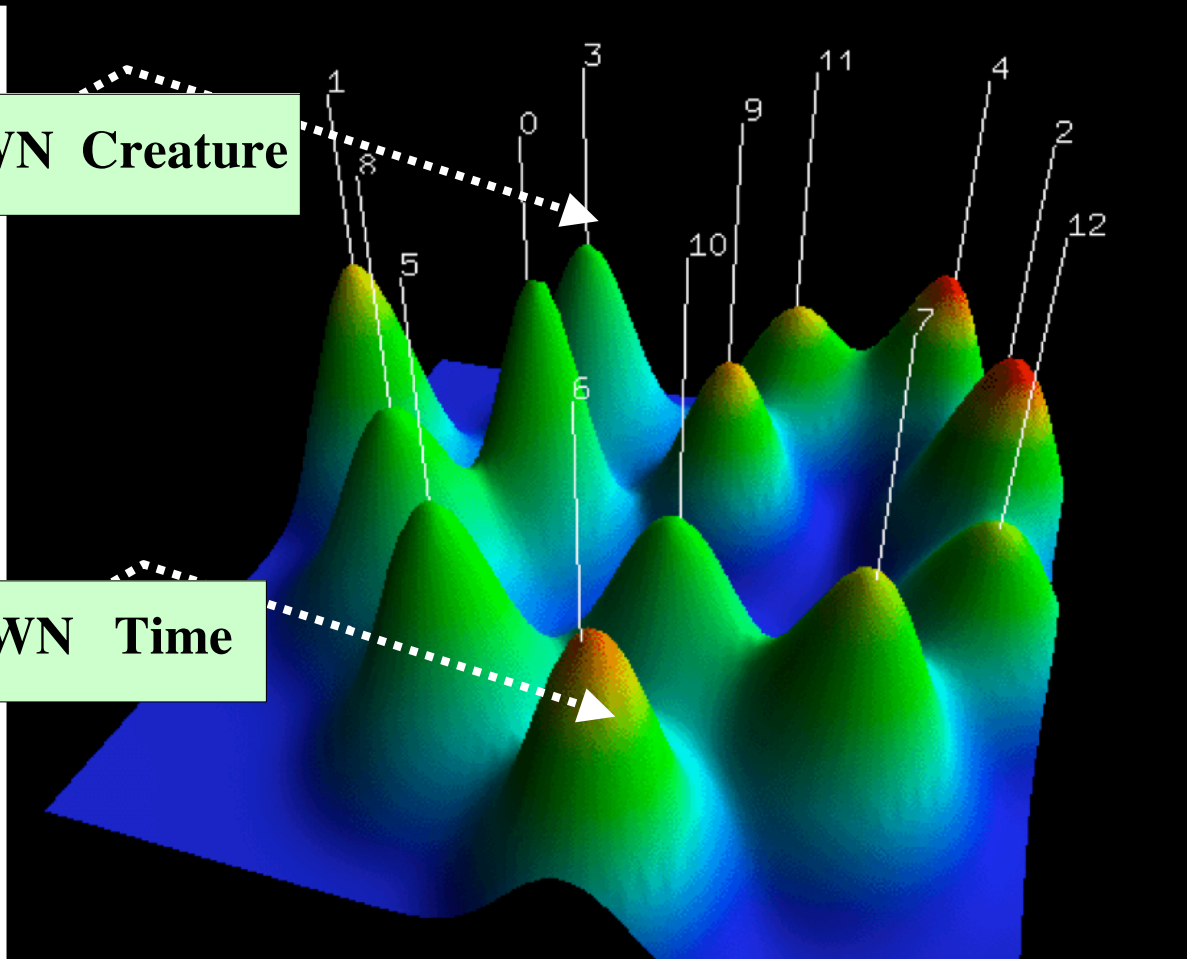
E.g., Almuhareb & Poesio (2004/2005), Veale and Hao (2007/2008)

## How Well Does Corpus/Web-based Categorization Do w.r.t. WordNet?

13-way clustering of 214 nouns, compared to WordNet

WN Creature

WN Time



Almuhareb & Poesio (2004)

*Weak text-derived features*

~ 60,000 features for 214 nouns

Result: 0.855 cluster purity

Veale & Hao (2007 / 2008)

*Strong simile-derived features*

~ 7,300 features for 214 nouns

Result: 0.902 cluster purity

Veale & Li (2009)

*Generic Clique-derived features*

~ 8,300 features for 214 nouns

Result: 0.934 cluster purity



## Semantics vs. Pragmatics: Similarity vs. Comparison

WordNets are a good source of word knowledge, lightweight semantics  
They must be used as a coherent part of an applied, pragmatic, NLP solution

- Pragmatic knowledge can come from large corpora of real “language use”  
Comparisons, similes and other tropes are a fluid source of tacit knowledge

- WordNet provides the *semantics of similarity* (e.g., Budinitsky & Hirst)  
Large corpora / usage data provide the *pragmatics of comparability*

- A simple framework, instantiated in many ways

The *Aristotle*, *Jigsaw Bard*, *Idiom Savant* and *Metaphorize* applications



*Afflatus.UCD.ie*



*Questions?*

*Fin*